Biophys Rev (2010) 2:137–145
DOI 10.1007/s12551-010-0036-1

REVIEW

A short survey on protein blocks

Agnel Praveen Joseph · Garima Agarwal · Swapnil Mahajan · Jean-Christophe Gelly · Lakshmipuram S. Swapna · Bernard Offmann · Frédéric Cadet · Aurélie Bornot · Manoj Tyagi · Hélène Valadié · Bohdan Schneider · Catherine Etchebest · Narayanaswamy Srinivasan · Alexandre G. de Brevern

Received: 20 April 2010 / Accepted: 2 July 2010 / Published online: 5 August 2010

Abstract Protein structures are classically described in terms of secondary structures. However, even if the regular secondary structures have relevant physical meaning, their recognition based on atomic coordinates has a number of important limitations, such as uncertainties in the assignment of the boundaries of the helical and β-strand regions. In addition, an average of about 50% of all residues are assigned to an irregular state, i.e., the coil. These limitations have led different research teams to focus on abstracting the conformation of the protein backbone in the localized short stretches. To this end, different geometric measures are being used to cluster local stretches in protein structures in a chosen number of states. A prototype representative of the local structures in each cluster is then generally defined.

Agnel Praveen Joseph and Garima Agarwal contributed equally to this article.

A. P. Joseph · J.-C. Gelly · A. Bornot · C. Etchebest · A. G. de Brevern (✉)
Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot Paris 7,
6, rue Alexandre Cabanel,
Paris Cedex 15 75739, France
e-mail: alexandre.debrevern@univ-paris-diderot.fr

A. P. Joseph · J.-C. Gelly · A. Bornot · C. Etchebest · A. G. de Brevern
Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), INSERM, UMR-S 665,
6, rue Alexandre Cabanel,
Paris Cedex 15 75739, France

A. P. Joseph · J.-C. Gelly · A. Bornot · C. Etchebest · A. G. de Brevern
Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Institut National de la Transfusion Sanguine (INTS),
6, rue Alexandre Cabanel,
Paris Cedex 15 75739, France

G. Agarwal · S. Mahajan · L. S. Swapna · N. Srinivasan
Molecular Biophysics Unit, Indian Institute of Science,
Bangalore 560012, India

S. Mahajan
National Centre for Biological Sciences,
Tata Institute of Fundamental Research,
UAS–GKVK Campus, Bellary Road,
Bangalore 560 065, India

B. Offmann · F. Cadet
INSERM, UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB),
15 Avenue René Cassin, BP 7151, 97715 Saint Denis Messag Cedex 09 La Réunion, France

B. Offmann · F. Cadet
Faculté des Sciences et Technologies, Université de La Réunion,
15 Avenue René Cassin, BP 7151, 97715 Saint Denis Messag Cedex 09 La Réunion, France

M. Tyagi
Computational Biology Branch, National Center for Biotechnology Information (NCBI),
National Library of Medicine (NLM),
8600 Rockville Pike,
Bethesda, MD 20894, USA

H. Valadié
UMR 5168 CNRS–CEA–INRA–Université Joseph Fourier,
Institut de Recherches en Technologies et Sciences pour le Vivant,
17 avenue des Martyrs, 38054 Grenoble Cedex 9, France

B. Schneider
Institute of Biotechnology AS CR,
Videnska 1083,
142 20 Prague, Czech Republic

These libraries of local structure prototypes are named "structural alphabets". We have developed a structural alphabet, denoted protein blocks, not only to approximate the protein structure but also to predict them from the sequence. Since its development, we and others have explored numerous new research fields using this structural alphabet. Here, we review some of the most interesting applications of this structural alphabet.

Keywords Protein structures · Secondary structures · Structural alphabet · Structure prediction · Structural superimposition · Mutation · Binding site

Introduction

Protein structures have been classically described in two regular states, the α-helix and β-strand, with the remaining unassigned regions described as an irregular state (coil) that corresponds to a large number of diverse conformations. However, the use of only three states oversimplifies the description of protein structures. A detailed description for 50% of the residues classified as coils is lacking even when they encompass a repeating local structure. To date, the description of local protein structures has focused on the elaboration of complete sets of small prototypes or "structural alphabets" (SAs), that help to approximate each part of the protein backbone (Offmann et al. 2007). Designing an SA requires the identification of a set of average recurrent local protein structures that (efficiently) approximates each part of the known structures. As each residue is associated to one of these prototypes, the whole three-dimensional (3D) protein structure can be translated into a series of prototypes (letters) in one dimension (1D) as the sequence of prototypes.

