Local backbone structure prediction of proteins.

Alexandre de Brevern, Cristina Benros, Romain Gautier, Hélène Valadié, Serge A. Hazout, Catherine Etchebest

Local backbone structure prediction of proteins

A.G. de Brevern [1]*, C. Benros [1], R. Gautier [1,2], H. Valadié [1,3], S. Hazout [1] and C. Etchebest [1]

[1] Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM E03-46, Université Denis Diderot - Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France.

[2] Université Pierre & Marie Curie - Paris 6, 75005 PARIS, France.

[3] CEA, 17 avenue des Martyrs, 38054 GRENOBLE, France.

* Corresponding author:
mailing address: Dr. de Brevern A.G., Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM E03-46, Université Denis DIDEROT-Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France
E-mail : alexandre.debrevern@ebgm.jussieu.fr
Tel: (33) 1 44 27 77 31
Fax: (33) 1 43 26 38 30

Running title: LocPred

Keywords: structure prediction, confidence index, Bayesian approach.

Abstract

Summary: A statistical analysis of the PDB structures has led us to define a new set of small 3D structural prototypes called Protein Blocks (PBs). This structural alphabet includes 16 PBs, each one is defined by the (Φ, Ψ) dihedral angles of 5 consecutive residues. The amino acid distributions observed in sequence windows encompassing these PBs are used to predict by a Bayesian approach the local 3D structure of proteins from the sole knowledge of their sequences. ***LocPred*** is a software which allows the users to submit a protein sequence and performs a prediction in terms of PBs. The prediction results are given both textually and graphically.

Introduction

A classical approach to simplify 3D protein structures consists in describing the protein backbone in terms of secondary structures with repetitive α-helices and β-strands and, everything else called coils. The use of neural networks and homologous sequences has increased the prediction rate to a value close to 80 % [1-3]. However, even with such a rate, the approximation of the three-dimensional structure by only 3 states is very crude: 50 % of the residues are assigned as "coil" whereas they correspond to very different local structures.

To go further, various teams have proposed to categorize the 3D structures through a *structural alphabet*, i.e. a set of small protein fragments frequently observed in a structural databank [4]. This structural description gives new insights into the relation 1D-3D, revealing peculiar sequence specificity [5-9].

We have defined in a previous study a structural alphabet composed of 16 average protein fragments of 5 residues in length, called Proteins Blocks (PBs, see figure 1) [6]. These PBs show a good 3D approximation of the local structures with an average *RMSD* of 0.42 Å.

