

---

## SHORT COMMUNICATION

---

# A Substitution Matrix for Structural Alphabet Based on Structural Alignment of Homologous Proteins and its Applications

Manoj Tyagi,<sup>1</sup> Venkataraman S. Gowri,<sup>2</sup> Narayanaswamy Srinivasan,<sup>1,2</sup> Alexandre G. de Brevern,<sup>3</sup> and Bernard Offmann<sup>1\*</sup>

<sup>1</sup>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

<sup>2</sup>Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

<sup>3</sup>Equipe de Bioinformatique et Génétique Moléculaire, INSERM U726, Université Paris 7, case 7113, 2, place Jussieu, 75251 Paris Cedex 05, France

**ABSTRACT** Analysis of protein structures based on backbone structural patterns known as structural alphabets have been shown to be very useful. Among them, a set of 16 pentapeptide structural motifs known as protein blocks (PBs) has been identified and upon which backbone model of most protein structures can be built. PBs allows simplification of 3D space onto 1D space in the form of sequence of PBs. Here, for the first time, substitution probabilities of PBs in a large number of aligned homologous protein structures have been studied and are expressed as a simplified  $16 \times 16$  substitution matrix. The matrix was validated by benchmarking how well it can align sequences of PBs rather like amino acid alignment to identify structurally equivalent regions in closely or distantly related proteins using dynamic programming approach. The alignment results obtained are very comparable to well established structure comparison methods like DALI and STAMP. Other interesting applications of the matrix have been investigated. We first show that, in variable regions between two superimposed homologous proteins, one can distinguish between local conformational differences and rigid-body displacement of a conserved motif by comparing the PBs and their substitution scores. Second, we demonstrate, with the example of aspartic proteinases, that PBs can be efficiently used to detect the lobe/domain flexibility in the multidomain proteins. Lastly, using protein kinase as an example, we identify regions of conformational variations and rigid body movements in the enzyme as it is changed to the active state from an inactive state. *Proteins* 2006; 65:32–39. © 2006 Wiley-Liss, Inc.

**Key words:** local protein structures; substitution matrix; structural alphabet; structure alignment and comparison; rigid body shift

## INTRODUCTION

It has been realized since long time that known protein structures can be regenerated by assembling fragments from a repertoire of short structural motifs. Many of these short structural motifs reoccur in a large number of proteins of diverse structure and function.<sup>1–5</sup> Analysis of protein structures based on these short structural motifs has been widely used by various groups and have been shown to be useful in protein structure prediction,<sup>6–9</sup> reconstruction of backbone,<sup>10–13</sup> description and prediction of small loops<sup>14–16</sup> and long fragments.<sup>17–19</sup> Following these early leads, a set of 16 pentapeptide structural motifs have been identified<sup>20,21</sup> as a set of basic backbone structural patterns known as “protein blocks” (hereafter referred as PBs). PBs represent basic structural motifs upon which backbone model of most protein structures can be built. They have been found to be very informative<sup>7,22</sup> and useful for preprocessing before *ab initio* and new fold recognition. Interestingly, in a recent work<sup>23</sup> PBs has been used for protein 3D structure prediction. Each PB is characterized by a set of 8 ( $\phi$ ,  $\psi$ ) values and is represented by a character symbol  $a, b, c, \dots$  to  $p$  (refer Materials and Methods). A known

---

The Supplementary Material referred to in this article can be found at <http://www.interscience.wiley.com/jpages/0887-3585/suppmat/>

Grant sponsor: Conseil Régional de La Réunion.

\*Correspondence to: Bernard Offmann, 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. E-mail: [bernard.offmann@univ-reunion.fr](mailto:bernard.offmann@univ-reunion.fr)

Received 23 February 2006; Revised 1 May 2006; Accepted 8 May 2006

Published online 7 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21087

protein structure can be encoded into PBs by sliding an overlapping window of five residues along the backbone, and PBs for each window could be assigned on the basis of the smallest root mean square deviation on angular values<sup>24</sup> between the observed ( $\phi$ ,  $\psi$ ) values in the window and the standard torsion angle values for various PBs. Hence, 3D information of protein structure can be represented (simplified) into a 1D sequence of PBs.

It is now well documented that simplification of 3D space onto 1D space is an efficient tool to understand the sequence-structure relationship<sup>18,25</sup> and opens up new front for structure analysis of proteins namely structure comparison and alignment. Reduction of 3D space onto 1D space using local structural properties, to align and identify structurally equivalent regions, have been exploited earlier.<sup>26</sup> Flexible protein structure alignment and comparison methods, like SSAP<sup>27,28</sup> and DALI<sup>29</sup> based on combination of distance matrices and dynamic programming technique have been present for long time and shown to be effective in both global<sup>28</sup> and local<sup>30</sup> structure alignment. Similarly, pairwise protein structure alignment using orientation independent backbone representation with dynamic programming has shown lot of promise.<sup>31</sup> Analysis of protein structures in terms of sequences of structural alphabets (SAs) combined with alignment algorithm to find structural similarities have been reported recently in the form of a web service called SA-Search.<sup>32</sup>

Even though alignment of SAs using classical alignment algorithm has been tried out to some effective way, a genuine substitution matrix for SAs is required, similar to amino acid substitution matrix, to fully exploit the potential of such an approach. This requirement has been tentatively addressed in SA-Search where substitution scores were derived only from emission probabilities of hidden states in Hidden Markov Model.<sup>32</sup> This approach diverges from more classical methods.

