Biochimie (2009) *in press*

Analysis of protein contacts into Protein Units

Guilhem Faure[1,#,+], Aurélie Bornot[1,2,#] & Alexandre G. de Brevern [1,2,*]

[1] INSERM UMR-S 726, Equipe de Bioinformatique Génomique et Moléculaire (EBGM), DSIMB, Université Paris Diderot - Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France.
[2] INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot - Paris 7, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France

Short title: peel it

* Corresponding author:
mailing address: Dr. de Brevern A.G., INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot - Paris 7, INTS, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
E-mail: alexandre.debrevern@univ-paris-diderot.fr
Tel: +33(1) 44 49 30 38
Fax: +33(1) 47 34 74 31

[#] both authors contribute equally to this work.

[+]Present adress : Laboratoire de Biologie Structurale et Radiobiologie, iBiTec-S (Institut de Biologie et de Technologie de Saclay), CEA/CNRS URA2096, Bâtiment 144, Point courrier 22, Commissariat à l'Énergie Atomique, 91191 Gif sur Yvette cedex, FRANCE

key words : amino acid; protein domain; side-chains; secondary structures; protein contacts, protein extremities, structural alphabet, Protein Blocks.

Abstract

Three-dimensional structures of proteins are the support of their biological functions. Their folds are maintained by inter-residue interactions which are one of the main focuses to understand the mechanisms of protein folding and stability. Furthermore, protein structures can be composed of single or multiple functional domains that can fold and function independently. Hence, dividing a protein into domains is useful for obtaining an accurate structure and function determination.

In previous studies, we enlightened protein contact properties according to different definitions and developed a novel methodology named Protein Peeling. Within protein structures, Protein Peeling characterizes small successive compact units along the sequence called protein units (PUs). The cutting done by Protein Peeling maximizes the number of contacts within the PUs and minimizes the number of contacts between them. This method is so a relevant tool in the context of the protein folding research and particularly regarding the hierarchical model proposed by George Rose.

Here, we accurately analyze the PUs at different levels of cutting, using a non-redundant protein databank. Distribution of PU sizes, number of PUs or their accessibility are screened to determine their common and different features. Moreover, we highlight the preferential amino acid interactions inside and between PUs. Our results show that PUs are clearly an intermediate level between secondary structures and protein structural domains.

1 Introduction

The knowledge of the three-dimensional (3D) structure of proteins is critical for understanding their biological functions. 3D structures are a valuable source of data for understanding their biological roles, their potential implication in some diseases mechanisms, and for progressing in drug design [1-3]. The interaction between residues composing proteins and their surroundings in the cell produces a well-defined folded protein, *i.e.*, the native state [4]. The resulting three-dimensional structure is determined by the amino acid sequence. Nonetheless, the mechanism of protein folding is not completely understood [5], neither is the protein aggregation [6]. Several models have been proposed for protein folding, *e.g.*, the framework model [7, 8], the diffusion-collision model [9], the hydrophobic collapse model [10] or the nucleation and growth mechanism [11]. The hierarchical model proposed by George Rose [12] is nowadays the most popular one. This principle is a hierarchical process [13-17] coupled with the hydrophobic effect as the driving force [18, 19]. Simulations based on this principle were done in a very elegant way by Srinivasan and Rose; they considered steric effects, conformational entropy with hydrophobic interactions and hydrogen bond formations [20-22]. In order to analyze the hierarchical process that conducts the protein folding, it is also possible to unfold proteins using molecular dynamics [23-26]. Plaxco and co-workers have shown that protein folding speeds correlate with the topology of the native protein [27]. Proteins which quickly fold are usually mostly stabilized by local structures, *e.g.*, turns, whereas slow folders usually present more non-local structures, *e.g.*, β-sheet [28].

Protein structures can be seen as composed of single or multiple functional domains that can fold and function independently. Dividing a protein into domains is useful for more accurate structure and function determination. Methods for phylogenetic analyses or protein modeling usually perform best for single domains [29]. The commonly used principle for automatic domain parsing is that interdomain interaction under a correct domain assignment is weaker than the intradomain interaction (PUU [30], DOMAK [31], 3Dee [32, 33], DETECTIVE [34], DALI [35], STRUDL [36], DomainParser [37, 38], Protein Domain Parser [39] and DDOMAIN [40]). Innovative approaches have been used in this context, *e.g.*, graph theory [41] and Normal Mode Analysis approach [42]. Most of the time, the size of protein domains remains important (often more than a hundred residues), these approaches maximized the number of contacts within a domain and are often benchmarked on a manual definition of structural domains [43]. A recent and well-designed analysis highlighted the complexity of defining automatically structural domains [44].

Some authors have proposed different methods to hierarchically split proteins into compact units smaller than protein domains [15, 45-48]. In this field, we should notice the most advanced research, namely DIAL [45, 47] and his accompanying database [49]. In this method, domains are considered to be clusters of secondary structure elements. Thus, helices and strands are first clustered using intersecondary structural distances between Cα positions. In a second step, dendograms based on this distance measure are used to identify sub-domains. Their goal was to describe the different levels of protein structure organization. Wetlaufer was the first to examine the organization of known structures and suggested that the early stages of 3D structure formation, *i.e.* nucleation, occur independently in separate parts of these molecules [50, 51]. These folding units have been proposed to fold independently during the folding process, creating structural modules which can be assembled to give the native structure.

We have likewise developed a method called Protein Peeling [52]. This algorithm dissects a protein into Protein Units (PUs). A PU is a compact sub-region of the 3D structure corresponding to one sequence fragment. The basic principle is that each PU must have a high number of intra-PU contacts, and, a low number of inter-PU contacts. Protein Peeling works from the Cα-contact matrix translated into contact probabilities. Based on the Matthews' coefficient correlation (MCC) [53] between contact sub matrices, an optimization procedure defines optimal cutting points. The latter separate into two or three PUs the examined region. The process is iterated until the compactness of the resulting PUs reaches a given limit, fixed by the user. The PU compactness is quantified by an index, *CI* (compaction index). This index is based on a correlation coefficient *R* between the mutual entropy of the contact submatrices [54-57]. Thus, organization of protein structures can be considered in a hierarchical manner: secondary structures are the smallest elements, and, Protein Units are intermediate elements leading to structural domains.

Protein contacts are essential for protein folding [58]. They have been used to develop energy potentials interesting for folding simulations [59, 60]. Inter-residue interactions can be characterized by contact order (CO) and long-range order (LRO) parameters that have a strong correlation with the folding rate of small proteins [27, 61-63].

In a recent work [64], we studied contacts within protein structures according to various criteria (lengths of proteins, SCOP classes, secondary structures, amino acid frequencies, accessibility). We showed that the distribution of the average contact number was clearly dependant to atoms taken as references. One of the most interesting results was the fact that contacts taken into account according to a given type of distance is not compulsorily taken into account by another one, *e.g.*, only 22% of the observed contacts considering side-chains are found if only alpha carbons (Cα) are considered [64]. Specificities were found according to the distance in the sequence between residues in contact and some differences were observed compared to the literature [65]. Moreover, we highlighted biases of the side-chain replacement methods [66-72].