Figure 1 shows an example of the encoding of a protein structure with an SA. The N-terminal extremity of *Aspergillus niger* acid phosphatase (Kostrewa et al. 1999) chain B is shown. A local protein structure prototype was associated to each residue, thereby enabling the precise description of the coil region as a succession of small protein prototypes instead of as a succession of identical states.

Protein blocks

Secondary structure assignments are widely used to analyze protein structures. However, such an approach often results in a coarse description of 3D protein structures, with about half of the residues being assigned to an undefined state (Bornot and de Brevern 2006). Moreover, the structural diversity observed in α-helices and β-strands is hidden. Indeed, α-helices are frequently not linear; rather, they are either curved (58%) or kinked (17%) (Martin et al. 2005). The absence of a secondary structure assignment for a significant proportion of the residues has led to the development of local protein structure libraries that are able to approximate all (or almost all) of the local protein structures without the need for classical secondary structures. These libraries have yielded prototypes that are representative of local folds found in proteins. The complete set of local structure prototypes defines an SA (Offmann et al. 2007).

Ten years ago, Serge Hazout developed a novel SA with two specific goals (de Brevern et al. 2000): (1) to obtain a good local structure approximation and (2) to predict local structures from the sequence. Fragments five residues in

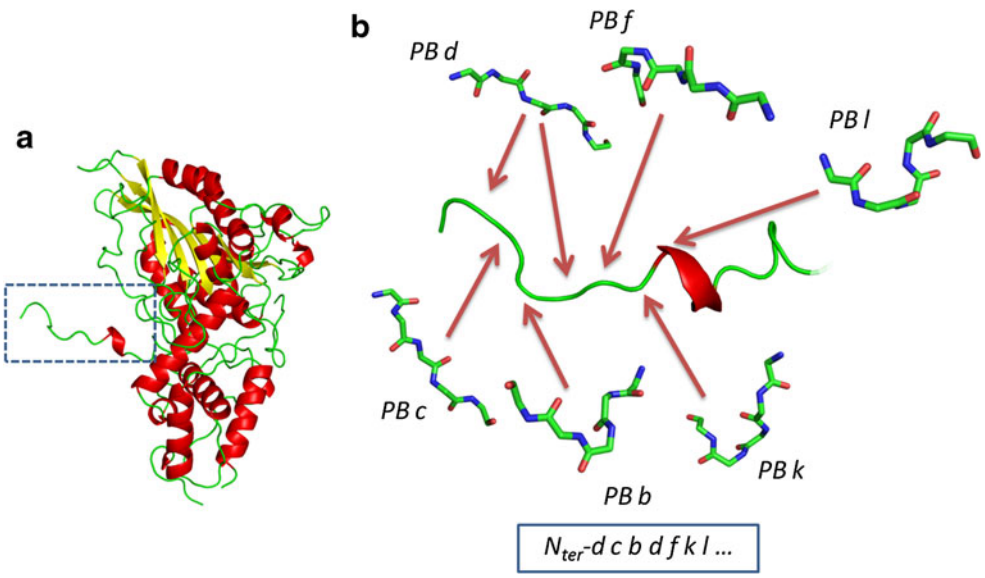

Fig. 1 Principle of encoding of protein structures using a "structural alphabet". The N-terminal extremity of chain B of *Aspergillus niger* acid phosphatase (Kostrewa et al. 1999) (**a**) is encoded in terms of a structural alphabet (**b**). Each residue is approximated by a specific prototype—here a protein block (*PB*). Hence, the crude description as a coil region (determined by any secondary structure assignment method) is replaced by a more precise series of PBs *dcbdfkl*

length were first coded in terms of the ϕ/φ dihedral angles, and then a root mean square deviation on angle (RMSDA) score was used to quantify the structural difference among the fragments (Schuchhardt et al. 1996). Using an unsupervised cluster analyzer related to self-organizing maps (SOM; Kohonen 1982, 2001), a three-step training process was carried out. The first step involved identifying the structural difference between fragments in terms of RMSDA; the second step took the transition probability (probability of transition from one fragment to another in a sequence) into consideration along with the RMSDA, i.e., in a similar way to the Markov model (Rabiner 1989). In the third step, the constraint based on transition probability was removed. Optimal prototypes were identified by considering both the structural approximation and the prediction rate. A set of 16 prototypes called protein blocks (PBs), represented as average dihedral vectors, was obtained at the end of this process (de Brevern et al. 2000).

These PBs are displayed represented in Fig. 2. PBs *m* and *d* can be described roughly as prototypes of the central α-helix and central β-strand, respectively. PBs *a* through *c* primarily represent β-strand N-caps and PBs *e* and *f*, β-strand 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 α-helix C-caps. This SA allows a good approximation of local protein 3D structures, with an average RMSD now evaluated at 0.42 Å (de Brevern 2005). PBs have been assigned using in-house software (available at http://www.dsimb.inserm.fr/DOWN/LECT/) or using PBE web server (http://bioinformatics.univ-reunion.fr/PBE/) (Tyagi et al. 2006b).