Here, we derive a substitution table for PBs on the basis of assignment of PBs to structurally aligned homologous proteins. These proteins are present in a large database, PALI,<sup>33,34</sup> of homologous protein families with structure-based alignment available for every family. The  $16 \times 16$  PB substitution table provides extent of preference of a PB in a protein for its retention or substitution by any other PB in an aligned homologue. Usage of such methodology to extract substitution matrix provides a more rational approach over the matrix used in SA-Search, namely to evaluate equivalence between homologous structures. Among several possible uses of PB substitution table, we demonstrate, in this study, its application to the following problems:

- Aligning structures and identifying structurally equivalent regions between homologous or distantly related proteins using a dynamic programming approach.
- Distinguishing between conformational differences and rigid body shifts among homologous protein structures.
- Characterization of changes in structures between active and inactive states of enzymes, by taking protein kinases as an example.

## MATERIALS AND METHODS

### Protein Blocks

A structural alphabet that is able to both approximate 3D structure and predict process has been identified.<sup>20</sup> It is composed of 16 folding patterns of five consecutive residues, protein blocks (PBs), representing local structural features of proteins. Description of how PBs were identified has already been documented.<sup>20</sup> Each of the PBs is represented by a vector of eight dihedral angles associated with five consecutive C $\alpha$  atoms and the PBs are denoted by letters  $a, b, \dots, p$ . These PBs represent distinct and most common backbone conformations of pentapeptide regions in proteins of known structure. As can be seen from Table I, dihedral angle values for the PBs  $d$  and  $m$  correspond to the prototype for the central  $\beta$ -strand and the central  $\alpha$ -helix, respectively. PBs  $a$  through  $c$  primarily represent  $\beta$ -strand N-caps and  $e$  and  $f$ , C-caps. PBs  $g$  through  $j$  are specific to coils,  $k$  and  $l$  to  $\alpha$ -helix N-caps, and  $n$  through  $p$  to  $\alpha$ -helix C-caps.

To assign PB to a pentapeptide region in a protein structure, root mean square deviations on angular values or *rmsda*<sup>24</sup> between observed ( $\phi$ ,  $\psi$ ) values in the pentapeptide and ideal ( $\phi$ ,  $\psi$ ) values of each one of 16 PBs (Table I) are calculated. The assigned PB to the pentapeptide region corresponds to the one with lowest *rmsda*. In this manner, a 3D protein structure is translated into a 1D sequence of PBs representing structure information as sequence of structural alphabets.

### Database of Aligned Homologous Protein Structures

The database of Phylogeny and Alignment of homologous protein structures (PALI)<sup>33,34</sup> comprises of structure-based pairwise and multiple alignments of homologous proteins of known three-dimensional structure. The rigid-body superposition program STAMP<sup>35</sup> has been used for this purpose. The database also consists of phylogenetic tree structures of various protein families derived using sequence-based and structure-based similarity measures. The PALI database is available at <http://pauling.mbu.iisc.ernet.in/~pali>.

All the structurally aligned homologous proteins from PALI database were translated into alignment of PB sequences. As one PB represents five C $\alpha$  atoms, we have used a convention of associating the PB to the middle residue of the pentapeptide. Therefore, for protein of length  $N$ , the length of PB sequence is  $N-4$ .

The dataset used in the current analysis consists of 1197 protein families with 6140 protein structures involved in 21,503 pairwise alignments. The derivation of  $16 \times 16$  PB substitution matrix is aided by 2,071,225 observations of PB substitutions in the homologous protein structures.

### Calculation of Substitution Matrix

The number of substitutions between any two PBs is counted based only on the alignment corresponding to

**TABLE I. Ideal values of  $\phi$  and  $\Psi$  Dihedral Angles (in degrees) That Characterize the 16 Protein Blocks as Described by de Brevern et al. (2000)**