In this study, we went deeper into the hierarchical organization of proteins by analyzing the contacts found inside and between protein sub-units defined by Protein Peeling, *i.e.*, Protein Units [52, 73]. We accurately analyzed the behaviors of Protein Peeling for various values of R (higher is the R value, deeper is the cutting). Distribution of PU sizes and number of PUs have been screened to determine if some common features could be obtained. The preferential amino acid interactions have been compared to the results previously obtained with complete proteins. This work enlightens that PUs are clearly an intermediate level between secondary structures and protein structural domains. Moreover, the major differences between the various ways to define protein contacts and thus potential repercussions on analysis were also taken into account and analyzed.

2 Material and methods

2.1 *Main principle of the analysis.*

Figure 1 shows the principle of the analysis. From the Protein DataBank (PDB) [74] was selected a non-redundant set of proteins (see below for the selection criteria). For an analysis purpose, protein structures were assigned in terms of secondary structure and Protein Blocks [54, 75]. Then, each protein, was cut into Protein Units (PUs) using the Protein Peeling approach (see Figure 1). Finally, a detailed analysis of the characteristics of PUs in terms of length, amino-acid composition and structure was realized. Moreover, a particular attention was given to contacts within and between protein units.

For comparison purpose, all analysis realized for protein units were also performed for complete proteins thus taken as reference.

2.2 *Databank.*

A non-redundant protein databank has been initially built using PDB-REPRDB [76, 77]. It was composed of 1,736 protein chains taken from the PDB. The set contained proteins with no more than 10% pairwise sequence identity. We selected chains with a resolution better than 2.5 Å and a R-factor less than 0.2. Pairwise root mean square deviation (*rmsd*) values between all chains were more than 10 Å. Only proteins with more than 99% of complete classical amino acids were conserved. Moreover, proteins that cannot be used by software used during analysis process have also been excluded. Thus, we retained 1,230 protein chains corresponding to 377,232 residues.

2.3 *Protein Peeling.*

The Protein Unit (PU) is an intermediate level between secondary structures and protein domains [52]. A PU has a great number of inner contacts (intra-PUs) and few contacts with other PUs of protein (inter-PUs). The principle of Protein Peeling is the following: the peeling starts from a matrix of contacts normalized in probabilities and looks to cut a protein into 2 or 3 PUs (or an already cut out PU). A partition index (*PI*) is calculated in each position. The *PI* is based on the Matthews Coefficient Correlation [78], it is thus maximal when the sum of the contacts of two matrices intra-PUs is high and that of inter-PUs is weak. The *PI* thus defines the regions to be cut out; parsing into 3 PUs is also tested with all positions. To characterize the compactness of PUs defined, a compactness index based on mutual information is calculated, it uses the sum of the probabilities associated with each PU and indicates when to stop cutting, when it reaches a given threshold *R* (see [52] for more details and Figure 2 for an example). A refinement of cutting is carried out thanks to the method of *pruning* which checks that PUs lately generated are compact [73].

2.4 *Contact definitions.*

Two residues are in contact if they are at a lower distance than a distance τ from one another (cf. Figure 1 of [64]). Various distances can be used [64]. Here, distances between Cα with threshold value equal to 8 Å (noted Cα^8) are used. The analyses are so comparable to those of [64] and applied to the principle of Protein Peeling [52, 73]. The short distance interactions in the sequence will not be taken into account, *i.e.* $D/2$ residues surrounding the studied residue are thus not considered ($D = 6$ [64]).

2.5 *Analysis of preferential contacts.*

Analysis of the observed contacts is carried out by computing the relative contact frequency (noted *rf* in the text) of the amino acid of type *i* found in contact (distance lower than τ) with the amino acid of type *j* :

$$r^{f} aa_{ij}^{contact} = \frac{faa_{ij}^{contact}}{f^{DB} aa_{j}} \tag{1}$$

with $r^{f} aa_{ij}^{contact}$ the relative contact frequency of the contacts of the amino acid of type *i* with amino acid of type *j* : $faa_{ij}^{contact} = Naa_{ij}^{contact} / Naa_{i}^{contact}$; $Naa_{ij}^{contact}$ is the number of contacts between residues of types *i* and *j*, and $Naa_{i}^{contact}$ the total number of contacts of amino acid of type *i*. This value is normalized by $f^{DB} aa_{j}$, the average frequency of amino acid of type *j* in the studied protein databank.

2.6 *Equivalent number (Neq).*

This index, we previously introduced for prediction purpose [54, 57, 79], is based on the information theory. It is used here to estimate the equilibrium between the lengths of the different PUs generated for each protein *P* and at each value of *R*. It is defined as the exponential of the Shannon entropy $H(P^R)$ [80]:

$$H\left(P^{R}\right) = -\sum_{i=1}^{n^{R}} l_{i}^{R} \times \ln l_{i}^{R} \tag{2}$$

$$N_{eq}\left(P^{R}\right) = \exp\left[H\left(P^{R}\right)\right] \tag{3}$$

with n^R, the number of Protein Units for a given *R* value, l_i^R is the normalized length of the i^{th} Protein Unit. This index denoted $N_{eq}(P^R)$, for "equivalent number of Protein Units" varies between 1 and n^R. For $N_{eq}(P^R) = 1$, only one PU represents the whole protein. For $N_{eq}(P^R) = n^R$, all PUs have exactly the same length (see Figure 3).

As there is no reason that the Protein Peeling cuts all proteins in the same number of PUs n^R, a normalization of $N_{eq}(P^R)$ is necessary for comparing the cutting of different proteins. Thus, each $N_{eq}(P^R)$ is normalized by $N_{eq}^{\max}(P^R)$, the maximal $N_{eq}(P^R)$ value, *i.e.*, n^R. Finally, the normalized values of $N_{eq}(P^R)$ vary between 0 and 1. Values closer to 0 correspond to a cutting with one main PU representing a large proportion of the protein and possibly other smaller PUs, whereas values closer to 1 characterize a cutting in several PUs of similar size.

2.7 *Analyses.*

Residue accessibility has been calculated with nAccess software (version 2.1.1) [81]. Secondary structure assignment has been done using DSSP software (version 2000, CMBI) [82]. As DSSP gives more than three states, we have reduced them: the α-helix contains α, 3_{10} and π-helices, the β-strand contains only the β-sheet and the coil corresponds to everything else (β-bridges, turns, bends, and coil). Software default parameters were used. Protein Peeling was carried out using the software of the Protein Peeling web server [52, 73]. Outputs were adapted for our study. Proteins were characterized according to the manually assigned classes of SCOP all-α, all–β, α/β and α +β [83]. The automatic categorization of Michie and co-workers was also used [84], it defines 3 classes: α, β and others. The first contains proteins having more than 40% of α-helices and less than 15% of β-sheets, the second less than 15% of α-helices and more than 30% of β-sheets, otherwise proteins are assigned to the *others* class.

2.8 *Protein Blocks.*

Protein Blocks (PBs) correspond to a set of 16 local prototypes [54, 56, 85-89], labeled from *a* to *p*, of 5 residues length based on Φ, Ψ dihedral angles description [90]. They were obtained by an unsupervised classifier similar to Kohonen Maps [91, 92] and Hidden Markov Models [93]. The PBs *m* and *d* can be roughly described as prototypes for central α-helix and central β-strand, respectively. PBs *a* through *c* primarily represent β-strand N-caps and PBs *e* and *f*, C-caps; PBs *g* through *j* are specific to coils, PBs *k* and *l* to α-helix N-caps, and PBs *n* through *p* to C-caps. This structural alphabet allows a reasonable approximation of local protein 3D structures [54]with a root mean square deviation (*rmsd*) now evaluated at 0.42 Å [75].

