

## **β-bulges: Extensive structural analyses of β-sheets irregularities**

Pierrick Craveur<sup>1,2,3,4,\*</sup>, Agnel Praveen Joseph<sup>5</sup>,  
Joseph Rebehmed<sup>1,2,3,4</sup> & Alexandre G. de Brevern<sup>1,2,3,4,\*</sup>

<sup>1</sup> INSERM, U665, DSIMB, F-75739 Paris, France.

<sup>2</sup> Univ. Paris Diderot, Sorbonne Paris Cité, UMR\_S 665, F-75739 Paris, France

<sup>3</sup> Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.

<sup>4</sup> Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.

<sup>5</sup> National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560065, India.

\* Corresponding authors: de Brevern Alexandre G. & Craveur Pierrick

mailing address: INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot, Sorbonne Paris Cité 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France

E-mail: [alexandre.debrevern@univ-paris-diderot.fr](mailto:alexandre.debrevern@univ-paris-diderot.fr) and [pierrick.craveur@inserm.fr](mailto:pierrick.craveur@inserm.fr)

Tel: +33(1) 44 49 30 38 / 30 73

Fax: +33(1) 47 34 74 31

*Running title:* β-bulges conservation

The manuscript contains 25 pages with 4 figures, 6 tables and 12 supplementary materials (8 tables and 4 figures corresponding to 16 pages).

## **Abstract**

β-sheets are quite frequent in protein structures and are stabilized by regular main-chain hydrogen bond patterns. Irregularities in β-sheets, named β-bulges, are distorted regions between two consecutive hydrogen bonds. They disrupt the classical alternation of side chain direction and can alter the directionality of β-strands. They are implicated in protein-protein interactions and are also introduced to avoid β-strand aggregation. Five different types of β-bulges are defined.

Previous studies on β-bulges were performed on a limited number of protein structures or one specific family. These studies evoked a potential conservation during evolution. In this work, we analyze the β-bulge distribution and conservation in terms of local backbone conformations and amino acid composition. Our dataset consists of 66 times more β-bulges than the last systematic study (Chan *et al.* Protein Science 1993, 2:1574-1590). Novel amino acid preferences are underlined and local structure conformations are highlighted by the use of a structural alphabet. We observed that β-bulges are preferably localized at the N- and C-termini of β-strands, but contrary to the earlier studies, no significant conservation of β-bulges was observed among structural homologues. Displacement of β-bulges along the sequence was also investigated by Molecular Dynamics simulations.

Key words: beta-sheets; beta-strand; structural irregularity; evolution; structural alphabet;

Protein Blocks; folds; structural comparison; protein structure; mining.

*50-75-word statement, written for a broader audience, outlining the importance and/or impact of the work presented in the manuscript:*

β-bulges are irregularities in hydrogen bond patterns of β-sheets; they bring in large structural effects and consequently the protein functions. Previous studies report β-bulge as a conserved sequence and structural motif. We performed a systematic assessment on a large protein databank to determine whether similar protein folds have conserved β-bulges. Contrary to the literature, no significant conservation of β-bulges was observed among structural homologues.

**Abbreviations and symbols**

AB : Anti-parallel Bent

AC : Anti-parallel Classic

AG : Anti-parallel G1

AS : Anti-parallel Special

ASTRAL : database for sequence and analysis

AW : Anti-parallel Wide

DSSP : Database of Secondary Structure Assignment

GDT\_TS : Global Distance Test Total Score

iPBA : improved Protein Block Alignment

LINCS : LINear Constraint Solver

MD : Molecular Dynamic

NMR : Nuclear Magnetic Resonance

PB : Parallel Bent

PB(s) : Protein Block(s)

PC : Parallel Classic

PW : Parallel Wide

PS : Parallel Special

RMSD : Root Mean Square Deviation

SCOP : Structural Classification Of Proteins

**Introduction**

Protein 3D structures are often described as a succession of repetitive secondary structures (1-2): (i)  $\alpha$ -helix (or 3.6<sub>13</sub> helix) characterized by intramolecular hydrogen bonds between amino acid residues  $i$  and  $i+4$  and (ii)  $\beta$ -sheet composed of extended chains with hydrogen bonds between adjacent chains. A major difference between these two main regular secondary structures is the non-local nature of hydrogen bonds. In case of  $\beta$ -sheet, hydrogen bonding partners can be far from each other in the sequence (3). Helical structures represent 1/3<sup>rd</sup> of residues while extended structures account for 1/5<sup>th</sup> (4-5). Depending on the strand orientation, a  $\beta$ -sheet can be parallel, anti-parallel or mixed, resulting in different hydrogen-bonded patterns (6). The planarity of  $\beta$ -sheet arrangement results in a periodicity in the side-chain orientation, pointing alternatively on both sides of the sheet. The sequence specificity of  $\beta$ -strands and their capping regions has been widely analyzed (7)-(8). Prediction of  $\beta$ -sheets structure is difficult (9-11) as  $\beta$ -sheet assembly is more complex than simple pair complementarities (6; 12-13).

Like the helices, which are often non-linear structures (14), the  $\beta$ -sheets also show irregularities, named  $\beta$ -bulges (15). A bulge is formed when extra residues are inserted between successive hydrogen bonds stabilizing the  $\beta$ -sheet, so that usually two or more residues are on one strand opposite a single residue on the other (which is named X) (16-17). They are mainly observed between anti-parallel  $\beta$ -strands and more than two bulges per protein are found on an average (16). Richardson and co-workers were the first to analyze these conformations identifying 91  $\beta$ -bulges in 28 proteins and proposed the first classification into 3 types: Classic, G1 and Wide. Milner-White analyzed the relation between G1  $\beta$ -bulge and the  $\beta$ -hairpins describing a particular G1  $\beta$ -bulge loop, related to turns (17-19). Thornton and co-worker made the first systematic study of  $\beta$ -bulges on a non-redundant dataset. They analyzed 362  $\beta$ -bulges and extended the earlier classification based on the

conformation and hydrogen-bond patterns. Consequently, they introduced two new classes: the Bent and the Special  $\beta$ -bulges.

**[Table I]**

$\beta$ -bulges are thus grouped into 5 classes according to backbone conformation and hydrogen-bonding patterns (see Table I for more details): (i) the Classic  $\beta$ -bulge occurs between a narrow pair of hydrogen bonds; (ii) the G1  $\beta$ -bulge occurs only between anti parallel  $\beta$ -strands; (iii) the Wide  $\beta$ -bulge occurs between the widely spaced pairs of hydrogen bonds; (iv) the Bent  $\beta$ -bulge, which do not have residue X, have the residues 1 and 2 inserted in each  $\beta$ -strand. and (v) the Special  $\beta$ -bulges that are formed when more than two residues are inserted in the bulged strand (16). The term “ $\beta$ -bulges residues” in the article refers to all residues which compose the  $\beta$ -bulges (at positions X, 1, 2, 3 and 4) unless otherwise indicated.