PBs (de Brevern et al. 2000) have been used both to describe the 3D protein backbones (de Brevern 2005) and to perform local structure prediction (de Brevern et al. 2000, 2007, 2002; Etchebest et al. 2005). Our earlier work on PBs revealed that PBs are effective in describing and predicting conformations of long fragments (Benros et al. 2006, 2009; Bornot et al. 2009; de Brevern and Hazout 2001, 2003; de Brevern et al. 2002, 2007;) and short loops (Fourrier et al. 2004; Tyagi et al. 2009a,b), analyzing protein contacts (Faure et al. 2008), in building a transmembrane protein (de Brevern 2005; de Brevern et al. 2009), and in defining a reduced amino acid alphabet to aid the design of mutations (Etchebest et al.

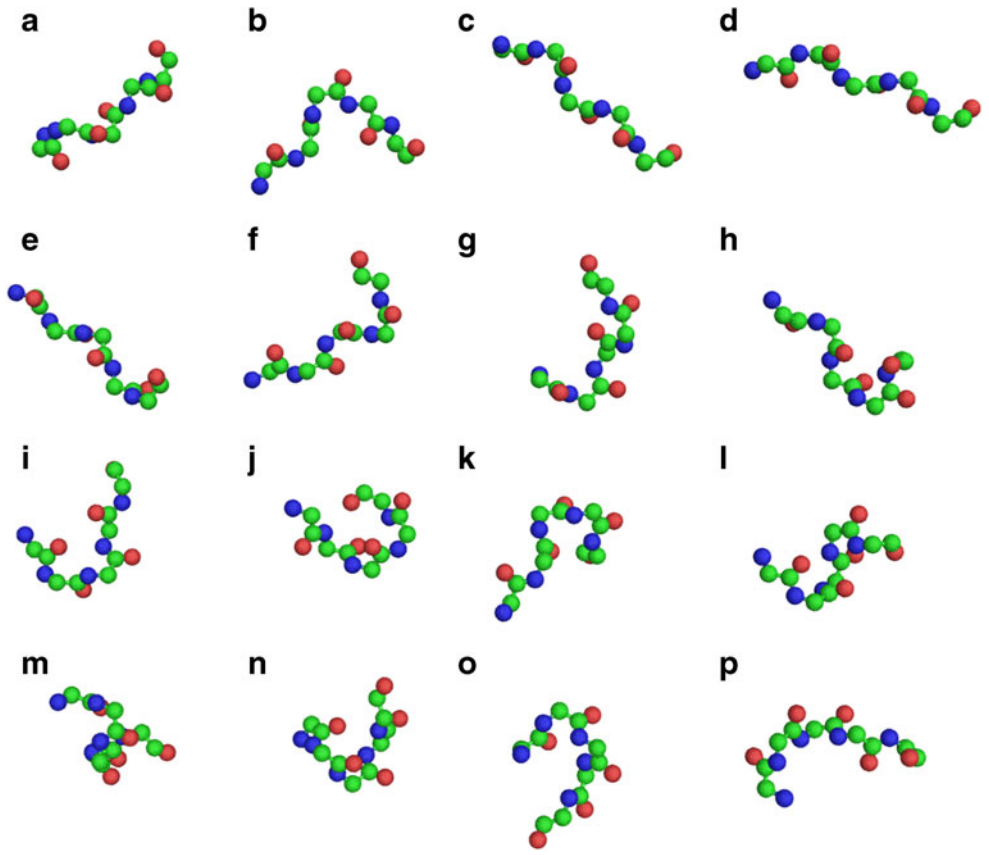

Fig. 2 The protein blocks. PBs from *a* to *p* are shown using PyMol software (DeLano 2002).. For each PB, the N -ap extremity is shown on the *left* and the C-cap on the *right*. Each prototype is five residues in length and corresponds to eight dihedral angles (ϕ,ψ). The PBs *m* and *d* are mainly associated to the central region of α-helix and the central region of β-strand, respectively (de Brevern 2005; de Brevern et al. 2000)

2007). This reduced amino acid alphabet was recently proved suitable for predicting protein families or sub-families and secretory proteins of *Plasmodium falciparum* (Zuo and Li 2009, 2010). We have also used PBs to superimpose and to compare protein structures (Tyagi et al. 2006a, b, 2008).