2.9 *Analysis of the over- and under-represented contacts.*

We can analyze the contacts of a PB (or a secondary structure state) (i) by assessing globally the specificity of each PB (secondary structure state, resp.), *i.e.*, which PB have the most informative contacts in terms of contact distribution, and (ii) by determining which PB is preferentially associated to a given PB. To deal with the first point, we have used the relative entropy or Kullback-Leibler asymmetric divergence measure [94].

$$K(\mathrm{p}, \mathrm{q}) = \sum_{i=1}^{j} p_i \ln\left(\frac{p_i}{q_i}\right)$$

It quantifies the contrast between the observed contacts frequencies between PBs (respectively, secondary structures) **p**: [42]*i*=1, . . . ,*j* and a reference probabilistic distribution **q** [95]. We have applied this expression for assessing the divergence *K* (**p**, **q**) of observed contacts distribution *p* observed for a given PB and a distribution of contacts *q* in the databank (taken as reference, *i.e.*, the frequency of PBs or of secondary structures). *j* equals 3 for the secondary structures and 16 for the Protein Blocks.

Concerning the second point, we have normalized the contact occurrences between each PB (secondary structure) into a *Z*-score. The contact occurrences for a given PB *x* (secondary structure state respectively) were normalized into a Z-score = ($n^{observed}(i,x)$ - $n^{theoretical}(i,x)$) / $\sqrt{n^{theoretical}(i,x)}$, with $n^{observed}(i,x)$ the number of contacts *i* observed in PB *x*, and $n^{theoretical}(i,x)$ the number expected. The product of the occurrence of PB *x* with the frequency of contacts *i* in the entire data bank equals $n^{theoretical}(i,x)$. Positive Z-scores (respectively negative) correspond to overrepresented amino acids (respectively underrepresented) in PB *x*.

3 **Results**

For analyzing the initial protein contacts that could potentially appear during the protein folding according to the hierarchical model, Protein Peeling is an interesting tool because it enlightens compact sequential units, namely Protein Units (PUs). Contrary to protein domain assignment, it is one of the few methods [47, 96] available that can go deep in the cutting process of the protein structure. It is based on a criterion *R* based on mutual information and that assesses the cutting process. The cutting of proteins is done recursively (see Figure 2 and Methods section). In this study, we address the hierarchical organization of proteins that is put in light using Protein Peeling.

In the first part, we review the influence of different parameters on the cutting process and characterized the generated PUs. Thus, we observed the number of PUs generated for a specific *R* value. It is important as the Protein Peeling cuts more than the classical protein domain assignment methodologies. Following this analysis, we analyzed the sizes of the PUs with regards to the sizes of the complete proteins and their structural SCOP classes. PUs were finally categorized according to their amino acid composition and their accessibility.

In a second part, we focused on contacts within and between PUs and structurally characterized their context in proteins and PUs.

3.1 *Analysis of the Protein Peeling cutting process and General characteristics of compact Proteins Units.*

Average Number of PUs. The criterion R of the Protein Peeling is based on the calculation of mutual information and conditions the cutting stop. Hence, first, we analyses the influence of this parameter on the number of PUs generated. Proteins were cut out for R values ranging between 20 and 90 (see Figure 4 for a distribution of the number of PUs). For low R values, peeling results in 2 to 3 PUs per protein. A 40 R value leads to the cutting of only a few proteins in 4 PUs. Actually, for R equals 70, the average number of PUs is 3. For an 80 R value, this average reaches 4. Finally, for a R value of 90, 6 PUs are observed in average but some proteins can be cut in more than 10 PUs, see Figure 5 for an example.

Influence of the Peeling Parameter on PUs sizes. The cutting process does not lead to equivalent sizes of PUs for all proteins. Nevertheless, the analysis of the PU size distribution (in terms of percentage of the protein length) does not show strong specificity to R value. All sizes of PUs are observed, ranging between 10 and 90% of the protein length. To observe mainly short PUs (15-20% of the protein length), R must be higher than 70. In order to go further, we search for groups of proteins with the same evolution of their PU size distribution according to R. In this purpose, for each protein, we considered the fraction of the protein length that each PU represents. The distribution of these different fractions for a given protein and a given R was characterized by the "equivalent number of Protein Units", namely $N_{eq}(P^R)$, based on the Shannon entropy (see Methods section). As proteins could have very different length and number of PUs, we normalized the $N_{eq}(P^R)$ value to unbias the statistical analysis (see Methods section). Then, a hierarchical clustering of proteins was done according to their normalized N_{eq} values for R ranging from 20 to 90. Finally, four groups of proteins were highlighted (see supplementary data 1).

The three first clusters have PUs with relatively similar sizes for low R values. Their normalized N_{eq} goes to a homogeneous distribution with value equals to 0.88 for R = 90. The fourth group represents only 8% of the protein of our databank and has a different profile. For low R values, their PUs have very different sizes (normalized N_{eq} of 0.64). For high R values, they keep this imbalance with a 0.77 normalized N_{eq} value. The characteristic of this group is to contain a percentage of big size proteins (>450 residues) much more important than the others (from 2 to 8 times more).

Influence of Protein characteristics on PUs sizes. A detailed analysis of the distribution of PU sizes in proteins does not show any particular behavior according to the length of protein. As expected, the small size proteins (less than 150 residues) have PUs of equivalent sizes whatever the R value; this behavior decreases slightly with the increase of protein size. No differences between SCOP classes are observed either. Thus, cutting performed by Protein Peeling does not systematically associates residues with a given type of secondary structure.

Amino Acid composition of PUs. A Sammon map [97] was used to analyze the relative amino acid frequency in PUs and in entire proteins. Each amino acid association is represented by a vector of 20 values (values are normalized). Thus a Sammon map can be computed from distance between the amino acids using the Euclidean distance between the amino acid vectors. Similar approaches have been applied to Protein Blocks in a recent study [89]. We found that, with the increase of R values, amino acid composition of PUs deviated more and more from the average composition of proteins. This phenomenon is partly due to the reduction in the size of fragments that could be associated to a compositional bias, *i.e.* not enough data. Nonetheless, as this light tendency begins at the first stages of cutting, it is partly a noteworthy behavior of the PUs.

Accessibility of PUs. The analysis of the average relative accessibility of PUs was carried out on biggest proteins of more than 400 residues, others having an important initial average accessibility. The first cuts, *i.e.*, R equals 20, do not generate a large number of PUs with high accessibility. Analysis of the most accessible PU for each protein shows that only 29% of these PUs have a relative accessibility higher than 35% (*rather accessible*) and 9% more than 50% (*very accessible*). For $R = 90$, more accessible PUs have been generated, 61% are *rather accessible*, but only 16% are *very accessible*. Concerning the least accessible PU of each protein, 50% have a relative accessibility still higher than 20%. Interestingly, Protein Peeling does not create exposed PUs distinct from buried PUs, *i.e.*, no PU would correspond to a protein “core”.