The extra residues do not only disrupt the normal alternation of side chain direction, but also have an effect on the directionality of  $\beta$ -strands; they tend to accentuate the typical right-handed twist of  $\beta$ -sheets. Hence,  $\beta$ -bulges are expected to be well conserved in proteins and a good example of this was reported for SH3 domain protein, where a  $\beta$ -bulge was found to be highly conserved (20).

$\beta$ -sheets are the least exposed and most rigid secondary structure regions (21). As  $\beta$ -bulges are more exposed than other strand residues, they seem to play a key role in protein-protein interaction (22), protein folding (23-24) and other functions (22; 25). It was also suggested that they can be associated with pathologies like neurodegenerative disorders resulted from protein aggregations (26) and kidney deficiencies (27). Nonetheless, most of these studies used a limited set of protein structures or a particular family (28).

The objective of this work is to study the distribution and the conservation of  $\beta$ -bulges in terms of types, local backbone conformation and amino acid composition, in proteins. For this purpose, we superimposed structurally similar proteins using iPBA tool. This methodology is based on the use of Protein Blocks (PBs), which is a structural alphabet composed of 16 prototypes that are 5 residues long and these prototypes were able to approximate a complete protein backbone conformation. This library of local conformations is currently the most widely used structural alphabet (29). iPBA uses the translation of protein structures as series of PBs, which can be compared using sequence alignment techniques. iPBA outperformed other established methods on a non-trivial benchmark dataset (30-31). Based on the superimposed structures, the distribution and conservation of  $\beta$ -bulges in terms of their types, local backbone conformation and amino acid composition were studied.

## **Results**

### **[Table II]**

*Analysis of the secondary structures.* 12,132 structures, representing 2,180,241 amino-acids, were used for this study, out of 16,712 structures in the SCOP dataset filtered at 95% sequence identity. The remaining protein chains comprise structures solved by Nuclear Magnetic Resonance, involve non-standard PDB file formats and those structures for which PROMOTIF failed to assign backbone conformations. Table II summarizes the secondary structure assignment for the SCOP95 dataset. Three structural classes *i.e.*,  $\alpha/\beta$ ,  $\alpha+\beta$  and all- $\beta$  represent a quarter of our dataset each, while all- $\alpha$  represents only 16.8% of the protein chains. Secondary structure assignment resulted in 35.1% of residues in  $\alpha$ -helical conformation, 18.5% in  $\beta$ -sheets and rest 46.4% in coils. Similar results were found for

SCOP40 dataset, and the secondary structure distributions are in agreement with previous studies (4; 32-33).

**[Table III]**

***β-bulge distribution.*** β-bulges represent 3.35% of the residues in the dataset (24,142 β-bulges representing 73,096 residues, see Table II). They are as recurrent as the PolyProline helix II conformation (34) or the frequent types of β-turns (35). Chan and coworkers (16) analyzed only 362 β-bulges from 170 proteins. In this work, the dataset used is 66 times bigger, nevertheless a similar distribution of β-bulges was found (see Table S1). Antiparallel β-bulges represent 92.3% of all β-bulges; parallel ones being less frequent (7.7%). On an average, each protein has two β-bulges, while the average occurrence reaches three for the all-β proteins.

As expected, the average number of β-bulges per protein is proportional to the length and the number of β-strands. More than 90.1% of β-bulges were found in structures which are shorter than 400 residues and having at least 3 β-strands.

Classic β-bulges are the majority with a frequency of 57.0%, followed by G1 β-bulges (32.8%), the other three types being rarer (Wide β-bulges: 4.9%, Special β-bulges: 3.3% and Bent β-bulges: 1.9%). A significant change compared to the previous study (16), is an increase in the frequency of antiparallel classic β-bulges from 46.7 to 52.6% and a decrease in antiparallel Wide β-bulges, from 8.3 to 3.7%.

PROMOTIF assigns β-bulges according to their conformation and hydrogen-bonding patterns. Hence, these characteristics could also apply to residues localized in a loop, as noticed for G1 β-bulge loop (17). β-bulges are observed with residues completely localized inside β-sheets, partly outside or completely outside the sheet.



As shown in Table III only 54.3% of  $\beta$ -bulges are composed entirely of residues localized inside the  $\beta$ -strands. The rest is mainly represented by antiparallel G1  $\beta$ -bulges with 98.1% having at least one residue localized outside  $\beta$ -strands, which often corresponds to G1  $\beta$ -bulge loop(17).

$\beta$ -bulges are usually localized close to  $\beta$ -strand extremities, *i.e.*, 90% of  $\beta$ -bulges are within 3 residues of a  $\beta$ -strand extremity (see Table S2). If we consider the  $\beta$ -bulges that are composed of residues lying exclusively in  $\beta$ -strands, the bulge residues are rarely localized in the middle of  $\beta$ -strands; this behavior can be observed for all types of  $\beta$ -bulges (see Figure S1). 75.35% of  $\beta$ -bulge residues are localized either within a strand or between two consecutive strands (see Table S3).

***$\beta$ -bulge in SCOP classes.*** As seen in Tables II and S4, the distribution of  $\beta$ -bulges is not similar in all SCOP classes.  $\beta$ -bulges were even found in the all- $\alpha$  class which by definition has a low  $\beta$ -sheet content. 30.7% of these  $\beta$ -bulges are entirely found inside  $\beta$ -sheets and are mainly antiparallel G1  $\beta$ -bulges (54.9%). As  $\alpha/\beta$  class is mainly composed of parallel  $\beta$ -sheets, it is expected to have the highest content of parallel Special, Wide, Bent and Classic  $\beta$ -bulges (3.4%, 5.0%, 2.9%, and 20.2% respectively).  $\alpha+\beta$  and all- $\beta$  classes exhibit roughly the same behavior with the dominance of antiparallel Classic  $\beta$ -bulges (60.5 and 55.2%, respectively), a significant representation of antiparallel G1  $\beta$ -bulges (31.0 and 35.0%, respectively) and a limited number of  $\beta$ -bulges outside  $\beta$ -strand (~15%), like  $\alpha/\beta$  class.

The multidomain protein and small protein classes have similar distributions with approximately 30% of  $\beta$ -bulges in  $\beta$ -strands, 30% outside  $\beta$ -strands and 38% are partly in  $\beta$ -strands. The membrane associated class has lower number of  $\beta$ -bulges, but has the highest number of antiparallel Wide  $\beta$ -bulge (9.3%, which is twice the frequency in the other classes); the other 5 types of  $\beta$ -bulges were never observed.