Protein Block	Dihedral angles							
	$\Psi_{n-2}$	$\phi_{n-1}$	$\Psi_{n-1}$	$\phi_n$	$\Psi_n$	$\phi_{n+1}$	$\Psi_{n+1}$	$\phi_{n+2}$
<i>a</i>	41.14	75.53	13.92	-99.80	131.88	-96.27	122.08	-99.68
<i>b</i>	108.24	-90.12	119.54	-92.21	-18.06	-128.93	147.04	-99.90
<i>c</i>	-11.61	-105.66	94.81	-106.09	133.56	-106.93	135.97	-100.63
<i>d</i>	141.98	-112.79	132.2	-114.79	140.11	-111.05	139.54	-103.16
<i>e</i>	133.25	-112.37	137.64	-108.13	133.00	-87.30	120.54	77.40
<i>f</i>	116.4	-105.53	129.32	-96.68	140.72	-74.19	-26.65	-94.51
<i>g</i>	0.40	-81.83	4.91	-100.59	85.50	-71.65	130.78	84.98
<i>h</i>	119.14	-102.58	130.83	-67.91	121.55	76.25	-2.95	-90.88
<i>i</i>	130.68	-56.92	119.26	77.85	10.42	-99.43	141.4	-98.01
<i>j</i>	114.32	-121.47	118.14	82.88	-150.05	-83.81	23.35	-85.82
<i>k</i>	117.16	-95.41	140.40	-59.35	-29.23	-72.39	-25.08	-76.16
<i>l</i>	139.20	-55.96	-32.7	-68.51	-26.09	-74.44	-22.60	-71.74
<i>m</i>	-39.62	-64.73	-39.52	-65.54	-38.88	-66.89	-37.76	-70.19
<i>n</i>	-35.34	-65.03	-38.12	-66.34	-29.51	-89.10	-2.91	77.90
<i>o</i>	-45.29	-67.44	-27.72	-87.27	5.13	77.49	30.71	-93.23
<i>p</i>	-27.09	-86.14	0.30	59.85	21.51	-96.30	132.30	-92.91

structurally conserved regions identified by STAMP superimposition. This caution is exercised, as the alignment of residues in the structurally variable regions is meaningless in the rigid body alignments. The raw frequencies are normalized with respect to the total number of two PBs in question as well as with respect to the total number of PB–PB substitutions in the dataset. These normalized frequencies are then expressed as the log-odds scores as follows<sup>36</sup>:

$$S_{ij} = \log_e \left[ \frac{N_{ij} / \sum_{j=1}^{16} N_{ij}}{\sum_{i=1}^{16} N_{ij} / \sum_{i=1}^{16} \sum_{j=1}^{16} N_{ij}} \right]$$

where  $N_{ij}$  is the raw frequency of replacing PB  $i$  with PB  $j$ .

### Data Set Used for Validation of PB Matrix

Validation of the substitution matrix was performed by using structural alignments to identify local equivalent regions. The initial goal was to benchmark our method (PB-ALIGN) using a comprehensive set of protein domain pairs. A total of 29 pairs was used in this evaluation, among which, 14 pairs were taken from Shindyalov and Bourne<sup>37</sup> and the other 15 pairs were taken randomly within SCOP families with a sequence identity cutoff of 40%. The complete list is provided in the supplementary data.

## RESULTS

The presented work is based on the concept of translating structurally aligned homologous proteins into aligned sequences of 16 types of PBs and calculation of PB substitution frequency to obtain a normalized  $16 \times 16$  matrix. This matrix gives a score to substitute

a given PB into another in topologically equivalent regions.

### Substitution Matrix

Analysis of pairs of proteins from PALI database was used to construct a PB substitution matrix. Table II provides the final substitution matrix, which is used extensively in the work described in this manuscript. It can be noted that most of the off-diagonal elements are negative suggesting that conformation changes of the pentapeptides in the homologous protein structures are occurring. The following PBs pairs *c-a*, *f-e*, *g-a*, *g-c*, *g-e*, *h-e*, *i-a*, *j-b*, *j-h*, *j-i*, *k-h*, *n-g*, *o-h*, *p-b*, *p-g*, *p-i*, *p-j*, *b-h*, *b-i*, *b-l*, *c-d*, *j-k*, *k-l*, *l-o*, *n-o*, *o-p*, and *p-n* are the off-diagonal elements with positive substitution scores i.e. favorable substitutions. Figure 1(a,b) show examples of *g-a* and *e-f* substitutions respectively and these PB pairs show similarity in their structures in the middle of the segments. Figure 1(c) gives an example of negative score substitution indicating differences in backbone structure. In total, there are 43 pairs with positive score including 16 diagonal elements.

Interestingly, the diagonal elements *m-m* and *d-d* substitutions that correspond to central  $\alpha$ -helix and central  $\beta$ -strands are not biased by their corresponding high frequencies owing to the normalization formula used.<sup>36</sup> Low frequencies among other PB substitutions and their good conservation, especially involving N and C caps residues of helices and strands, results in high scores in other diagonal elements.