Characterizing protein extremities. A particular interest was focused on protein extremities. Actually, extremities of proteins are often considered as “mobile” [98], because they have fewer constraints than the hydrophobic core of the protein. We thus studied all the protein extremities using the Protein Peeling. If a PU, representing less than 20% of the size of protein, is cut early in the process of peeling and is not cut again, we considered it as *mobile*. This way, we found that half of the proteins have mobile extremities, 23% have mobile N-termini (noted *Nt*), 27% mobile C-termini (*Ct*) and 2% the two mobile ends. Globally, the mobile ends can be all-α, all-β or others (according to [84]), but a more important tendency for the all-α category is observed (37.4% for *Nt* and 48.3% for *Ct*). α-helices are not conditioned by long range contacts within the sequence like β-sheets; this tendency seems logical.

3.2 *Analysing contacts between and within Protein Units*

In the second part of the results, we focus on the contacts observed within and inside the Protein Units. Firstly, as we previously did on global structures [64], we globally, analyze the contacts within the Protein Units. Then, we refine our analysis by considering the different protein fold classes.

For an analysis purpose, we performed this research by categorizing PUs according to their fold class, *i.e.* mainly α, mainly β or other. Furthermore, we took advantage of the accurate description given by the Protein Blocks alphabet [99] to analyze precisely the structural environment of contacts .

Preferential contacts within PUs. The dissection of proteins using Protein Peeling was analyzed in terms of relative contact frequencies or *rf* (see Methods section) for the contacts within PUs, namely intra-PUs, and also for contacts between PUs, namely inter-PUs (see Table 1 and supplementary data 2). These *rf* values were compared to the *rf* values observed for the whole databank [64]. Surprisingly, for all R values, the variations observed between the *rf* of the databank and *rf* of intra-PUs are very limited. The Cysteine increases by 0.3 its interaction with itself and with the Tryptophan by 0.1, all others couples have a variation less than 0.1. Only 14% of the couples have a difference more than 0.05 and, on average the *rf* differences are 0.024. In comparison to previous analyses, these values are very low [64].

This conservation of preferential contacts is striking as the percentage of intra-PU contacts represents only 60% of the protein databank contacts for R = 90. Protein Peeling thus creates PUs which preserves an interactions distribution similar to the one observed in whole proteins. From this point of view, PUs could be characterized as small protein domains.

The inter-PU contacts present much more variations in comparison (see Table 2), but these variations remain generally weak. Only 10 interaction couples have *rf* differences more than 0.2. The involved amino acids are Cysteine (6 times), Glycine (2 times), Methionine (1) and Tryptophan (1). Their differences are more important for weak *R* values, and then they tend to approach the *rf* values observed in the databank. The other implicated amino acids are Valine, Isoleucine, Histidine and Proline.

PUs within SCOP classes. The distributions of *rf* within SCOP classes showed some important (but not drastic) changes compared to the *rf* of protein databank [64]. For the intra-PU contacts of α/β class proteins, only 1 *rf* presents a difference higher than 0.2 compared to the *rf* of its SCOP class, 2 for the class all-β and 3 for the class all-α. For this last class 39 *rf* have a difference more than 0.1; the amino acids mainly concerned are Tryptophan, Phenylalanine, Tyrosine, Cysteine, and Methionine. Only the class α+β has a true specificity (see Table 2), 32 *rf* have a difference more than 0.2 and 7 of more than 0.5. The most affected couple is Asparagine- Asparagine, its *rf* increases from 0.93 before cutting to 1.50 for R = 90. Cysteine (8 times), Tryptophan (8 times), Methionine (5 times) and Histidine (4 times) are then the most affected. A general tendency is observed: the more the *R* values increase, the more the *rf* differences increase, even if that is sometimes weak. Analysis of the inter-PU contacts is more complex because proteins associated with a SCOP class represent only a part of proteins and, inter-PU contacts an even weaker part. Strong changes are observed, but the limited number of data does not authorize us to draw conclusions.

Categorization of PUs in α, β and other classes. Like proteins, PUs can be classified using the criteria of Michie and collaborators [84]. Thus a PU-α, a PU-β and a PU-other classes can be defined. This approach makes possible to characterize inter-PUs, *e.g.* α-β. For R = 20, the number of contacts within PUs-α represents 17.5% of contacts, within PUs-β represents 15.8% and within others 50.6%, the remaining are the inter-PU contacts. The distribution of the number of contacts within the intra-PUs remains enough equilibrated for each level of R value. For R = 90, the percentage of β intra-PUs increases to 24%, the last stages of Protein Peeling process gives more PUs rich in β-strands. For this last class and this high R value, the analysis of *rf* shows many changes. A clear increase of *rf* for the couples implying two residues of opposed charges is observed, but also at a lower level, those implying residues of same charges. A diminution of *rf* for all amino acid couples implying Cysteine, as well as the aromatic ones with themselves is highlighted.

For inter-PUs, the distribution of contacts is unbalanced in favor of the inter-PUs α-α contacts (1.8 times than random values), β-β contacts (1.6) and α-other contacts (1.5). Those observations are done whatever the R value. Between Protein Units, *rf* values could have strong variations.

Contacts of Local Protein Structures. Following the same kinds of approaches, we have analyzed the interactions between classical secondary structures and between a structural alphabet letters, namely the Protein Blocks (PBs) [54, 75, 79, 88]. These latter correspond to a set of 16 local prototypes; they approximate every part of the protein structures (at the opposite of secondary structures). The PBs *m* and *d* can be roughly described as prototypes for central α-helix and central β-strand, respectively. PBs *a* through *c* primarily represent β-strand N-caps and PBs *e* and *f*, C-caps; PBs *g* through *j* are specific to coils, PBs *k* and *l* to α-helix N-caps, and PBs *n* through *p* to C-caps. Thus, we can precisely analyze the protein structures at every residue.

In a comparison purpose, analyses are firstly presented for complete protein structures. Relying on this reference, specificities observed with protein units can then be presented more clearly.

In complete proteins, at the level of the secondary structures, as expected, we observe that interactions between β-strands are the most important observed interactions (Z-score value more than 500) while their interaction with coil residues and α-helices are quite limited (Z-score value less than -200). For the α-helices, interactions with loops and β-strands are disfavored (negative Z-scores) and favored for itself. Coil residue interactions are strongly disfavored with α-helices (Z-score value of -170) and favored with coil (Z-score value of 131). Other interactions are less significant. It must be noticed that the strong values of Z-scores are also due to the limited number of states. In the same way, if one Z-score is very high and positive (*resp*. negative), the two other Z-scores must be also important and negative (*resp*. positive).

The same analyses were made with the PBs in complete proteins (see Figure 6). As expected, it is finer than with simple 3-states alphabet. The PBs mainly associated to β-strands, *i.e.*, PBs *c*, *d* and *e*, are in favorable partnership with themselves, and in general for PBs *a* to *f* with PB *d*. In compensation, these last PBs are not found associated to PB *m*, the helical PB. We must notice a very high Z-score for the association of PB *d* with itself (= 491). Computation of Kullback - Leibler asymmetric divergence measure (*KLd*) quantifies the disequilibrium of association. Hence, PB *d* has the highest *KLd* value (=0.39), *i.e.*, it is associated to a limited number of PBs, while *KLd* values are very low for PBs *g*, *k*, *l*, *n* and *o*. For PBs *k*, *l*, *n* and *o*, they are found mainly associated with PB *m*. PB *g* is the only PB with no strong favored or disadvantaged associations.