**[Table IV]**

***Amino-acid preferences of  $\beta$ -bulge.*** Table IV shows the amino acid and the Protein Block preferences of the different  $\beta$ -bulges (see also Table S5). Aspartate is found over-represented, mainly at positions 1 and 2. For all  $\beta$ -bulge types, a significant preference for the amino acids Glycine, Asparagine and Proline is seen. Fewer amino acid preferences were found in all parallel  $\beta$ -bulges, and in anti-parallel Bent  $\beta$ -bulges.

Previous work on  $\beta$ -bulges mainly focuses on three  $\beta$ -bulge types, i.e. antiparallel Classic, G1 and antiparallel Wide [see Table 4, 5 and 6 of Chan's study(16)]. About 85, 69 and 30 times more  $\beta$ -bulges of each type are identified in this work leading to a new characterization (see bold and underlined labels in Table IV). For the antiparallel Classic  $\beta$ -bulge, four new over-represented residues are observed, Leucine at position X, Asparagine and Lysine at position 1 and Arginine at position 2. On the other hand, three over-represented residues reported earlier, had lesser preference in this study, Tyrosine at position X and Valine, Glutamate and Serine at position 1 (see Table S5). In the case of antiparallel G1  $\beta$ -bulges, striking differences observed were (a) lesser preferences for Histidine and Serine at position X, (b) over-representation of Aspartate at position 1 and (c) enrichment of three residues at position 2, i.e., Serine, Threonine and Asparagine, which are otherwise common in all  $\beta$ -bulge types.

Finally, the antiparallel Wide  $\beta$ -bulges showed preference for Glycine & Asparagine at positions X and 1, and Serine & Tryptophan at position X. Position 2 had two new over-represented amino acids, namely Glycine and Proline, apart from Aspartate and Asparagine. Another major difference is the absence of preference for Glutamine.

***Local conformation preferences of  $\beta$ -bulge.*** We also investigated the preference for local backbone conformations associated with  $\beta$ -bulges. Protein Blocks is currently the most widely

used structural alphabet (29). It gives a finer description of local backbone conformations when compared to the classical secondary structures (36). PBs *b* and *i*, are found strongly over-represented (21 and 14 fold each) in  $\beta$ -bulges. PB *b* is found at the N-cap of  $\beta$ -strands and *i* is frequent in loop region (37). These two PBs are found in successive positions and strongly mark the irregularity in the  $\beta$ -strand. These PB conformations are not easily altered, as seen in the PB substitution matrix developed in our earlier work (30-31). Interestingly, both parallel (minority) and anti-parallel (majority) types favored similar Protein Blocks are found overrepresented for both antiparallel (majority) and parallel (minority) types. Residue X is mainly in  $\beta$ -strand conformation reflected by the preferences for PB *d*, which correspond to the central region of classical  $\beta$ -strand, PBs *b* and *c* corresponding to the N cap, and PB *f* to the C-cap. In our previous studies, we have shown that certain preferential transitions are observed between PBs (36-37), *i.e.*, PBs (letters) have preferred successions (words). In G1  $\beta$ -bulge PBs *i* or *p*, at position 1, are seen to be followed by PB *a*, at position 2.

Antiparallel Classic  $\beta$ -bulges were characterized by distinct preferences for amino acids. They also have distinct PBs' patterns with PBs *d* or *h* favored at position X. PB *h* is mainly associated with residues following the end of a  $\beta$ -strand while PB *a* at position 1 and PBs *g* and *j* at position 2 are mainly loop associated PBs. AC  $\beta$ -bulges are mainly found within  $\beta$ -strands (~75% see Table III) with a preference to strand termini, an assignment in agreement with the observed PBs, mainly associated to loops.

***Protein superimpositions*** Analysis of  $\beta$ -bulges in specific protein families, *e.g.* the WD40 family (28) and the immunoglobulin family (16), has suggested that  $\beta$ -bulges could be more conserved than other parts of protein structures. 950,793 structure superimpositions were carried out using iPBA program. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could just arise from the

physico-chemical properties of proteins favoring certain packing arrangements and chain topologies. The average GDT\_TS score is 33.25 with a peak at 31 (see Figure S2). Even though superimpositions were performed at the level of SCOP fold, a non negligible proportion of structural alignments shares a very low GDT\_TS, *i.e.*, some structures, classified in same SCOP fold cannot be properly superimposed. Hence, we selected only superimpositions with GDT\_TS score better than 15, a threshold already used in a previous study (38); corresponding to an average RMSD lower than 2.69Å (see Figure S3); reflecting superimpositions of structures sharing similar global conformation. Consequently, 716,346 superimpositions were selected.

Similarly, a strong correlation exists between GDT\_TS and sequence identity, as seen between GDT\_TS and RMSD (see Figure S3). The GDT\_TS and the sequence identity have an exponential correlation. As the sequence is less conserved than the structures, we observed that a GDT\_TS score of 60 corresponds to a sequence identity of about 35% on an average. We observed 1,665,200 non-superimposed β-bulges and 531,567 aligned β-bulges (full and partial). As certain folds contained more experimental protein structures, the distribution of β-bulges was normalized.

#### **[Tables V and VI]**

*Are the β-bulges conserved among homologous structures?* On an average, a β-bulge has 42% chances of being conserved (superimposed: 30% fully and 12% partially). In majority of the cases, the two β-bulges superimposed share the same type and secondary structure localization. A β-bulge has 33.0% probability to be conserved with the same type, 31.4% with the same secondary structure localization, and 27.4% with both (see Table V). Conservation of β-bulge with change of types (9%) was observed 5 times more for partial than for full

superimposition (probability of 7.5% and 1.5% respectively). This result emphasizes that changes in the hydrogen bond pattern of  $\beta$ -bulge goes hand in hand with modification of the local structural conformation. . It must be noticed that it often changed the  $\beta$ -bulge type and the superimposition in such case can only be partial.

In the following sections, the residues which form the bulge (positions 1 to 4) are named the bulged residues. A  $\beta$ -bulge has only 13.2% probability to be structurally conserved with the same X-residue, and only 5.1% chance with same bulged residues. For different  $\beta$ -bulge types (except Bent) at equivalent positions, the probabilities to be structurally conserved with the same X-residue decreases to 0.71%, and down to 0.11% for the same bulged residues. Hence, the amino acid composition of a conserved  $\beta$ -bulge in homologous structures is not necessarily similar. However, conserved  $\beta$ -bulges with same bulged amino acid sequence have 91% probability to share the same  $\beta$ -bulge type. Figure 1(a) shows the conservation of bulged residues in regards to the whole sequence. It highlights without any doubt that the  $\beta$ -bulges are not more conserved (on average) than any other regions of the proteins. As expected (Figure 1(b)), sequence identity variation is correlated with GDT\_TS, underlining no specific generic constraints specific to  $\beta$ -bulge. This result is not in contradiction with specific studies that underline the conservation of  $\beta$ -bulge, as we see better structural conservation of  $\beta$ -bulge for alignments with higher GDT\_TS score and a higher sequence identity. The probability to observe a conserved  $\beta$ -bulge in homologous structures increases from 38.48% (for sequence identity lower than 35%) to 75.98% (for higher rate), and from 29.81% (for GDT\_TS between 15 and 40) to 63.29% (for higher GDT\_TS, see Table VI).