### Application of the Substitution Matrix to Identify Structurally Equivalent Regions

One of the most convenient ways to validate the substitution matrix is to use it for protein structural alignment and to compare the results with those obtained

**TABLE II. Normalized Substitution Frequencies Expressed as Log-Odds Scores Between Any Two Protein Blocks as Determined by Structure-Based Pairwise Alignments of Homologous Proteins of Known Three-Dimensional Structure From PALI Database**

Protein blocks	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>	<i>j</i>	<i>k</i>	<i>l</i>	<i>m</i>	<i>n</i>	<i>o</i>	<i>p</i>
<i>a</i>	2.28															
<i>b</i>	-0.12	2.49														
<i>c</i>	0.54	-0.21	1.69													
<i>d</i>	-0.29	-0.44	0.17	1.35												
<i>e</i>	-1.59	-0.48	-1.10	-0.36	3.05											
<i>f</i>	-0.54	-1.53	-0.39	-0.49	0.75	2.21										
<i>g</i>	0.31	-0.73	0.18	-1.29	1.37	-0.33	3.25									
<i>h</i>	-1.14	0.20	-1.63	-1.20	0.66	-0.34	-0.74	3.07								
<i>i</i>	0.39	0.24	-1.11	-1.12	-1.15	-1.07	-0.19	-0.92	3.37							
<i>j</i>	-1.15	0.32	-1.03	-0.92	-0.76	-0.34	-0.51	1.18	1.54	3.74						
<i>k</i>	-1.75	-0.03	-2.45	-2.63	-0.38	-0.04	-1.39	0.51	-0.15	0.07	2.52					
<i>l</i>	-0.60	0.04	-2.21	-1.56	-1.76	-0.33	-0.74	-0.36	-0.22	-0.12	0.19	2.24				
<i>M</i>	-2.40	-2.98	-2.70	-5.20	-4.75	-2.14	-1.10	-2.93	-3.15	-2.00	-1.02	-0.68	1.06			
<i>n</i>	-1.40	-0.83	-1.68	-3.07	-0.58	-1.99	1.07	-1.07	-0.97	-0.44	-0.56	-0.27	-0.77	3.65		
<i>o</i>	-0.54	-0.55	-0.65	-2.66	-2.48	-1.41	-0.01	0.96	-0.89	-0.48	-1.71	0.06	-1.26	0.26	3.36	
<i>p</i>	-0.36	0.33	-0.01	-2.10	-2.22	-1.91	0.47	-1.81	1.32	0.60	-1.35	-1.23	-1.10	0.36	0.24	2.83

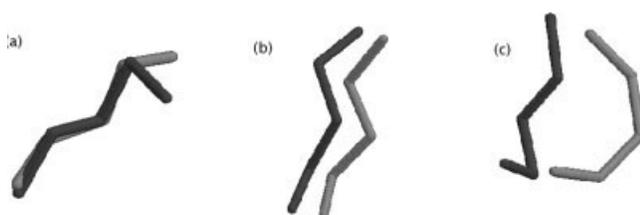


Fig. 1. Backbone comparison of PBs. (a) Superimposed backbone structures of PB *a* (black) and *g* (grey), where matrix gives positive score for substitution between *a* and *g*. (b) Backbone structure of PB *e* (black) and *f* (grey) having positive substitution score. (c) Backbone structure of PB *a* (black) and *j* (grey) having negative substitution score.

with other well-established methods. Alignment of protein structure in terms of PBs using the substitution table and a dynamic programming approach (hereafter called PB-ALIGN) was benchmarked against the standard structural alignment methods implemented in DALI and STAMP.

When aligning two structures, PBs are assigned to the two proteins in consideration. Then, using dynamic programming approach, sequences of PBs from two proteins are optimally matched rather like amino acid sequence alignments. To quantify the extent of substitution between PBs the newly generated PB substitution matrix is used.

To identify a gap penalty with optimal performance a large number of PB alignments were generated with a number of gap penalties in the range of  $-1$  to  $+1$ . Alignment quality was estimated by analyzing the correspondence between known equivalent regions and between known nonequivalent regions. From manual analysis of these alignments we found that positive penalty often produced potentially erroneous alignments, whereas negative penalty performed better in aligning equivalent regions. We fixed empirically the gap penalty to  $-0.5$ , which often resulted in reasonable alignments. Analysis

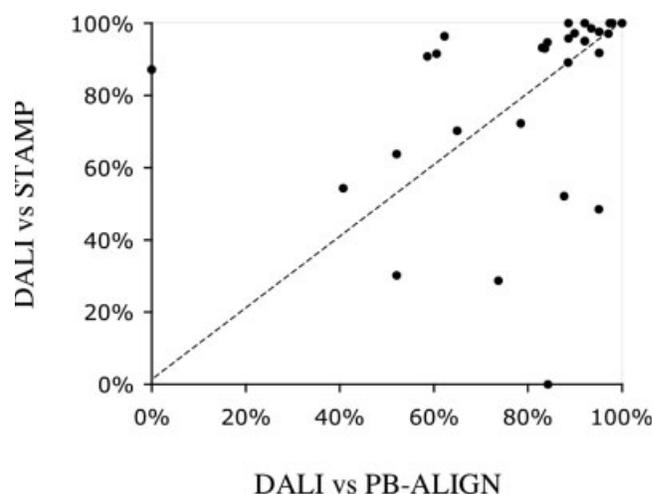


Fig. 2. Comparison of PB-ALIGN, STAMP, and DALI. Comparison of structural alignment against DALI using STAMP (*y*-axis) and PB-ALIGN (*x*-axis). Each axis represents the percentage of aligned positions that are in agreement with alignment from DALI.

of the resulting alignment provides a direct way to identify structures which are equivalent or variable.