Other laboratories have taken advantage of PBs to reconstruct globular protein structures (Dong et al. 2007), design peptides (Thomas et al. 2006), and define binding site signatures (Dudev and Lim 2007). Novel prediction methodologies (Li et al. 2009; Rangwala et al. 2009; Zimmermann and Hansmann 2008) and fragment-based local statistical potentials (Li et al. 2009) have also been developed. The features of this alphabet have been compared by Karchin et al. (2003) with those of eight other SAs, revealing that our PB alphabet is highly informative, with the best predictive ability of those tested. Among the currently available SAs, it is the most widely at the present time.

Applications

Binding site signature

Protein blocks enable the detection of structural similarity between proteins with excellent efficiency. Dudev, Lim and co-workers (Dudev and Lim 2001; Yang et al. 2008) used this concept to locate structural motifs of metal/ligand-binding sites in proteins (Dudev and Lim 2007). These researchers encoded a protein structure databank in terms of PBs and subsequently located PBs encompassing specific metal-binding sites. They then defined a discontinuous PB pattern, similar to a PROSITE pattern. First, the structural motifs of the Cys_4 Zn-finger domains, which are known to adopt a specific structure, were analyzed. They then focused on structural motifs of the Mg^{2+}-binding sites in a set of non-redundant Mg^{2+}-binding proteins. Four Mg^{2+}-structural motifs were identified that showed important binding site relationships were identified, and then other features of the proteins were defined (Dudev and Lim 2007). This strategy can be easily extended to other cases. These researchers have recently extended the approach to DNA and RNA binding sites, highlighting a novel non-specific motif enabling diverse interactions with DNA and RNA as with proteins (Wu et al. 2010).

Definition of a reduced amino acid alphabet

The reduced amino acid alphabet is a popular concept that has been explored by many research teams. Indeed, the appropriate selection of an amino acid type in a reliable set is a particularly helpful approach to limit the number of experiments. Most of such approaches are mainly based on sequence properties (Akanuma et al. 2002; Clarke 1995; Kamtekar et al. 1993).

In this area, PBs not only help in describing protein structures, but they are also useful in extracting sequence–structure relationships. Based on this relation, we have proposed an association of amino acids in a limited number of clusters. This approach permits an exchange of amino acids that are equivalent in terms of sequence–structure relationship, while still maintaining local protein structure conformation (Etchebest et al. 2007). Zuo and Li used this reduced amino acid alphabet to predict different properties through a learning approach (Zuo and Li 2009, 2010).

Long structural fragments

Protein blocks are 5-residue-long fragments. To assess the structural stability of these short fragments, we identified the most frequent series of five consecutive PBs which are nine residues long. We then selected the 72 most frequent series and named them structural words (SWs). Interestingly, SWs encompass 92% of the residues (nearly 100% of the repetitive structures and 80% of secondary structure coil). By using most of the SWs, we were able to create a simple network describing most of the transitions between the SWs in proteins (de Brevern et al. 2002). The study of SWs yields a pertinent description of a large part of 3D structures, but as they constitute a sub-set of all five PBs combinations, they do not allow for a description of every part of the protein structure. We therefore have developed a novel approach named the Hybrid Protein Model (HPM; de Brevern and Hazout 2000). This innovative approach allowed us to create longer prototypes comprising 10–13 residues Benros et al. 2002, 2003, 2006, 2009; de Brevern and Hazout 2001, 2003). This resulted in a higher structural variability for the longer fragments through a significant increase in the number of prototypes, e.g., 100–130 prototypes (Benros et al. 2006, 2009; de Brevern and Hazout 2001, 2003). These longer fragments were used to perform structural superimposition (de Brevern and Hazout 2001), methodological optimization (Benros et al. 2003; de Brevern and Hazout 2003), and analyses of the sequence–structure relationship (Benros et al. 2006, 2009; Bornot et al. 2009; de Brevern and Hazout 2001). A modified version of HPM proposed by Serge Hazout has led to the construction of networks of local protein structures (Hazout 2005).

Structural alignment

The SA allows 3D protein structures to be translated into a series of letters (see Fig. 1a). Consequently, it is possible to use classical sequence alignment methodology to perform structure-based alignment (see Fig. 1b). The

main difficulty lies in obtaining a pertinent substitution matrix in order to find the similarity score between PBs for alignments. Using the homologs of known 3-D structures in the PALI database (Gowri et al. 2003) encoded in terms of PBs, it has been possible to compute a PB substitution matrix (Tyagi et al. 2006b). A dedicated webserver has been developed (http://bioinformatics.univ-reunion.fr/PBE/) that performs optimal alignments of a query protein structure with entries of 3-D structures in a database by using PBs and the substitution matrix (Tyagi et al. 2006a). A recent benchmark has proved that this method is most efficient in mining PDB and identifying proteins of a similar 3-D structure (Tyagi et al. 2008).