***Stability of  $\beta$ -bulges*** To understand the structural significance and stability of  $\beta$ -bulges, we studied a particular case of  $\alpha$ -lytic protease (PDB code 1SSX (39)) using Molecular Dynamics (MD) simulations. This protein is 198 residues long and is characterized by a large number of

about 15  $\beta$ -bulges [see Fig. 2(b)]. Three different temperatures were applied during the MD simulations to analyze the stability of  $\beta$ -bulges. 34 different  $\beta$ -bulges have been observed in the different simulations (see Table S6). The 15  $\beta$ -bulges remain stable and are present at least 90% of the times for the three temperatures. Nonetheless, formation and disappearance of  $\beta$ -bulges (see Figure S4) and changes in their types were also observed. It is mainly due to significant local conformational changes that lead to the loss or gain of hydrogen bonds. Some  $\beta$ -bulges were quite transient (seen less than 29% of the simulation time) at all temperatures. These transient  $\beta$ -bulges are mainly composed of residues which are also observed in stable  $\beta$ -bulges. This could be due to the local structural environment that alters the protein flexibility resulting in the  $\beta$ -bulge shift. For example, the two  $\beta$ -bulges found at position 44-51-52 and 45-50-51 correspond to a stable and transient  $\beta$ -bulge respectively, with overlapping positions in the sequence. Interestingly these putative displacements are accompanied by a change of  $\beta$ -bulge type.

## **Discussion**

This study is based on the analysis of a larger dataset of  $\beta$ -bulges compared to previous works. We provide a new description of amino acid preferences [see Table IV(a) and Table S5(a)] associated with  $\beta$ -bulges. Various studies on specific protein families have shown the importance of conserved  $\beta$ -bulges (40-42). Higher conservation of  $\beta$ -bulges was observed with the increasing degree of homology in terms of both sequence and structure.

However our results highlight the observation that  $\beta$ -bulges in structurally similar proteins are not necessarily conserved. We found that a  $\beta$ -bulge has only 42% probability to be conserved in structures sharing the same fold. We also showed that different types of  $\beta$ -bulges are conserved among similar structures and transition between  $\beta$ -bulge types is more frequent

than expected (around 1 in 10  $\beta$ -bulges). Nonetheless, we do not observe any significant preferences for certain type to be conserved [see Fig. 1(a) and Table S7]. For instance, different numbers of  $\beta$ -bulges were observed in two structures corresponding to the matrix protein VP40 of Ebola virus (43-44) [see Fig. 2(a) and 2(b)]. These structures (SCOP domain: d1es6a1 and d1h2ca\_) share the same amino acid sequence and are very close in terms of global conformation (GDT\_TS = 46.62 and RMSD = 2.06 Å). 3  $\beta$ -bulges were found conserved between both structures and one was missing in one structure [see Fig. 2(a)]. This observation highlights the possible effects of protein flexibility or crystal packing on  $\beta$ -bulges formation.

Finally, this work highlights that the conservation of  $\beta$ -bulges is not significantly influenced by conservation of the fold, contrary to the speculations in previous studies. However higher is the homology in terms of sequence and structure, higher is the probability to find conserved  $\beta$ -bulges. Molecular dynamic studies on bulged protein allow the observation of stable and transient nature of  $\beta$ -bulges. This behavior needs to be investigated at a greater detail on related proteins to quantify its impact on conservation studies.

## **Materials and Methods**

*Overview of the method.* Figure 3 shows the workflow for the analysis of  $\beta$ -bulges. Protein structures were taken from the ASTRAL SCOP dataset (45) [Fig. 3(a)]. Secondary structures, including  $\beta$ -bulges were assigned by PROMOTIF software (46) [Fig. 3(b)]. Protein structural domains containing at least one  $\beta$ -bulge were superimposed with protein structural domains belonging to the same SCOP folds. Pairwise structural alignments were performed using iPBA (30) [Fig. 3(c)]. The conservation of  $\beta$ -bulges at structurally equivalent positions were then analyzed [Fig. 3(d)].

**Structural datasets.** Two sets of protein structures were extracted from Protein DataBank (47) based on the ASTRAL SCOP dataset (45), filtered at 40% and 95% sequence identity. The proteins were classified into folds and classes based on the SCOP classification (48). All NMR structures were excluded from the analysis. SCOP95 dataset contained 16,712 structures representing 1,195 folds and 7 classes.

**Analysis of local backbone conformation.** Secondary structures have been assigned using PROMOTIF software (46); it is based on DSSP methodology (49) and used backbone hydrogen bond patterns. PROMOTIF also gives assignments of different types of turns and  $\beta$ -bulges. The assignment of  $\beta$ -bulges is based on Chan *et al.* classification (16) that defines 5 main types Classic (C), Bent (B), Wide (W), G1 and Special (S). PROMOTIF is used for the secondary structure assignment in PDBsum (50) and it is the only currently available software that allows distinguishing and assigning the different types of  $\beta$ -bulges.

Protein Blocks (PBs(37)) were also used to have a finer and different view of the local backbone conformation. They correspond to a set of 16 pentapeptide conformations, labeled from *a* to *p*, described as a series of ( $\phi, \psi$ ) dihedral angles(29; 51) (see Table S8). This library was obtained by clustering all pentapeptide conformations using an unsupervised classifier similar to Kohonen Maps (52-53) and Hidden Markov Models (54). The PBs *m* and *d* can be roughly described as prototypes for the central region of  $\alpha$ -helix and  $\beta$ -strand, respectively. PBs *a*, *b* and *c* primarily represent the N-cap of  $\beta$ -strand while *e* and *f* correspond to C-caps. PBs *g* to *j* are specific to coils. *k* and *l* correspond to N cap of  $\alpha$ -helix while PBs *n* to *p* are associated with C-caps. This structural alphabet of 16 prototypes allows a reasonable approximation of local protein 3D structures (37) with an average root mean square deviation (RMSD) of about 0.42 Å (36).