A comprehensive data set that consists of 15 protein pairs belonging to homologous families were sampled following documented test cases,<sup>37</sup> and a further 14 other protein pairs were sampled from SCOP families with a 40% sequence identity cutoff (see complete list in the online supplementary material). Structural alignments were performed using DALI, STAMP, and PB-ALIGN.

We analyzed the extent of overlapping of structurally equivalent regions by comparing (i) DALI vs. STAMP and (ii) DALI vs. PB-ALIGN. We simply counted the number of positions within each aligned region, which were in agreement with the reference alignment from DALI.

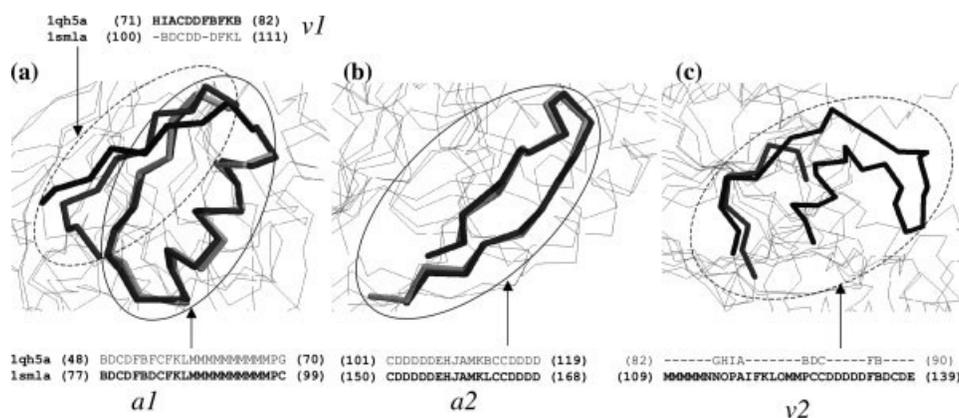


Fig. 3. Identification of structurally equivalent regions between 1QH5 and 1SML. Superposed structures of 1QH5 and 1SML are used here to highlight structurally equivalent and variable regions identified by PB-based alignment. (a) The regions from the two proteins circled with a solid line are identified as equivalent according to the alignment using PBs shown at the bottom of the figure. The C terminal extension of this region is shown in another circled region (dashed line). (b) The region *a2* of the two proteins that is also identified, using PBs, as equivalent are shown in black for 1QH5a (region 101–119) and in grey for 1SMLa (region 150–168). (c) Identification of variable regions by PB alignment for 1QH5 (region 82–90 in grey) and 1SML (region 109–139 shown in black). STAMP superposition indeed shows that these two regions are structurally nonequivalent. Files are provided for this figure (see supplementary material for details).

Interestingly, PB-ALIGN was able to align as much as 75.9% of the positions that are aligned in DALI, and this is comparable with the performance of STAMP which scores 77.6%. Results for each of the pairwise alignments of homologous proteins from DALI and PB-ALIGN show that in more than two-third of the cases, at least 80% of the aligned positions are in common between the two methods (Fig. 2). First, this demonstrates that structural alignment based on the use of our PB substitution matrix and dynamic programming is giving reasonably comparable results when compared to standard rigid or flexible structure alignment methods. Second, this shows that PB-ALIGN can be used as a fast method to identify structural equivalences in homologous proteins.

Attempts have been made to use PB substitution scores to locate local portions in distantly related proteins that are structurally equivalent. In the first instance, we compared two distantly related  $\alpha$  and  $\beta$  protein ( $\alpha + \beta$ ) class based on SCOP classification<sup>38</sup> from the metallohydrolase superfamily (1QH5a and 1SMLa) sharing 18% sequence identity. The lengths of these two chains are almost the same (260 and 266, respectively).