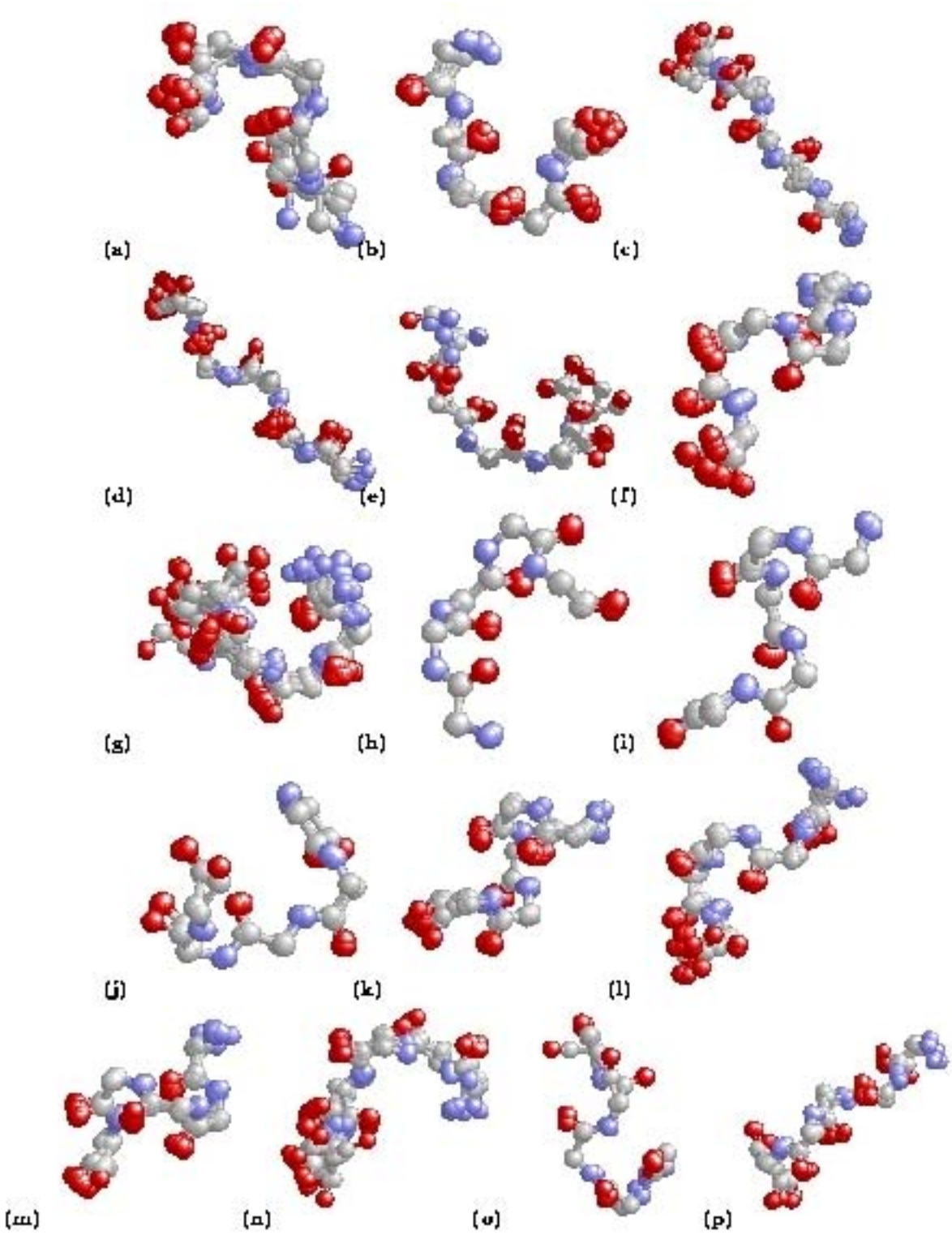

Figure 1: Superimposition of fragments of the structural alphabet. The backbone is superimposed with the atoms C_α, N, O, and C along the five amino acids of the fragments for the 16 protein blocks (PBs). PB *a* to PB *p* are displayed from left to right and from top to bottom. PBs *m* and *d* correspond roughly to the core of α-helices and β-strands, respectively.

They have also proved their reliability to describe long length fragments [10 - 13]. The main structural characteristics of the Protein Blocks are briefly pointed out in the following. PBs *a* to *f* may be related to the β-strand secondary structure, PB *d* corresponds to the more regular central part, PBs *a*, *b* and *c* to the N-caps and *e, f* to the C-caps. The PBs *k* to *p* may be related to the α-helix secondary structure, with PB *m* describing the central part of a right-handed helix, PBs *k* and *l* for the N-caps and PBs *n* to *p* for the C-caps. Finally, PBs *g* to *j* may mainly be associated with coil structures. A Bayesian approach based on the relationship between Protein Blocks and their amino acid propensities is used to perform a local structure prediction [6].

Thus, the prediction of the PB series from the sole knowledge of the protein sequence allows predicting every region of the protein without ignoring the local conformations of the coil state. Moreover, it gives a precise description of the repetitive structures [13]. Bayesian prediction gives a lower prediction rate than more sophisticated method like Artificial Neural Networks [1 – 3]. Nevertheless, it permits to analyze the role of each amino acid in the prediction and to compute an index which is directly correlated with the quality of the prediction (see *Prediction confidence index* section).

The purpose of this project was to develop a software named ***LocPred*** (*Loc*al structure *Pred*iction) based on this alphabet. ***LocPred*** is written in Java and can be used under many different platforms. The user can submit a protein sequence either in single letter amino acid code format or in *Fasta* format (Figure 2a).

Bayesian prediction

The prediction is based on the observed distributions of the amino acids in sequence windows encompassing each PB. Three options are available: (i) A *Bayesian prediction*: Tested with more than 300 sequences belonging to the Protein Databank, we have obtained an average prediction rate of 34.4%.