Contrary to the analysis of secondary structures, PBs analysis highlight the behaviors of some PB considered as 'loop' PBs have significant interaction with 'loop' PBs, but also with β-strand PBs. For instance, PBs *a* and *f* interacts with PB *d*; while PBs *h* and *i* have strong interactions with themselves.

PB *m*, representative of the regular α-helix is in favorable partnerships with itself (Z-score equals to 121.9). With this only favorable association, its *KLd* is highest of alpha-helical

PBs. Hence, PB analysis highlights that β-strand and α-helix structures have different behaviors. Indeed, N and C capping regions of β-strands have also an important number of contacts, while only the regular region of α-helix structures have a large number of contacts.

In Protein Units, the same studies were carried out at an early cutting stage of Protein Peeling (R=20) and at a late stage (R=90). As previously, we have looked at inter- and intra-PU contacts.

At the early stage, for the secondary structures, no drastic change in terms of Z-scores between the secondary structures within PUs is observed. It is highly similar to the observation done with the complete databank. Interestingly, between the PUs, *i.e.*, inter-PU contacts, the behaviors of α-helices are not highly discriminating. With Z-score values near 0, no particular associations are found. Hence, Protein Peeling cuts inside β-sheets while he does not favor the cut between interacted α-helices. No significant difference can be observed between parallel and anti-parallel β-sheets.

At the late, no more significant difference is found with the observation done with the databank. Protein Peeling would thus not cut out proteins on the simple criteria of the secondary structures. Examples of Figure 5 and supplementary data 1 [100] show the coloring of different PUs characterized by Protein Peeling at the late stage. The majority of PUs is composed of several distinct secondary structures. For instance yellow PUs is composed of one α-helix and one β-strand.

For the PBs, at early and late stage of Protein Peeling, the contacts within PUs are very close to the one of the databank, showing no particularities. At the opposite, inter-PU contacts have some intensity differences. At first, a strong fall of Z-scores between the PBs related to β-strands is observed. Only the Z-score of PB *c* with PB *d* and PB *d* with itself is more than 50, as PB *m* with itself. At the last stage, Z-score value increases, mainly for the PB *d*.

Nonetheless, the protein contacts of 'loops' PBs are no more significant. Thus, the decrease of early stage is mainly due to the low number of observations, while at the last stage inter-PU contacts represent half of the contacts. At the latter stage (R=90), PBs *d* is associated to PBs *a* to *f* while PB *m* only with PB *m*. The association of PB *h* and *i* is not significant in this case, contrary to the intra-PU contacts where they are found. They represent so an interesting signature of Peeling.

Discussion and conclusion

The analysis of PUs brought surprising results. First of all, during cutting, all the sizes of PUs are observed and this without influence of the protein size or its class. Moreover, it has been shown that Protein peeling is a relevant tool to characterize protein ends which can be considered as mobile for half of the proteins. Our calculations show a lower percentage of mobile ends compared to those calculated by Jacob and Unger [98]. This difference comes partly from the databank they used. Their proteins where constituted of no more than 200 residues, but especially their analyses were based on a calculation of solvent accessibility. It must be also considered that entire ends of protein are often not crystallized.

The amino acid distribution of PUs does not diverge too much from the distribution of the protein databank. Indeed, PUs generated for R = 90 have smaller length, and so the *very far* contacts [64] (distant of more than 50 residues in the sequence, they represent one third of the contacts) are poorly represented. Nonetheless, observed *rf* of PUs do not correspond to *rf* of the *near* (5 to 20 residues) and *far* (21 to 50 residues) contacts in the sequence, but to classical *rf* of the whole proteins. Only class SCOP $\alpha+\beta$ has lightly specific *rf*, potentially due to interfaces between the two "domains". In the same way, analysis of PUs in terms of secondary structures did not show any significant tendency in comparison to the entire databank. The Protein Blocks [56, 75, 85, 88, 99, 101-108] give a finer description than secondary structures [109], they so give more precise description of protein contacts and even some insights in signature of specific regions of protein loops in the PUs (*e.g.*, PBs *h* and *i*). As one limit of our analysis was low counts of contacts for some local structure category (see result section), the use of non-redundant databanks with a higher redundancy rate should be a solution for improving statistics and so analyzing more precisely inter- and intra-PUs interactions.

Hence, this study shows that cutting by Protein Peeling creates true Protein Units, a structural intermediate level between secondary structures and protein domain, having the characteristics of these latter in term of contacts features. They so can be considered as an extension of the idea of Protein Domains. Figures 7 and 8 show two representative examples of this principle. These two examples have already been used to explain the difficulty of cutting protein into domain (cf. Figure 5 of [110]). Figure 7 deals with the ubiquitin (PDB code 1ubi [111]). It is considered as only one domain by SCOP, CATH, DALI or DDOMAIN (see Figures 7a and 7b), while Protein Peeling cuts it into 3 PUs (for a *R* value of 20, see Figures 7c and 7d) or into 4 PUs (for a *R* value of 80, see Figures 7e and 7f). Figure 8 focus on the monomer of unadenylated glutamine synthetase from *Salmonella typhimurium* (PDB code 1lgr [112], see Figures 8a and 8b). SCOP considers two domains (see Figures 8c and 8d), CATH also two (see Figures 8e and 8f) and DALI four (see Figures 8g and 8h). In the last two cases, some regions are not associated to any structural domains. DDOMAIN (with the AUTHORS annotation approach) cuts it into three domains (see Figures 8i and 8j). The Protein Peeling considers 2 Protein Units for a low *R* value of 20 (see Figures 8k and 8l), 4 Protein Units for a *R* value of 80 (see Figures 8m and 8n) and 7 Protein Units for a *R* value of 90 (see Figures 8o and 8p). A detailed analysis of the position of these different domains and Protein Units gives interesting insights: SCOP and CATH assignment of two domains corresponds to the two first created PUs ($R = 20$), while the delimitation of DALI corresponds to the four PUs (R = 80). With DDOMAIN, a clear equivalence is found between first PU and first domain. But the cutting of DDOMAIN is neither encompassed by Protein Peeling, nor CATH (or another approach). These examples show that the Protein Peeling can easily cut domain into smaller sub-units.

Protein Peeling performs a linear cutting in the sequence. As noted early by Wetlaufer, some protein domains are not sequential [50]. Comparison with other related approaches such as DIAL would be also quite interesting [45]. In the same way, behavior of Protein Peeling can be compared to protein domain assignment, *e.g.* [40], however, extensive comparison is a difficult task [44]. We have showed that the distance measurement and threshold values are very sensitive parameters to define a contact [64]. Thus, a Protein Peeling based not only on Cα but taken into account other atoms could be interesting. A Voronoï approach is a good alternative that can be considered. In the same way, the research of amino acid clusters [113, 114] and atom density within the PUs could give insight to some determinants, such as the Most Interaction Residues and Tightened End Fragments [115-118]. These last researches could be directly linked to the fact that no PU corresponding to protein core could be found by the Protein Peeling approach. It means that within protein folds (after the folding process), the proteins do not present a core with so many contacts that it can be distinguished of the rest of the protein. This could be explained by the Protein Peeling methodology which proceeds to the cutting sequentially. An interesting hypothesis is also the possibility that PUs folds before the protein core formation. It can be so very interesting for the prediction of protein structures using hierarchical approach as *CombDock* [119].