An alignment of the PB sequences from these two metallohydrolases was calculated as described earlier. Several regions of interest in this alignment are detailed in Figure 3(a–c). Highlighted are four distinct PB alignment stretches. According to our PB based alignment featured at the bottom of Figure 3(a,b), these regions (*a1* and *a2*) would be structurally equivalent, while the other two PB alignments featured in Figure 3(a,c) would correspond to structurally variable regions (*v1* and *v2*). Superposed coordinates of 1QH5 and 1SML from STAMP are used only to highlight structural equivalent and variable regions identified by dynamic programming based alignment of PBs. Indeed, *a1* and *a2* regions circled with a solid line in Figure

3(a,b) are well superimposed. On the other hand, structures in the *v1* and *v2* regions circled with a dashed line in Figure 3(a,c), as expected, are not well superimposed and they are not considered as structurally equivalent in the PB alignment. Simple comparison of the results of our approach with that of the SSAP method<sup>28</sup> shows interesting results. SSAP method was successful in identifying region *a1* but fails to identify structurally equivalent region *a2*. Also compared to SSAP based alignment of region *v2* (not shown), the structurally variable region is more evident from PB alignment, because of poor PB scores in the region. In the case of DALI region *a1* was well identified as equivalent but surprisingly there was no demarcation of region *v2* from *a2* as variable. In addition, the C-terminal extension (*v1*) adjacent to the *a1* region displays a region that is almost equivalent between the two proteins but it appears that there are subtle conformational changes identified from the PB alignment [Fig. 3(a)]. When examining this C-terminal extension in the STAMP superimposition, slight differences in backbone conformation are indeed observed [Fig. 3(a)]. Once again this subtle change is not directly evident from SSAP or DALI based structure alignments.

In the second instance, two distantly related proteins (1BNKa and 1FMTb) from the all  $\beta$  class FMT C-terminal domain like superfamily were studied. The lengths of these two chains are very different (200 and 108, respectively). We analyzed three different local regions from the PB alignment [Fig. 4(a–c)]. Examination of the first region [Fig. 4(a)] indicates that they would be structurally equivalent regions but the 1BNK fragment is shorter than the equivalent 1FMT fragment. This is illustrated in the superimposition of the two structures. Indeed, N-cap and furthest C-cap region are well superimposed while the central helix, which is significantly longer in the 1FMT structure, is only poorly superim-

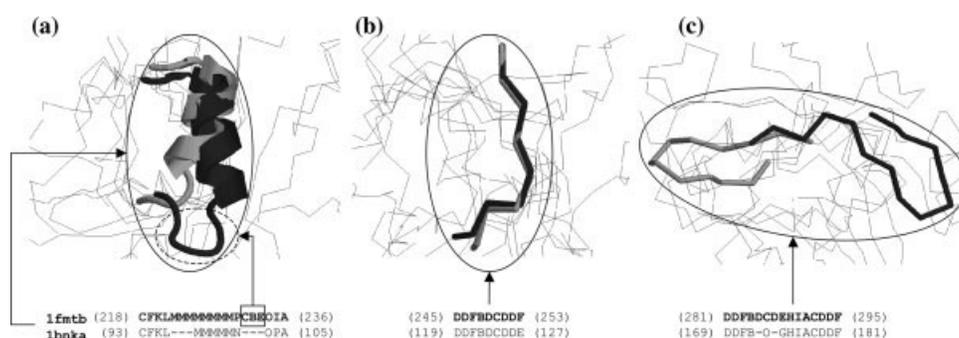


Fig. 4. Identification of structurally equivalent regions between 1FMT and 1BNK. Superposed structures of 1FMT and 1BNK are used here to highlight structural equivalent and conformationally variable regions identified by dynamic programming based alignment of PBs. (a) Aligned helical region from 1FMTb (218–236 region shown in black) and 1BNKa (region 93–105 shown in grey) indicating longer helical and loop region in 1FMTb as identified by PB alignment. Circled region with dashed line shows the presence of extra loop in C termini of 1FMTb (region 231–233), which corresponds to the extra “CBE” PB-motif in the alignment. (b) The highlighted region from the two proteins is identified as equivalent from PB alignment and is shown in black for protein 1FMTb (region 245–253) and in grey for protein 1BNKa (region 119–127). (c) Similar structural regions of 1FMTb (region 281–295 shown in black) and 1BNKa (region 169–181 shown in grey) but which are not superimposable by rigid body superposition. Regions discussed above uses original residue numbering as given in PDB coordinate files. Files are provided for this figure (see supplementary material for details).

posed. Figure 4(a) also shows the presence of extra loop region in 1FMT as identified by the “CBE” PB-motif in the PB alignment. Examination of the second region [Fig. 4(b)] from the two proteins shows that they are structurally equivalent and are indeed perfectly superimposed. Interestingly, the third region [Fig. 4(c)] is predicted as structurally equivalent by PB alignment with positive PB substitution scores. However, the  $C\alpha$  positions are not superimposable and share no equivalent residues. Close examination of this particular region shows that the backbones are almost identical but poor superimposition is due to a rigid body shift. This result thus indicates that despite absence of good superimposition, PB alignment is able, in a flexible manner, to detect structurally equivalent regions in proteins.

### Distinguishing Conformational Differences and Rigid-Body Shifts in Homologous Proteins

When two homologous structures are aligned using rigid body superposition, high  $C\alpha$ - $C\alpha$  deviations can result either due to conformational differences between the aligned regions or due to differences in the spatial positioning of identical conformational motifs. For example, alignment of a  $\alpha$ -helical region and  $3_{10}$  helical region in two homologues correspond to conformational difference. On the other hand, the difference in the relative position of two conserved  $\alpha$ -helical regions in the two homologues corresponds to rigid-body shift. Both these changes can result in high  $C\alpha$ - $C\alpha$  deviations.