**Structure superimposition.** Abstraction of structures in terms of PBs helps to encode 3D information into a 1D sequence (29; 37; 51). We used classical amino acid sequence alignment strategies to align PBs sequences used to compare protein structures (55-57). The alignment approach was refined with the use of an anchor-based dynamic programming algorithm which first identifies all high scoring and structurally favorable local alignments (anchors). The segments between these anchors are then aligned to obtain a global alignment. This improved PB based structure alignment approach, namely iPBA, outperformed other established methods as seen with different robust benchmark datasets (30-31). ProFit (version 3.1) (58) is used to obtain the final 3D superimposition of two protein structures (based on the PB-based sequence alignment). ProFit performs least squares fit of protein structures based on the residue equivalences in a given sequence alignment. Only C $\alpha$  atoms of equivalent residues were used for the calculations.

iPBA provides two measures of the structural superimposition: the RMSD, and the Global Distance Test Total Score (GDT\_TS defined by Zemla in 2003 (59)). This latter varying between 0 and 100, and reflects the global similarity of two protein structures. It was used for model assessment in the last rounds of Critical Assessment of Techniques for Protein Structure Prediction (CASP (60)).

**Sequence alignment.** Amino acid sequence alignment was performed using CLUSTAL-W (version 2.1)(61). Default parameters were used with Gonnet substitution matrix(62).

**Characterization of  $\beta$ -bulge superimposition.** Proteins classified into the same fold were superimposed with iPBA. The structure based sequence alignment generated by iPBA was used to evaluate the structural conservation of  $\beta$ -bulges. Based on this alignment output, we can distinguish three cases of  $\beta$ -bulges superimposition (Fig. 4):

1. *Full Superimposition*: All the residues of first  $\beta$ -bulge are aligned with those in the second  $\beta$ -bulge. [see Fig. 4(a)], the number of aligned residues correspond to the number of residues of the shorter  $\beta$ -bulge.
2. *Partial Superimposition*: only some residues of the smaller  $\beta$ -bulge are aligned with the residues of the other  $\beta$ -bulge [see Fig. 4(b)].
3. *Non Superimposed  $\beta$ -bulges*: A  $\beta$ -bulge found in one protein structure is not aligned with a  $\beta$ -bulge on the other protein [see Fig. 4(c)].

**Amino acid and PB composition.** Occurrence of each amino acid and Protein Block in a particular  $\beta$  bulge type, have been normalized into a Z-score:

$$Z(n_{i,j}) = \frac{n_{i,j}^{obs} - n_{i,j}^{th}}{\sqrt{n_{i,j}^{th}}}$$

with  $n_{i,j}^{obs}$  the observed occurrence number of amino acid or PB  $i$  in position  $j$  (residue position X, 1, 2, 3, or 4) for a given particular  $\beta$ -bulge type and  $n_{i,j}^{th}$ , the expected number.

The expected values correspond to the product of the occurrence in position  $j$  and the frequency of amino acid  $i$  in the entire databank (or from  $\beta$ -strands of the databank). Positive Z-scores correspond to over-represented amino acids or PBs, and the threshold values of 4.42 and 1.96 were chosen to indicate the level of significance (p-value less than  $10^{-5}$  and  $5 \cdot 10^{-2}$  respectively). This measure was used in our previous studies to analyze the amino acid representation in Protein Blocks (37; 63).

**Molecular simulations.** Molecular dynamics (MD) simulations were performed with GROMACS 4.5.4 (64-67) using Amber 03 force field(68) for proteins and the explicit TIP3P solvent model for water molecules was used (69). The structure was immersed in a water box with periodic boundary conditions and neutralized with  $\text{Na}^+$  or  $\text{Cl}^-$  counterions. The energy of

each system was then minimized with a steepest-descent algorithm for 2000 steps. MD simulation was performed in NPT ensemble, with temperature and pressure kept constant, at three different temperatures (298, 310 and 353 K) and 1 bar pressure using Berendsen algorithm (67). The coupling time constants were  $\tau=0.1$  ps and  $\tau=4$  ps for temperature and pressure respectively. Bond lengths were constrained with the LINCS algorithm (70), which allowed an integration step of 2fs. The Particle-mesh Ewald summation (71) was used to handle long-range electrostatic interactions using a cut-off of 1.4 nm for non-bonded interactions. An equilibration step was first performed for 500 ps, with protein atom positions constrained while ions and water molecules were free to move, followed by an unrestrained production step of 50 ns. The coordinates were recorded at every 10 ps interval. The MD simulation was checked and analyzed using Gromacs tools.

### **Electronic supplementary material**

The supplementary material could be found in the file named “Supplementary\_Material.doc”.

It is composed by 8 tables and 4 figures which correspond to 16 pages.

The used data for this study are available under the following url:

<http://www.dsimb.inserm.fr/~craveur/Beta-Bulge/>

### **Funding**

These works were supported by grants from the Ministry of Research (France), University Paris Diderot, Sorbonne Paris Cité, National Institute for Blood Transfusion (INTS, France), and Laboratoire of Excellence GR-Ex by the program “Investissements d’avenir”, (France), Institute for Health and Medical Research (INSERM, France) to PC, JR and AdB, ANR Naturadyre to JR (France) and HFSP and NCBS to APJ (India).