Acknowledgments

This work was supported by French Institute for Health and Medical Care (INSERM) and University Denis Diderot Paris 7. AB benefits from a grant of the Ministère de la Recherche.

Figure legends

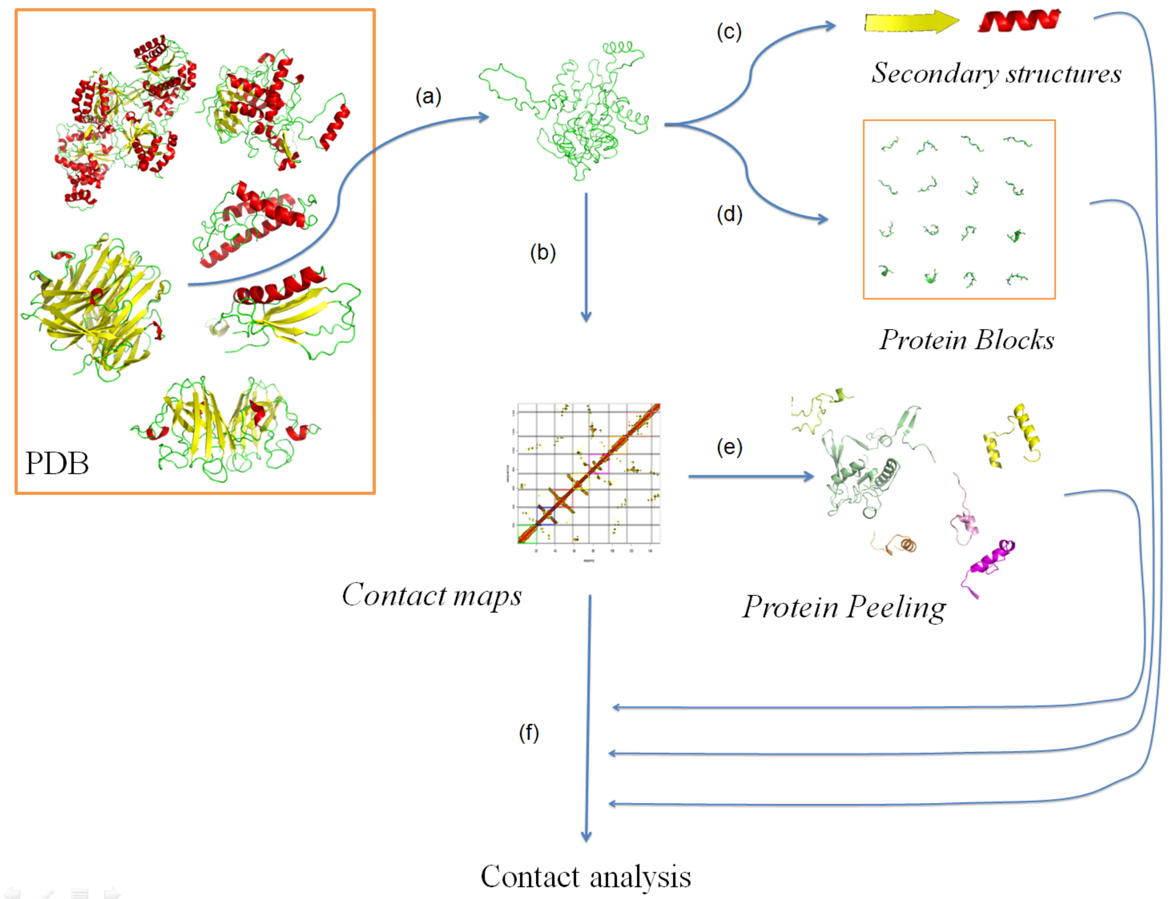

Figure 1. *Main Principle.* (a) A set of non-redundant proteins are selected from the Protein Databank [74]. (b) A contact map is computed for each protein. (c) The secondary structures and (d) the Protein Blocks are assigned. (e) The Protein Peeling gives series of Protein Units. (f) The analysis of protein contact is done using the contact maps previously computed. Different analyses are done using secondary structure and Protein Blocks assignment, and the Protein Units.