The basic premise in distinguishing between conformational difference and rigid-body shift is that conformational differences are characterized by high root mean square deviations (RMSD) of pentapeptide regions and poor PB substitution scores. Difference in spatial orientation of structurally conserved segments is charac-

terized by high  $C\alpha$ - $C\alpha$  deviations in the pentapeptides, but, good (favored) PB substitution scores. Thus, simple transformation of superimposed structures into PB alignment provides a novel and direct way to rapidly detect these two situations.

Detailed description of how PBs can be used to address this issue is documented in the supplementary material of this manuscript. Three examples are provided; (i) relative orientation of C-terminal lobe with respect to N-terminal lobe in endothiasepsins (E.C. 3.4.23.6); (ii) rigid body shifts and conformational changes in two distantly related proteins of class II aminoacyl tRNA synthetase N domain; and (iii) structural alterations between active and inactive states of cyclic AMP dependent protein kinase.

## DISCUSSION

Arriving at a meaningful measurement of the probabilities for short structural motifs to change conformation in topologically equivalent regions is only possible if local backbone of a set of structurally aligned proteins is decoded in terms of a structural alphabet. This issue is, to our knowledge, addressed for the first time here using PBs.<sup>20</sup> Interestingly, because of the methodology used to construct the matrix, direct evaluation of equivalences between homologous structures is possible which is not the case for the HMM based matrix derived from SA-Search.<sup>32</sup> The derived substitution matrix here suggests that perceivable conformational changes occur even in topologically equivalent regions of homologous proteins. This is indicated by negative scores of most of the off diagonal elements (Table II) despite considering topologically equivalent regions in rigid body superimposition from PALI database.

Because PBs allow encoding of protein 3D structures into 1D sequences, these can, interestingly, be manipulated rather like amino acids sequences. This approach has been explored here, namely for structure comparison and alignment. For PB alignment to be relevant, the availability of a biologically meaningful PB substitution matrix was a prerequisite. This was ensured by the methodology used for its construction. The matrix derived required further validation in terms of its performance to align pairs of protein structures in comparison to well-established methods. Aligning protein structures by aligning PB sequences using dynamic programming is different from aligning secondary structural elements because PBs describe more precisely the backbone conformation in coil regions and N or C caps of regular helices or strands.<sup>18</sup> Hence it is expected to be more efficient than a  $3 \times 3$  matrix. Alignment, using PBs, is achieved here in a flexible manner and performs comparably to other robust flexible structural alignment methods like in DALI<sup>29</sup> or SSAP.<sup>28</sup> However, it is noteworthy that the actual implementation of PB alignment is expected to fail in detecting domain swapping situations. The originality in our approach resides in the methodology used, which, besides being very intuitive, is very different from those implemented in DALI<sup>29</sup> or SSAP.<sup>28</sup>

Importantly, these two methods are being routinely used for structure comparison on a large-scale basis via web services. Concurrently, application of PB substitution matrix in protein fold recognition is expected to be a useful venture. This has recently been tested on a large-scale basis. We showed that the efficiency rate to mine similar fold proteins from SCOP using 1D representation as sequence of PBs varies from 86.1 to 93.6%,<sup>39</sup> thus further validating our approach and substitution matrix. A web service that implements this approach<sup>39</sup> is available at <http://bioinformatics.univ-reunion.fr/PBE/>.

## CONCLUSIONS

In this study, we demonstrated, using a structural alphabet, the usefulness of encoding 3D structure into 1D space through the use of a substitution matrix.

Such a substitution table is shown to be useful in distinguishing conformational changes and rigid-body shifts of structural motifs in homologous proteins. Its application is also demonstrated in terms of characterization of structural differences involving rigid body movements between active and inactive forms of an enzyme.

Using 1D representation of 3D structure combined with our substitution matrix and simple dynamic programming, we were, even in the case of two homologues with very different sequence lengths, able to locate regions of structural similarity, highlight subtle change in conformations within aligned regions, and to identify regions of no structural similarity. Though robust methods such as DALI<sup>29</sup> and SSAP<sup>28</sup> are quite sensitive and effective in rapid detection of common folds and structural motifs, the applications presented in this study clearly highlights the original and informative nature of

the derived  $16 \times 16$  substitution matrix and gives good indication of its strength in protein structure analysis.

This work has important implications in comparative modeling of loops. Besides, it can be used to add meaning to nonsuperimposed (variable) regions, in databases of structurally aligned proteins like PALI, by finding structural equivalence in these regions.