**References**

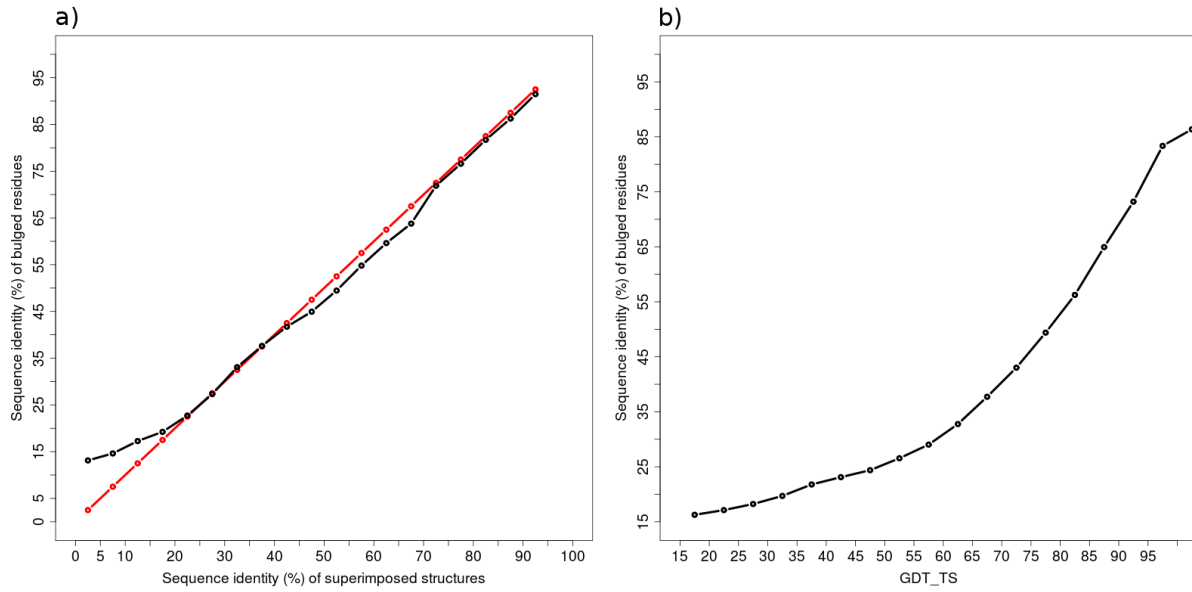
1. Eisenberg D (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A* 100:11207-11210.
2. Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* 37:205-211.
3. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* 37:251-256.
4. Tyagi M, Bornot A, Offmann B, de Brevern AG (2009) Analysis of loop boundaries using different local structure assignment methods. *Protein Sci* 18:1869-1881.
5. Fourrier L, Benros C, de Brevern AG (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* 5:58.
6. Penel S, Morrison RG, Dobson PD, Mortishire-Smith RJ, Doig AJ (2003) Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. *Protein Eng* 16:957-961.
7. Regan L (1994) Protein structure. Born to be beta. *Curr Biol* 4:656-658.
8. Colloc'h N, Cohen FE (1991) Beta-breakers: an aperiodic secondary structure. *J Mol Biol* 221:603-613.
9. Wouters MA, Curmi PM (1995) An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins* 22:119-131.
10. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN (1998) Determinants of strand register in antiparallel beta-sheets of proteins. *Protein Sci* 7:2287-2300.
11. Fujiwara K, Toda H, Ikeguchi M (2012) Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. *BMC Struct Biol* 12:18.
12. Mandel-Gutfreund Y, Zaremba SM, Gregoret LM (2001) Contributions of residue pairing to beta-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J Mol Biol* 305:1145-1159.
13. Ho BK, Curmi PM (2002) Twist and shear in beta-sheets and beta-ribbons. *J Mol Biol* 317:291-308.
14. Kumar S, Bansal M (1998) Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys J* 75:1935-1944.
15. Richardson JS, Getzoff ED, Richardson DC (1978) The beta bulge: a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci U S A* 75:2574-2578.
16. Chan AW, Hutchinson EG, Harris D, Thornton JM (1993) Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci* 2:1574-1590.
17. Milner-White EJ (1987) Beta-bulges within loops as recurring features of protein structure. *Biochim Biophys Acta* 911:261-265.
18. Leader DP, Milner-White EJ (2009) Motivated proteins: a web application for studying small three-dimensional protein motifs. *BMC Bioinformatics* 10:60.
19. James Milner-White E, Poet R (1987) Loops, bulges, turns and hairpins in proteins. *Trends in biochemical sciences* 12:189-192.
20. Larson SM, Davidson AR (2000) The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci* 9:2170-2180.
21. Bornot A, Etchebest C, de Brevern AG (2011) Predicting protein flexibility through the prediction of local structures. *Proteins* 79:839-852.
22. Richardson JS, Richardson DC (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A* 99:2754-2759.

23. Capraro DT, Roy M, Onuchic JN, Gosavi S, Jennings PA (2012) beta-Bulge triggers route-switching on the functional landscape of interleukin-1beta. *Proc Natl Acad Sci U S A* 109:1490-1493.
24. Zhao L, Ericksen B, Wu X, Zhan C, Yuan W, Li X, Pazgier M, Lu W (2012) Invariant gly residue is important for alpha-defensin folding, dimerization, and function: a case study of the human neutrophil alpha-defensin HNP1. *J Biol Chem* 287:18900-18912.
25. Axe DD, Foster NW, Fersht AR (1999) An irregular beta-bulge common to a group of bacterial RNases is an important determinant of stability and function in barnase. *J Mol Biol* 286:1471-1485.
26. Siepen JA, Radford SE, Westhead DR (2003) Beta edge strands in protein structure prediction and aggregation. *Protein Sci* 12:2348-2359.
27. Azinas S, Colombo M, Barbiroli A, Santambrogio C, Giorgetti S, Raimondi S, Bonomi F, Grandori R, Bellotti V, Ricagno S, Bolognesi M (2011) D-strand perturbation and amyloid propensity in beta-2 microglobulin. *FEBS J* 278:2349-2358.
28. Wu XH, Wang Y, Zhuo Z, Jiang F, Wu YD (2012) Identifying the Hotspots on the Top Faces of WD40-Repeat Proteins from Their Primary Sequences by beta-Bulges and DHSW Tetrads. *PLoS One* 7:e43005.
29. Joseph AP, Agarwal G, Mahajan S, Gelly J-C, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadié H, Schneider B, Cadet F, Srinivasan N, de Brevern AG (2010) A short survey on Protein Blocks. *Biophysical Reviews* 2:137-145.
30. Joseph AP, Srinivasan N, de Brevern AG (2011) Improvement of protein structure comparison using a structural alphabet. *Biochimie* 93:1434-1445.
31. Gelly JC, Joseph AP, Srinivasan N, de Brevern AG (2011) iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res* 39:W18-23.
32. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng* 6:377-382.
33. Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Structural Biology* 5:17.
34. Mansiaux Y, Joseph AP, Gelly JC, de Brevern AG (2011) Assignment of PolyProline II conformation and analysis of sequence--structure relationship. *PLoS One* 6:e18401.
35. Bornot A, de Brevern AG (2006) Protein beta-turn assignments. *Bioinformatics* 1:153-155.
36. de Brevern AG (2005) New assessment of a structural alphabet. *In Silico Biol* 5:283-289.
37. de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 41:271-287.
38. Joseph AP, Srinivasan N, de Brevern AG (2012) Cis-trans peptide variations in structurally similar proteins. *Amino Acids* 43:1369-1381.
39. Fuhrmann CN, Kelch BA, Ota N, Agard DA (2004) The 0.83 Å resolution crystal structure of alpha-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain. *J Mol Biol* 338:999-1013.
40. Kim CW, Michaels ML, Miller JH (1992) Amino acid substitution analysis of E. coli thymidylate synthase: the study of a highly conserved region at the N-terminus. *Proteins* 13:352-363.
41. Dion-Schultz A, Howell EE (1997) Effects of insertions and deletions in a beta-bulge region of Escherichia coli dihydrofolate reductase. *Protein Eng* 10:263-272.