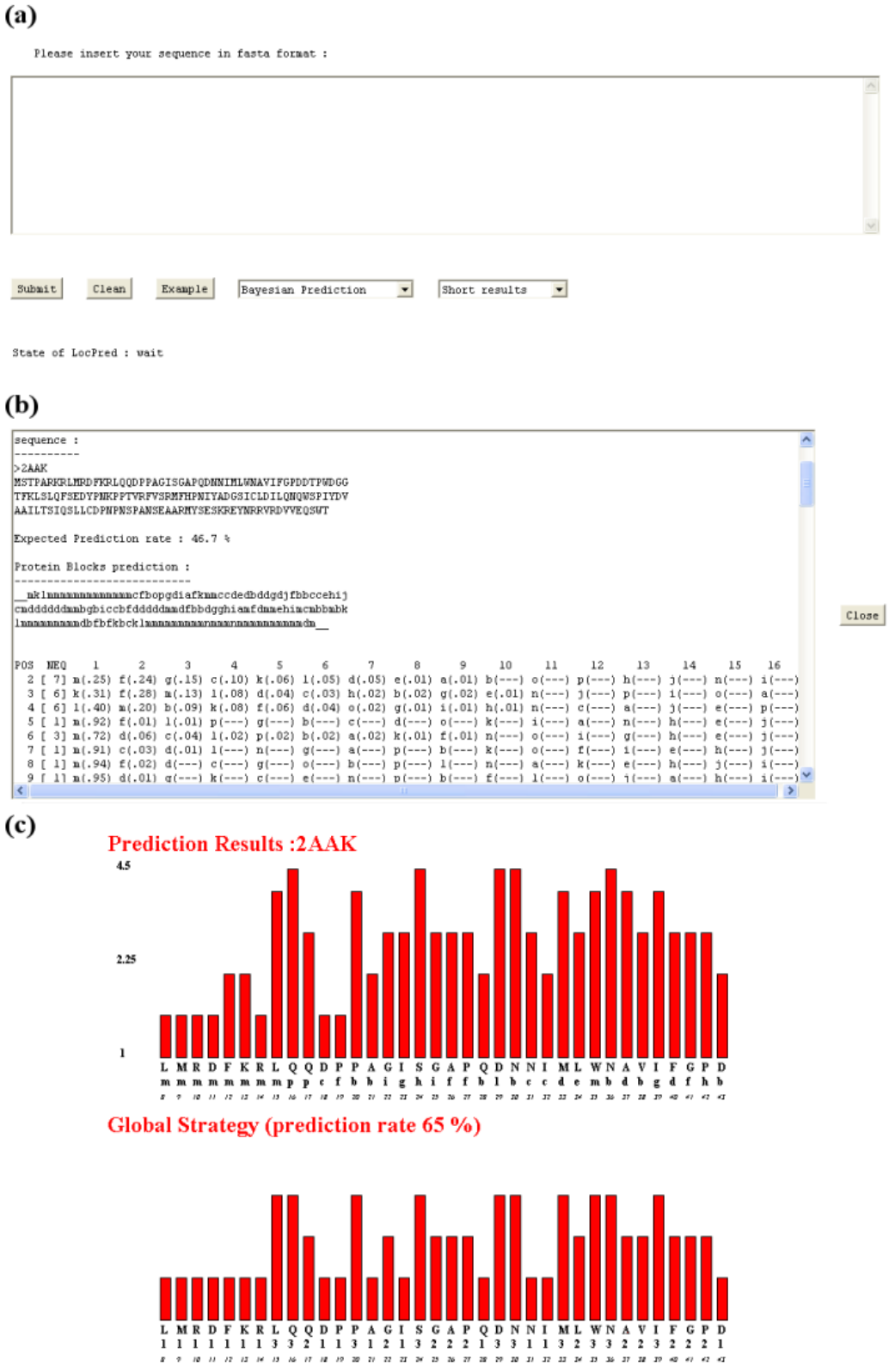

Figure 2: Prediction for the protein-conjugating enzyme (code PDB: 2aak) with (a) a window for pasting the sequence, (b) the prediction results in terms of Protein Blocks with from top to bottom the sequence, the expected prediction rate, the prediction in Fasta format with only the most probable PB and the complete prediction with all the Protein Blocks and their probabilities, and (c) top: *Neq* variation along the protein sequence and, bottom: *global strategy* with the number of selected PBs for a chosen prediction rate of 65% (2aak N-ter, residues 8 to 43).

(ii) *Sequence families* approach. This approach has been developed to optimize the sequence-structure relationship. Indeed, for one given PB, the Bayesian approach implies the use of one amino acid occurrence matrix. However, a same local fold, e.g. a PB, can be associated with different sequence clusters. So, using an optimization close to Kohonen's Self-Organizing Maps (SOM [14]), we have defined several new occurrence matrices for the most frequent PBs (for more precise details see [6]). They permit to increase the sequence – structure relationship of these PBs. This clustering in different sequence families has led to an improvement of the prediction rate to 40.7% on average. (iii) *New sequence families* approach. Moreover, we have recently improved this approach with the use of a method related to simulated annealing simulations. The prediction rate now reaches 48.7%.