New developments in the field of protein structure have originated from this work. One of these relates directly to the use of the substitution matrix and concerns the characterization of conformational patterns in active and inactive forms of kinases. A comparison of closely related kinases revealed a higher global similarity between the active state kinases compared to the inactive states (as reflected from their PB scores) (Agarwal et al. 2010). The second axis focuses on the database, which is the basis for generating the substitution matrix, namely, the PALI database. The superimposed structures of PALI show regions with the correct alignments linked with regions more difficult to align (named variable regions). A novel optimization of the superposition based on PBs shows a global improvement in the variable regions. Hence, PBs improve PALI database alignments (Agarwal et al. in preparation). The last axis concerns the alignment approach. Even though the recent benchmark has demonstrated the quality of the methodology (Tyagi et al. 2008), some structural alignments still show poor consistency. Optimization of the substitution matrix and a novel alignment methodology have improved both the mining and the superimposition of protein structures (Joseph et al. in preparation). Figure 3 illustrates a very difficult case of superimposition of the 3D structures of *Aspergillus niger* acid phosphatase (Kostrewa et al. 1999) and *Escherichia coli* periplasmic glucose-1-phosphatase (Lee et al. 2003). In comparison, the superimposition obtained with our previous approach (Tyagi et al. 2006b) is quite poor. Our novel procedure allows the recognition of highly similar regions (shown on Fig. 3e) and provides a spectacular improvement in the superimposition. Figure 3c, d show that the bottom part of the protein structures can truly be superimposed. Figure 4 gives the alignment in terms of PBs.

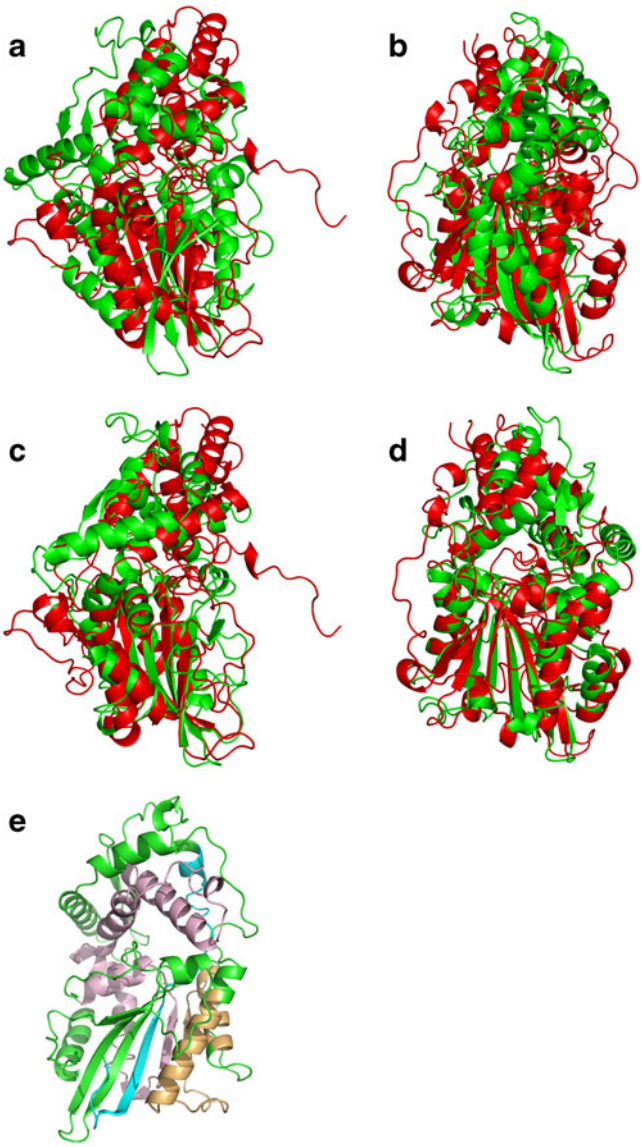

Fig. 3 An example of difficult superimposition of three-dimension (3D) protein structures using PBs. The 3D structure of the *Aspergillus niger* acid phosphatase (Kostrewa et al. 1999) chain B has been superimposed on the 3D structure of *Escherichia coli* periplasmic glucose-1-phosphatase chain A (Lee et al. 2003). **a**, **b** Superimposition using a previous approach (Tyagi et al. 2006b), **c**, **d** superimposition with the novel approach. Using regions of high similarity as seeds (*blue*, *gold* and *pink*), as shown in **e**, the root mean square deviation is 17 Å lower than the value previously computed