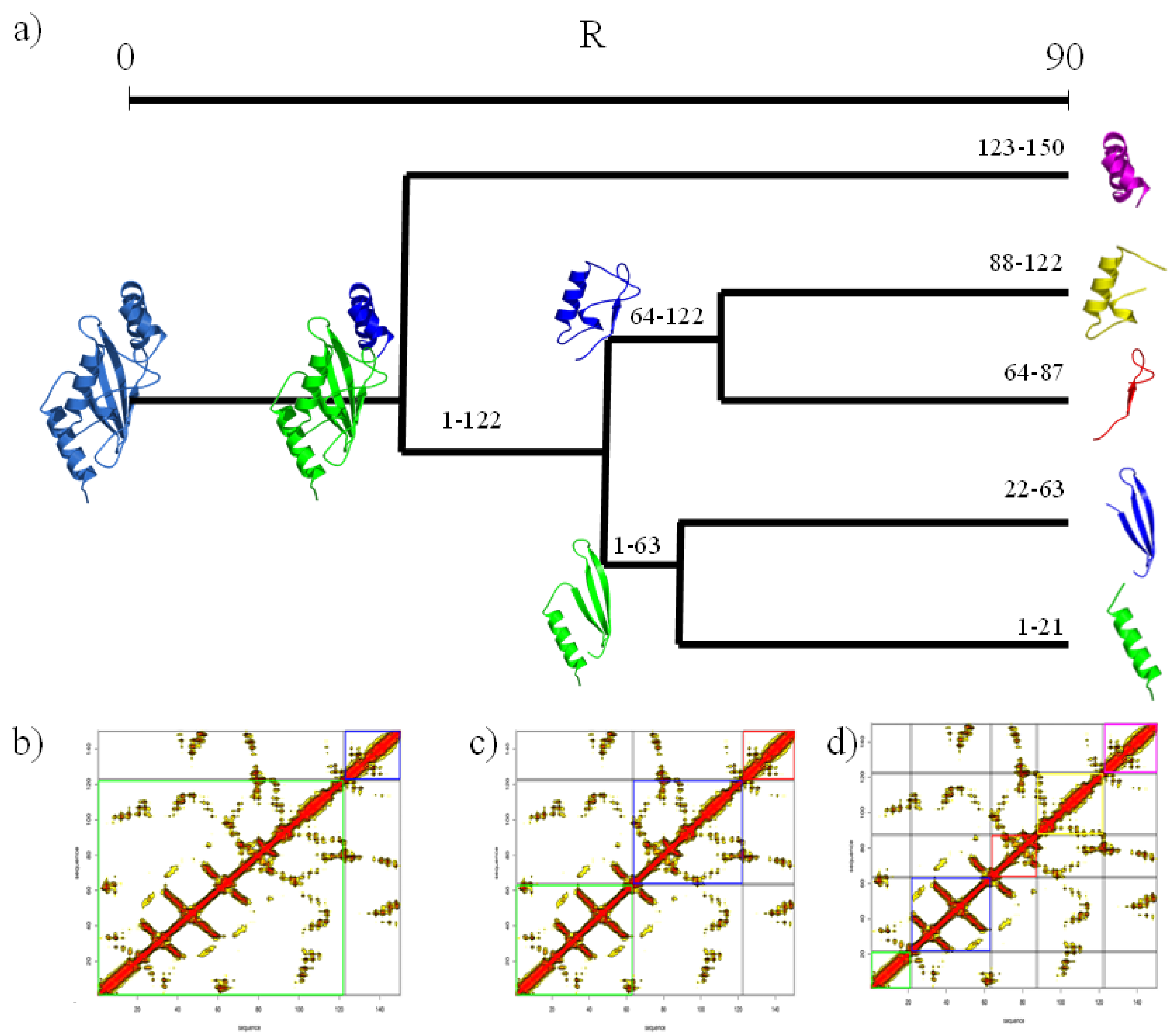

Figure 2. *Example of Protein Peeling*. (a) Protein Peeling [52, 73] of the ubiquitin conjugating enzyme from *Arabidopsis thaliana* (PDB code 2aak [120]). In a first step, the protein is cut into two Protein Units from 1-122 and the C termini region 123-150. This last is composed of two crossing helices that have few contacts. Then, in a second step, the first part of the protein is cut into two equilibrated PUs (1-63 and 64-122). The first is then cut into two PUs: 1-21 composed of a single helix, which has few contacts with the second PU, 22-63, a three layer beta-sheet. Finally, the Protein Unit 64-122 is split into two PUs: 64-87 and 88-122. (b-d) Contact matrices used for the Protein Peeling, (b) after the first cut, (c) after the second cut and (d) delimitating the 5 PUs intra- and inter-contact matrices. Protein structures are visualized with PyMol software [121].

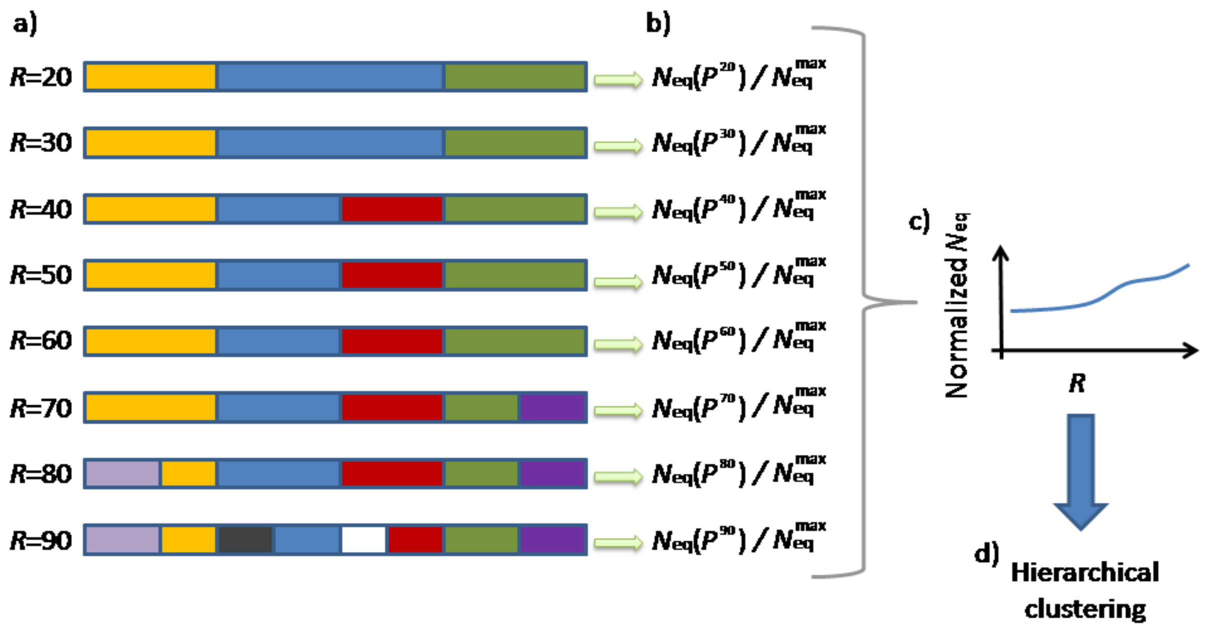

Figure 3. *Equivalent number (N_{eq}).* This Figure is an illustrative example of the computation of N_{eq} measure and how it was used to cluster N_{eq} profile. (a) Protein Peeling is performed for each protein P of the non-redundant databank and for each values of R ranging from 20 to 90. (b) a $N_{eq}(P^R)$ value is obtained; to ensure comparability, it is normalized with regards to the maximal theoretical $N_{eq}(P^R)$ value. (c) Thus, each protein P is represented by a profile of normalized N_{eq}. (d) To cluster the profiles, a hierarchical clustering is done with R software [122].

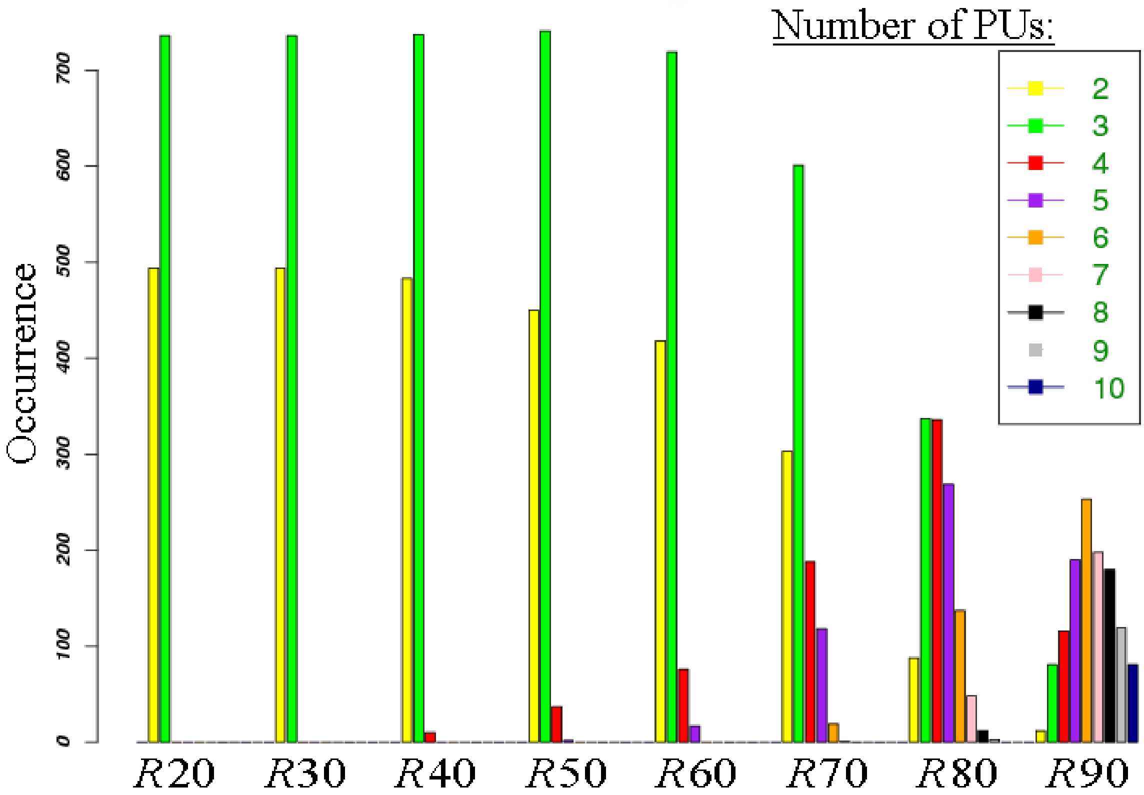

Figure 4. *Number of PUs for values of R increasing between R = 20 and R = 90*. The occurrence number has been computed on all the non-redundant databank.