The prediction score is computed along a sliding sequence window of 15 residues in length. For each sequence position, ***LocPred*** gives as outputs the most probable PBs as well as the distribution of the probabilities associated with each PB (Figure 2b).

Prediction confidence index

From this information, it is possible to define an entropy-based index called *Neq* (for *eq*uivalent *N*umber of Protein Blocks), close to the one proposed in PSIPRED [15]. The *Neq* allows one to locate strongly (*Neq* ~ 1) versus weakly (*Neq* ~ 16) informative sequence regions. We have shown that a strong correlation exists between the *Neq* values and the PB prediction success in each position. Thus, *Neq* helps to distinguish putative well predictable regions versus misleading regions.

A user would like to know if the performed prediction in terms of PBs will be correct. So, we have used the average *Neq* value taken from the prediction and a linear regression model to compute the expected prediction rate for a protein (only available for *New sequence families* approach). This latter has a standard deviation of only 5%.

Prediction strategies

In the same way, we have assessed the quality of the prediction at each position by taking into account the local *Neq* value and then proposed two distinct strategies. Both use a fixed prediction rate.

(i) The "*global strategy*": it consists in the computation of the optimal number of PBs in each position to insure a given prediction rate. So, the number of selected PBs may be variable along the sequence. Figure 1c shows the results of the prediction for the protein-conjugating enzyme with the global strategy for a prediction rate of 65%. For instance, the 7 first residues have been associated with one single PB, the next two with 3 PBs.

(ii) The "*local strategy*": the protein sequence is predicted with a constant number of PBs per position (Figure 2c). This strategy determines the regions able to be predicted with this prediction accuracy [6]. The corresponding PBs selected by each method can be downloaded.

Moreover, an online help is available on http://www.ebgm.jussieu.fr/~debrevern/LOCPRED/, as well as the 3D structures of the PBs. These strategies are interesting as a first step in an *ab initio* method [16] and could help to analyze and align appropriately sequences with low similarity. For the homology modeling with an available 3D structure or a 3D model, a rasmol script [17] can be obtained to visualize the *Neq* variations along the structure. In the same way, a comparison of a 3D structure or model translated in terms of Protein Blocks can be done.

Availability

LocPred is freely available for use through the Internet at the URL: http://www.ebgm.jussieu.fr/~debrevern/LOCPRED and can also be installed locally (same URL). It can be executed over the World Wide Web on any Java compatible Web Browser. The Java files are available at the same URL.

Acknowledgments

We would like to thank Estelle Calvez, Maxime Huvet, Laurent Fourrier and Aurélie Urbain for different tests and analyses, Joelle Hochez for the data-processing support, Patrick Fuchs and Anne-Claude Camproux for fruitful discussions.

This work was supported by a grant from the Ministére de l'Enseignement Supérieur et de la Recherche and from "Action Bioinformatique inter EPST" 2001-2002 (number 4B005F) and 2003-2004 (“Outil informatique intégré en Génomique Structurale. Vers une prédiction de la structure tridimensionnelle d'une protéine à partir de sa séquence.” and “Plateforme de bioinformatique structurale - RPBS”). AdB was supported by a grant from the Fondation de la Recherche Médicale. CB and RG have grants from the Ministère de la Recherche. HV has a grant from the Centre d’Essai Atomique (CEA). CE and SH are Professors at the University Paris 7 - Denis-Diderot, Paris. AdB is a researcher at the French Institute for Health and Medical Research (INSERM).

References

[1] Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**, 195-202.

[2] Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P., and Lund, O. (2000). Prediction of protein secondary structure at 80% accuracy. Proteins. **41,** 17-20.

[3] Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002). Improving the prediction of secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins. **47**, 228-235.

[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] 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.

[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] Bystroff, C., Thorsson, V. and Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol. **301**, 173-190.

[8] Camproux, A.C., de Brevern, A.G., Hazout, S. and Tuffery, P. (2001). Exploring the use of a structural alphabet for a structural prediction of protein loops. Theor Chem Acc. **106(1/2)**, 28-35.

[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] Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**, 59-69.

[15] Guffin, L.M., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein structure prediction

server. Bioinformatics. **16**, 404-405.

[16] Bystroff, C. and Shao, Y. (2002). Fully automated *ab initio* protein structure prediction using I-Sites, HMMSTR and Rosetta. Bioinformatics. **18**, S54-S61.

[17] Sayle, R.A. and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem Sci. **20** ,374-378.