Prediction

As with secondary structure prediction, it is possible to predict local structures in terms of the SA (see Table 1 for a summary of all prediction approaches). Indeed, concomitant to an accurate description of the local 3D structure, the definition of PBs is driven by prediction capabilities (de Brevern et al. 2000). The prediction principle is based on Bayes' theorem. First, a set of protein chains used in training were encoded in terms of PBs using the minimal RMSDA criterion. Then, sequence windows of 15 residues length were considered for calculating the propensities associated with each PB. For each PB, the probability of

Fig. 4 An example of a hard case of superimposition of 3D protein structures using PBs. The 3D superimposition shown in Fig. 3 is presented here as the PB alignment. Repetitive PBs and direct N-cap and C-cap are shown in color

occurrence of an amino acid at each position in the sequence window was calculated and an occurrence matrix generated, i.e., 16 for the 16 PBs. Bayes theorem was applied to predict the structure of the new sequences. A prediction rate of 34.4% was achieved (de Brevern et al. 2000, 2004). Nonetheless, as only one amino acid occurrence matrix is associated to each PB, the sequence information is averaged. A clustering approach related to SOM (Kohonen 1982, 2001) performed on PBs sequences revealed well-defined sequence families for some PBs; an amino acid occurrence matrix was then computed for each sequence family. This strategy increased sequence specificities for some PBs and permitted an improved prediction rate of 40.7% to be achieved (de Brevern et al. 2000, 2004). Finally, a simulated annealing approach in the process of sequence family generation helped to improve the overall prediction to 48.7% (Etchebest et al. 2005). Importantly, this approach did not bring any biased or unbalanced improvements between the PBs. Combining the secondary structure information with the Bayesian prediction did not result in any significant improvement in the prediction rate. A website, LocPred (http://www.dsimb.inserm.fr/~debrevern/LOCPRED/), which includes most of the tools developed to date, is available to facilitate researchers in performing these predictions (de Brevern et al. 2004). Predictions were also performed with the SWs (de Brevern et al. 2002, 2007)

Table 1 Summary of prediction methods in terms of approaches, year of publication, prediction rate, and remarks

Approach	Year	Information	Prediction rate (%)	Reference	Web server	Remarks
Bayesian prediction	2000	One sequence	34.4	de Brevern et al. [10]	LocPred	First method
Sequence families	2000	One sequence	40.7	de Brevern et al. [10]	LocPred	Based on Bayesian prediction
Bayesian prediction	2002	One sequence	34.4	de Brevern et al. [16]	None	Prediction of structural words
Hidden Markov Model	2003	One sequence	None	Karchin et al. [34]	None	Fold recognition
Sequence families	2005	One sequence	48.7	Etchebest et al. [20]	LocPred	Improved sequence families
Pinning strategy	2007	One sequence	43.6	de Brevern et al. [18]	None	Prediction of structural words
knowledge-based prediction	2007	One sequence	62.0	Offmann et al. [42]	pb_prediction	Pentapeptide match/SCOP class
Two-layer support vector machine	2008	Evolutionnary	61.0	Zimmermann and Hansmann [57]	LOCUSTRA	First use of evolutionary information
Database-matching approach	2009	One sequence	45.3	Li et al. [40]	LSSRAP	Use also accessibility and secondary structure
svmPRAT	2009	Evolutionnary	68.9	Rangwala et al. [44]	svmPRAT	Protein residue annotation toolkit

and specifically for short loops (Fourrier et al. 2004; Tyagi et al. 2009a, b).

A knowledge-based approach for predicting local backbone structure was also developed. In this case, overlapping fragments of five residues from a query sequence are extracted and queried against a pentapeptide database. In this database, which was built from the SCOP database culled at 95% identity, each pentapeptide is mapped to a PB. In absence of any "hit" in the database, pentapeptides in which constraint of identity in the central position (position 3) is relaxed are considered. Overall performance of the approach was about 62%.

Recent developments have been made by other teams. Li and co-workers proposed an innovative approach for PB prediction that takes into account information from secondary structure and solvent accessibility (Li et al. 2009). Prediction rates were significantly improved (http://sg.ustc.edu.cn/lssrap/). Interestingly, this approach was found to be useful for fragment threading, pseudo-sequence design, and local structure predictions.

Support vector machine (SVM) methodology coupled with evolutionary information greatly improved the prediction rates. Hence, Zimmermann and Hansmann were able to develop a method for PB prediction using SVMs with a radial basis function kernel, leading to an improvement of the prediction rate to 60–61% (Zimmermann and Hansmann 2008). This method, called Locustra, is available online at http://www.fz-juelich.de/nic/cbb/service/service.php. In a very recent work, Rangwala, Kauffman, and Karypis developed a novel tool named svmPRAT (Rangwala, et al. 2009) that involves formulating the annotation problem as a classification or regression problem using SVMs (http://www.cs.gmu.edu/~mlbio/prosat/). An impressive increase in prediction rate of about 69% is achieved using such an approach. PB prediction is part of MONSTER (Minnesota prOteiN Sequence annoTation servER, http://bio.dtc.umn.edu/monster/). Thus, in less than a decade, the prediction rate of PBs has doubled in a very efficient way.