42. Chen PY, Gopalacushina BG, Yang CC, Chan SI, Evans PA (2001) The role of a beta-bulge in the folding of the beta-hairpin structure in ubiquitin. *Protein Sci* 10:2063-2074.
43. Dessen A, Volchkov V, Dolnik O, Klenk HD, Weissenhorn W (2000) Crystal structure of the matrix protein VP40 from Ebola virus. *EMBO J* 19:4228-4236.
44. Gomis-Ruth FX, Dessen A, Timmins J, Bracher A, Kolesnikowa L, Becker S, Klenk HD, Weissenhorn W (2003) The matrix protein VP40 from Ebola virus octamerizes into pore-like structures with specific RNA binding properties. *Structure* 11:423-433.
45. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res* 32:D189-192.
46. Hutchinson EG, Thornton JM (1996) PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* 5:212-220.
47. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235-242.
48. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536-540.
49. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.
50. Laskowski RA (2009) PDBsum new things. *Nucleic Acids Res* 37:D355-359.
51. Joseph AP, Bornot A, de Brevern AG. 2010. Local Structure Alphabets. *In Protein Structure Prediction* (H. Rangwala and G. Karypis, eds.) pp. 75-106. wiley.
52. Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern* 43:59-69.
53. Kohonen T. 2001. *Self-Organizing Maps* (3rd edition). Springer.
54. Rabiner LR (1989) A tutorial on hidden Markov models and selected application in speech recognition. *Proceedings of the IEEE* 77:257-286.
55. Tyagi M, de Brevern AG, Srinivasan N, Offmann B (2008) Protein structure mining using a structural alphabet. *Proteins* 71:920-937.
56. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, Offmann B (2006) A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins* 65:32-39.
57. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG, Offmann B (2006) Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res* 34:W119-123.
58. Martin A, Porter C (2010) <http://www.bioinf.org.uk/software/profit/>.
59. Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31:3370-3374.
60. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T (2011) Assessment of template based protein structure predictions in CASP9. *Proteins* 79 Suppl 10:37-58.
61. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.
62. Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. *Science* 256:1443-1445.
63. 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.
64. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.* 7:306-317.

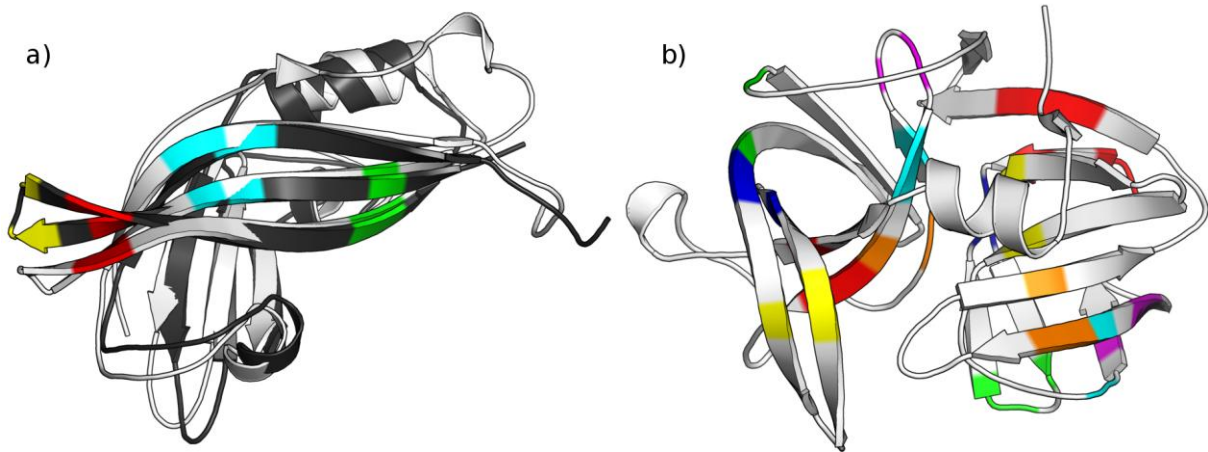
65. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comp.* 4:435-447.
66. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: Fast, Flexible and Free. *J. Comp. Chem.* 26:1701-1718.
67. Berendsen HJC, Postma JPM, van Gunsteren WF, Di Nola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.
68. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24:1999-2012.
69. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926-935.
70. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.
71. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh ewald potential. *J. Chem. Phys.* 103:8577–8592.
72. Morgunova E, Dauter Z, Fry E, Stuart DI, Stel'mashchuk V, Mikhailov AM, Wilson KS, Vainshtein BK (1994) The atomic structure of Carnation Mottle Virus capsid protein. *FEBS Lett* 338:267-271.
73. Hopper P, Harrison SC, Sauer RT (1984) Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. *J Mol Biol* 177:701-713.