As an extension of our work, we are working towards arriving at gross flexible global and local structural alignment of distantly related proteins with optimized gap penalties. At present we are testing batteries of gap penalties on comprehensive data set by analyzing the effect of penalty on overall quality of alignment and efficiency rate to nine similar structure proteins. It will lead us to define an optimal gap penalty which will provide balance between structure mining efficiency and overall alignment quality. Similarly, general consideration on the rationale for constructing such a matrix is presently being addressed like amino acid substitution matrices<sup>40</sup> namely in the field of PBs compositional bias. We are also currently investigating the distribution properties of raw PB alignment scores against randomized datasets in both local and global alignment schemes to arrive at a genuine statistical measurement of alignment significance.

## ACKNOWLEDGMENTS

We thank Mr. C. Sairam Swamy for providing the code for dynamic programming and Rajesh Thangudu for his suggestions on the work. NS is an International Senior Fellow of the Wellcome Trust, London. He thanks the authorities of Reunion University for the visiting professorship in the laboratory of BO. We thank anonymous reviewers for fruitful suggestions.

## REFERENCES

1. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. *EMBO J* 1986;5:819–822.
2. Levitt M. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* 1992;226:507–533.
3. Unger R, Sussman JL. The importance of short structural motifs in protein structure analysis. *J Comput Aided Mol Des* 1993;7:457–472.
4. Han KF, Baker D. Recurring local sequence motifs in proteins. *J Mol Biol* 1995;251:176–187.
5. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 1989;5:355–373.
6. Bystrhoff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 1998;281:565–577.
7. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* 2003;51:504–514.
8. Hunter CG, Subramaniam S. Protein local structure prediction from sequence. *Proteins* 2003;50:572–579.
9. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: improved prediction methods. *Proteins* 2005;59:810–827.
10. Camproux AC, Gautier R, Tuffery P. A hidden Markov model derived structural alphabet for proteins. *J Mol Biol* 2004;339:591–605.

11. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* 2002;323:297–307.
12. Tuffery P, Guyon F, Derreumaux P. Improved greedy algorithm for protein structure reconstruction. *J Comput Chem* 2005;26:506–513.
13. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol* 1995;249:493–507.
14. Fourier L, Benros C, de Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* 2004;5:58.
15. Camproux AC, Brevern AG, Hazout S, Tuffery P. Exploring the use of a structural alphabet for structural prediction of protein loops. *Theor Chem Acc* 2001;106:28–35.
16. Camproux AC, Tuffery P, Buffat L, Andre C, Boisvieux JF, Hazout S. Analyzing patterns between regular secondary structures using short structural building blocks defined by a hidden Markov model. *Theor Chem Acc* 1999;101:33–40.
17. de Brevern AG, Hazout S. 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics* 2003;19:345–353.
18. de Brevern AG, Valadie H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Sci* 2002;11:2871–2886.
19. de Brevern AG, Hazout S. Compacting local protein folds with a "hybrid protein model". *Theor Chem Acc* 2001;106:36–47.
20. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 2000;41:271–287.
21. de Brevern AG. New assessment of a structural alphabet. *In Silico Biol* 2005;5:26.
22. Karchin R. Evaluating local structure alphabets for protein structure prediction, PhD Thesis, Computer Science Department, University of California, 2003.
23. de Brevern AG, Wong H, Tournamille C, Colin Y, Le Van Kim C, Etchebest C. A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochim Biophys Acta* 2005;1724:288–306.
24. Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng* 1996;9:833–842.
25. Rost B. Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem Anal* 2003;44:559–587.
26. Lo Conte L, Smith TF. Visible volume: a robust measure for protein structure characterization. *J Mol Biol* 1997;273:338–348.
27. Taylor WR, Orengo CA. A holistic approach to protein structure alignment. *Protein Eng* 1989;2:505–519.
28. Taylor WR, Orengo CA. Protein structure alignment. *J Mol Biol* 1989;208:1–22.
29. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993;233:123–138.
30. Orengo CA, Taylor WR. A local alignment method for protein structure motifs. *J Mol Biol* 1993;233:488–497.
31. Ye J, Janardan R, Liu S. Pairwise protein structure alignment based on an orientation-independent backbone representation. *J Bioinform Comput Biol* 2004;2:699–717.
32. Guyon F, Camproux AC, Hochez J, Tuffery P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res* 2004;32(Web Server issue):W545–W548.
33. Balaji S, Sujatha S, Kumar SS, Srinivasan N. PALI-a database of Phylogeny and Alignment of homologous protein structures. *Nucleic Acids Res* 2001;29:61–65.
34. Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res* 2003;31:486–488.
35. Russell RB, Barton GJ. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol* 1994;244:332–350.
36. Johnson MS, Overington JP. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol* 1993;233:716–738.
37. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998;11:739–747.
38. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
39. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, Brevern AG, Offmann B. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucl Acids Res*, in press.
40. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK. Protein database searches using compositionally adjusted substitution matrices. *FEBS J* 2005;272:5101–5109.