As emphasized by Li and co-workers (Li et al. 2009), it is often difficult to accurately compare the different studies because of the different definitions considered for local structures or different dataset and/or different criteria used for evaluating success predictions.

HPM strategy was used to construct a new library of local structures in order to extend the analyses to long structural fragments. As a result, 120 structural clusters (named local structure prototypes, LSPs) were then proposed to describe fragments that are 11 residues long (Benros et al. 2006). An original prediction method based on logistic regressions was first developed for predicting local structures from a single sequence. This method proposed a short list of the best structural candidates among the 120 LSPs of the library. A prediction rate of 51.2% was reached within the framework of a geometrical assessment. This result was quite significant, given the fragment length and the high number of classes (Benros et al. 2006). An improved prediction method based on SVMs and evolutionary information has recently been proposed. A global prediction rate of 63.1% was achieved and the prediction for 85% of the proteins was improved. This method has been shown to be among the most efficient of cutting-edge local structure prediction strategies (Bornot et al. 2009).

Conclusions and perspectives

Since 1989, nearly 12 different SAs have been developed (see, for example, Fetrow et al. 1997; Ku and Hu 2008; Sander et al. 2006; Unger et al. 1989; Unger and Sussman 1993; for dedicated reviews, see Joseph et al. 2010; Offmann et al. 2007). However, almost none of these have been used outside their the laboratories where they were developed. The PB alphabet is the only exception, and the reason for this rests with the ease with which protein 3D structures can be encoded as PBs. The PB alphabet can be considered as the classical standard of the SA, similar to DSSP being the classical standard for secondary structure assignment (Kabsch and Sander 1983).

PBs have been utilized in numerous different applications, such as the modeling of a transmembrane protein implicated in malarial infection (de Brevern 2005, 2009; de Brevern et al. 2009). It has also led to the development of an excellent superimposition method (Tyagi et al. 2008) and is now used by various research teams (Joseph et al. 2010). We have also developed confidence indexes associated to prediction accuracy (Bornot et al. 2009; de Brevern et al. 2000; Etchebest et al. 2005). We now link the uncertainties with the prediction of protein flexibility, looking at data from X-ray analysis, molecular dynamics (Bornot et al. submitted) and nuclear magnetic resonance. In the same way, superimposition approaches are constantly being improved. We have also examined protein–protein interactions from the standpoint of PBs (Swapna et al. in preparation).

Acknowledgments The authors would like to thank the reviewers for their comments that help improve the manuscript. The research was supported by grants from the French Ministry of Research, University of Paris Diderot–Paris 7, University of Saint-Denis de la Réunion, French National Institute for Blood Transfusion (INTS), French Institute for Health and Medical Research (INSERM), and Indian Department of Biotechnology. APJ and GA are supported by CEFIPRA number 3903-E and Council of Scientific and Industrial Research, respectively. AB had a grant from the French Ministry of Research, MT has a post-doctoral fellowship from NIH, and HV had a post-doctoral fellowship from CEA. NS and AdB acknowledge CEFIPRA for collaborative grant (number 3903-E). BS and AdB acknowledge Partenariat Hubert Curien Barrande (2010–2011). BS is supported by grant AV0Z50520701.

References

Agarwal G, Dinesh D, Srinivasan N and de Brevern AG (2010) Characterization of conformational patterns in active and inactive forms of kinases using Protein Blocks approach. In: Maulik U, Bandyopadhyay S, Wang J (eds) *Computational Intelligence and Pattern Analysis in Biological Informatics*. Wiley, in press

Akanuma S, Kigawa T, Yokoyama S (2002) Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA 99:13549–13553

Benros C, Hazout S, de Brevern AG (2002) Extension of a local backbone description using a structural alphabet. "Hybrid Protein Model": a new clustering approach for 3D local structures. In: International Workshop on Bioinformatics ISMIS. Lyon, pp 36-45

Benros C, de Brevern AG, 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. In: IEEE Workshop on Neural Networks for Signal Processing. IEEE Int Work 1:53–72

Benros C, de Brevern AG, Etchebest C, Hazout S (2006) Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62:865–880

Benros C, de Brevern AG, Hazout S (2009) Analyzing the sequence-structure relationship of a library of local structural prototypes. J Theor Biol 256:215–226

Bornot A, de Brevern AG (2006) Protein beta-turn assignments. Bioinformation 1:153–155

Bornot A, Etchebest C, de Brevern AG (2009) A new prediction strategy for long local protein structures using an original description. Proteins 76:570–587