**Figure Legends**

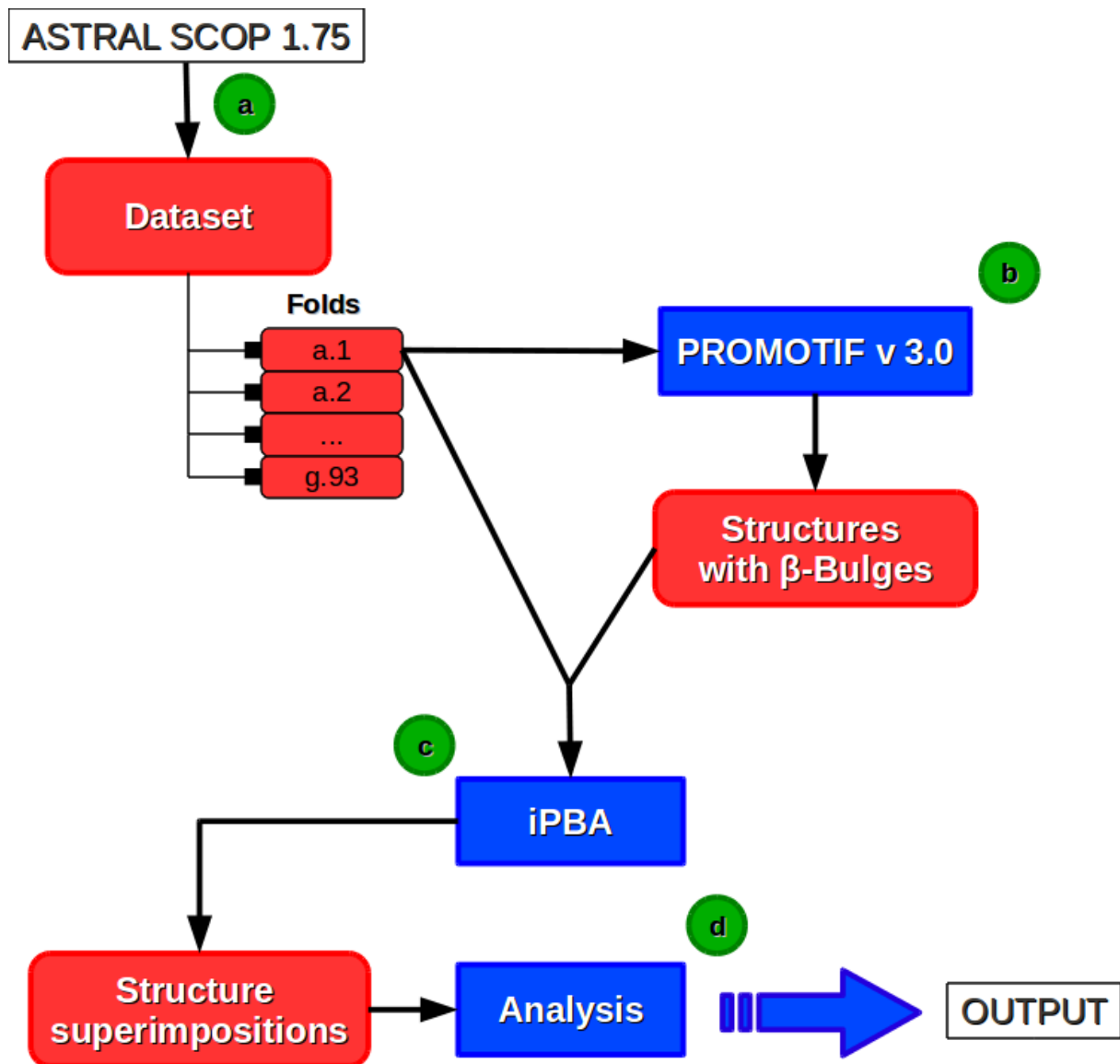


**Figure 1.** Conservation of bulged residues. The average sequence identity of non-X residues of superimposed  $\beta$ -bulges with respect to (a) sequence identity of superimposed structures (black line), the red line represents the reference line  $f(x)=x$ , and (b) GDT\_TS of structural alignments.

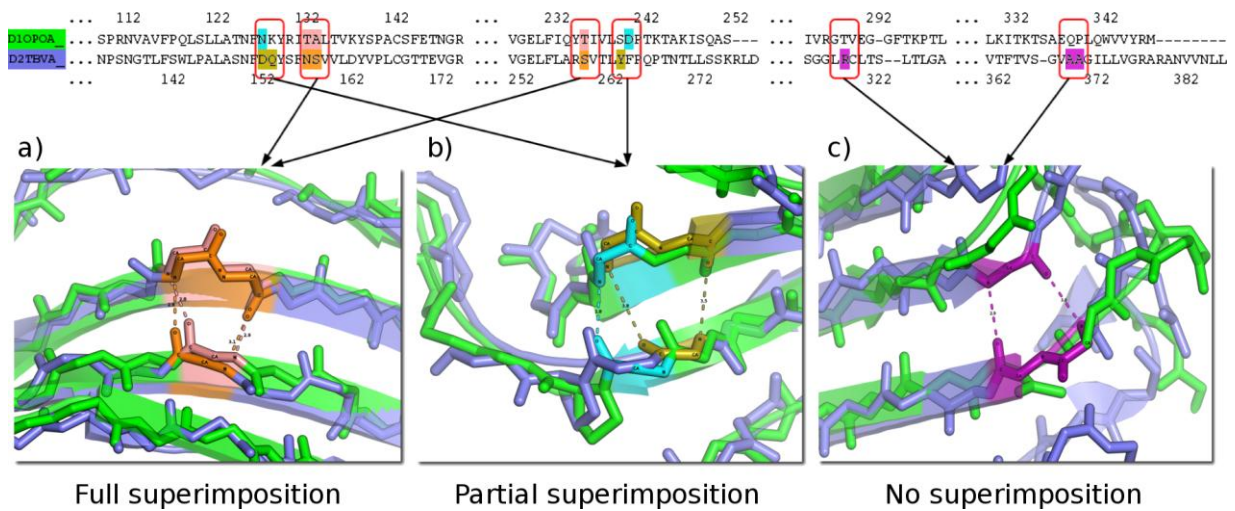




**Figure 2.** *Stability of  $\beta$ -bulges.* (a) The two SCOP domains d1es6a1 (in white) and d1h2ca\_ (in black) of Ebola virus matrix protein share identical sequence and similar 3D structure (GDT\_TS = 46.62 and RMSD = 2.06Å). However they exemplify variable number of  $\beta$ -bulges (3 for d1es6a1 and 4 for d1h2ca\_). Three  $\beta$ -bulges are conserved between these homologous structures (two partial superimpositions in red and blue, and one full superimposition in green) and one  $\beta$ -bulge is missing in d1es6a1 (in yellow). This observation highlights the possible effect of protein flexibility or crystal packing on  $\beta$ -bulges formation. (b) The  $\alpha$ -lytic protease (1SSX), 198 residues long, is characterized by a large number of  $\beta$ -bulges (15 in number) compared to its length. This structure was used in three molecular dynamic simulations (at temperatures 298, 310 and 353 K) to study the stability of  $\beta$ -bulges (see Figure S4 and Table S6).



**Figure 3.** Principle of the analysis of  $\beta$ -bulges. (a) Protein structures are taken from the ASTRAL SCOP dataset(45). (b) Classical secondary structures and  $\beta$ -bulges are assigned with PROMOTIF software (46). All the protein structures containing at least one  $\beta$ -bulge and belonging to same SCOP fold are superimposed, using (c) iPBA software(30-31), and then (d) the selected  $\beta$ -bulges superimpositions are analyzed.



**Figure 4.** *The three possible cases of  $\beta$ -bulge structure superimposition.* The superimposition of carnation Mottle virus capsid protein chain A (PDB code 1OPO(72)) and coat protein of tomato bushy stunt virus chain A (PDB code 2TBV(73)). Structure comparison results in a GDT\_TS score of 64.0, (RMSD equals to 1.72Å and 26.6 % sequence identity). Each protein had 6  $\beta$ -bulges including 4 which are structurally equivalent. It exhibits three possible cases: (a) *Full superimposition*: all residues in  $\beta$ -bulges are found to be perfectly aligned, e.g. for the Classic  $\beta$ -bulge of (pink colour) carnation mottle virus capsid protein at residue Thr-236 (residue-X), Thr-133 and Ala-134, and coat protein of tomato bushy stunt virus (orange colour) with residues Ser-260 (residue-X), Asn-158 and Ser-159; (b) *Partial Superimposition*: only some of the residues composing  $\beta$ -bulges are aligned, e.g. the Bent  $\beta$ -bulge of 1OPO (in blue), composed by residues Asn-128 and Asp-241, with the Classic  $\beta$ -bulge of 2TBV (in gold), composed by residues Tyr-264 (residue-X), Asp-153 and Gln-154. Only 2 residues are aligned with each other, and (c) *No superimposition*:  $\beta$ -bulge residues found in one protein structure but not aligned with  $\beta$ -bulge residues on the second one, e.g. the Classic  $\beta$ -bulge of 2TBV (in purple), composed of residues Arg-319 (residue-X), Ala-370 and Ala-371, have no direct counterparts.