Clarke ND (1995) Sequence 'minimization': exploring the sequence landscape with simplified sequences. Curr Opin Biotechnol 6:467–472

de Brevern AG (2005) New assessment of a structural alphabet. In Silico Biol 5:283–289

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

de Brevern AG, Hazout S (2001) Compacting local protein folds with a "hybrid protein model". Theor Chem Acc 106:36–47

de Brevern AG, Hazout S (2003) 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19:345–353

de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41:271–287

de Brevern AG, 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 AG, 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 AG, 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 AG, Autin L, Colin Y, Bertrand O, Etchebest C (2009) In silico studies on DARC. Infect Disord Drug Targets 9:289–303

DeLano WLT (2002) The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos. Available at: http://www.pymol.org.

Dong QW, Wang XL, Lin L (2007) Methods for optimizing the structure alphabet sequences of proteins. Comput Biol Med 37:1610–1616

Dudev T, Lim C (2001) Modeling Zn^{2+}-cysteinate complexes in proteins. J Phys Chem 105:10709–10714

Dudev M, Lim C (2007) Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinform 8:106

Etchebest C, Benros C, Hazout S, de Brevern AG (2005) A structural alphabet for local protein structures: improved prediction methods. Proteins 59:810–827

Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG (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 AG (2008) Protein contacts, inter-residue interactions and side-chain modelling. Biochimie 90:626–639

Fetrow JS, Palumbo MJ, Berg G (1997) Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. Proteins 27:249–271

Fourrier L, Benros C, de Brevern AG (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinform 5:58

Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res 31:486–488

Hazout S (2005) Une nouvelle méthode d'apprentissage: "Self-Learning by Information Share-Out" (SLISO). Sixièmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM) pour la génomique Lyon, pp 483-488

Joseph AP, Bornot A, de Brevern AG (2010) Local Structure Alphabets. In: Rangwala H, Karypis G. (eds), *Protein Structure Prediction*. Wiley, in press

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680–1685

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

Kohonen T (2001) Self-organizing maps, 3rd edn. Springer, Berlin Heidelberg New York

Kostrewa D, Wyss M, D'Arcy A, van Loon AP (1999) Crystal structure of *Aspergillus niger* pH 2.5 acid phosphatase at 2. 4 A resolution. J Mol Biol 288:965–974

Ku SY, Hu YJ (2008) Protein structure search and local structure characterization. BMC Bioinform 9:349

Lee DC, Cottrill MA, Forsberg CW, Jia Z (2003) Functional insights revealed by the crystal structures of Escherichia coli glucose-1-phosphatase. J Biol Chem 278:31412–31418

Li Q, Zhou C, Liu H (2009) Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. Proteins 74:820–836

Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5:17

Offmann B, Tyagi M, de Brevern AG (2007) Local protein structures. Curr Bioinf 3:165–202

Rabiner LR (1989) A tutorial on hidden Markov models and selected application in speech recognition. Proc IEEE 77:257–286

Rangwala H, Kauffman C, Karypis G (2009) svmPRAT: SVM-based protein residue annotation toolkit. BMC Bioinf 10:439

Sander O, Sommer I, Lengauer T (2006) Local protein structure prediction using discriminative models. BMC Bioinform 7:14

Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P (1996) Local structural motifs of protein backbones are classified by self-organizing neural networks. Protein Eng 9:833–842

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

Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, 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 CS, Cadet F, Srinivasan N, de Brevern AG, Offmann B (2006b) Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 34:W119–W123

Tyagi M, de Brevern AG, Srinivasan N, Offmann B (2008) Protein structure mining using a structural alphabet. Proteins 71:920–937

Tyagi M, Bornot A, Offmann B, de Brevern AG (2009a) Analysis of loop boundaries using different local structure assignment methods. Protein Sci 18:1869–1881

Tyagi M, Bornot A, Offmann B, de Brevern AG (2009b) Protein short loop prediction in terms of a structural alphabet. Comput Biol Chem 33:329–333

Unger R, Sussman JL (1993) The importance of short structural motifs in protein structure analysis. J Comput Aided Mol Des 7:457–472

Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5:355–373

Wu CY, Chen YC and Lim C (2010) A structural-alphabet-based strategy for finding structural motifs across protein families, Nucleic Acids Res. doi: 10.1093/nar/gkq478

Yang TY, Dudev T, Lim C (2008) Mononuclear versus binuclear metal-binding sites: metal-binding affinity and selectivity from PDB survey and DFT/CDM calculations. J Am Chem Soc 130:3844–3852

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

Zuo YC, Li QZ (2009) Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides 30:1788–1793

Zuo YC, Li QZ (2010) Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino Acids 38:859–67