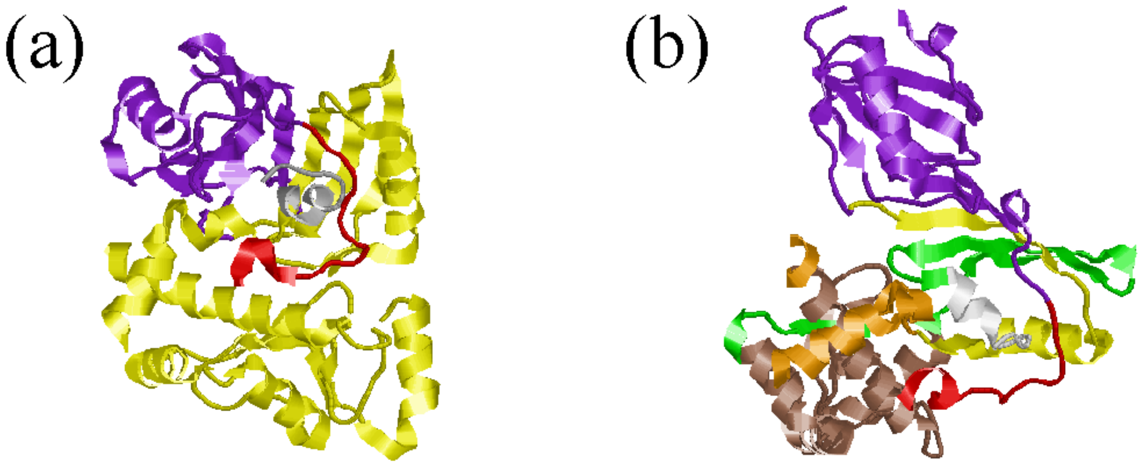

Figure 5. *Example of peeling resulting in PUs of different sizes*. The protein *Escherichia Coli* tRNA pseudouridine synthase TruD (PDB code 1SZW) is a protein of 329 amino acids [123]. It is a representative example of protein with PUs of very different sizes, several PUs are more than 50 long residues while the smallest only 20 residues long. Two levels of peeling are shown: (a) 3 PUs for R equal 20, (b) 9 PUs for R equal 90. Proteins are shown using Rasmol software [124].

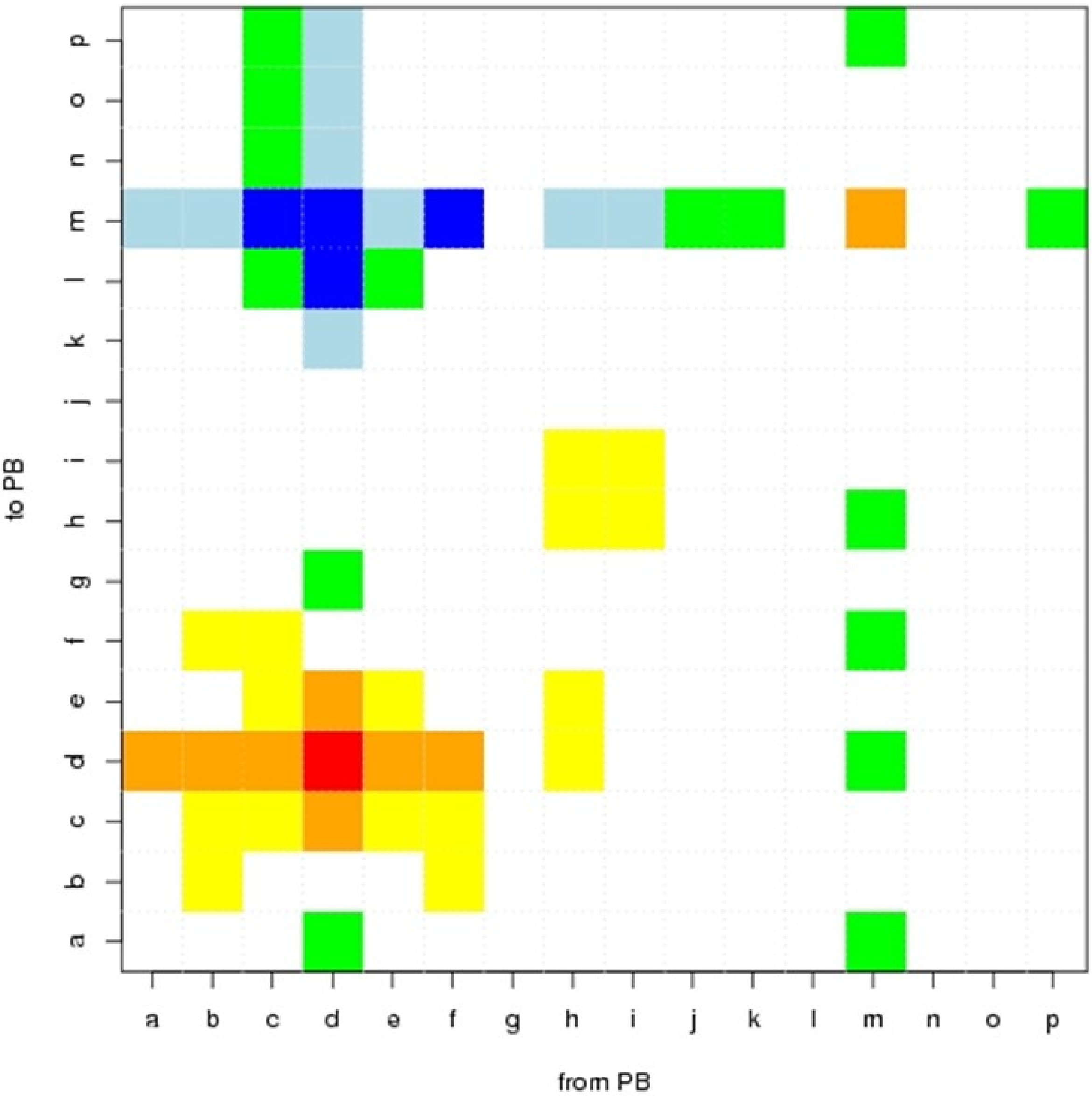

Figure 6. *Over- and Under-represented contacts between Protein Blocks.* These over- and under-representations are given in terms of Z-scores: (red): Z-score > 150, (orange): 50 <Z-score < 150, (yellow): 25 <Z-score < 50, (white): -25 <Z-score < 25, (light blue): -50 <Z-score < -150, and, (dark blue): -150 >Z-score.

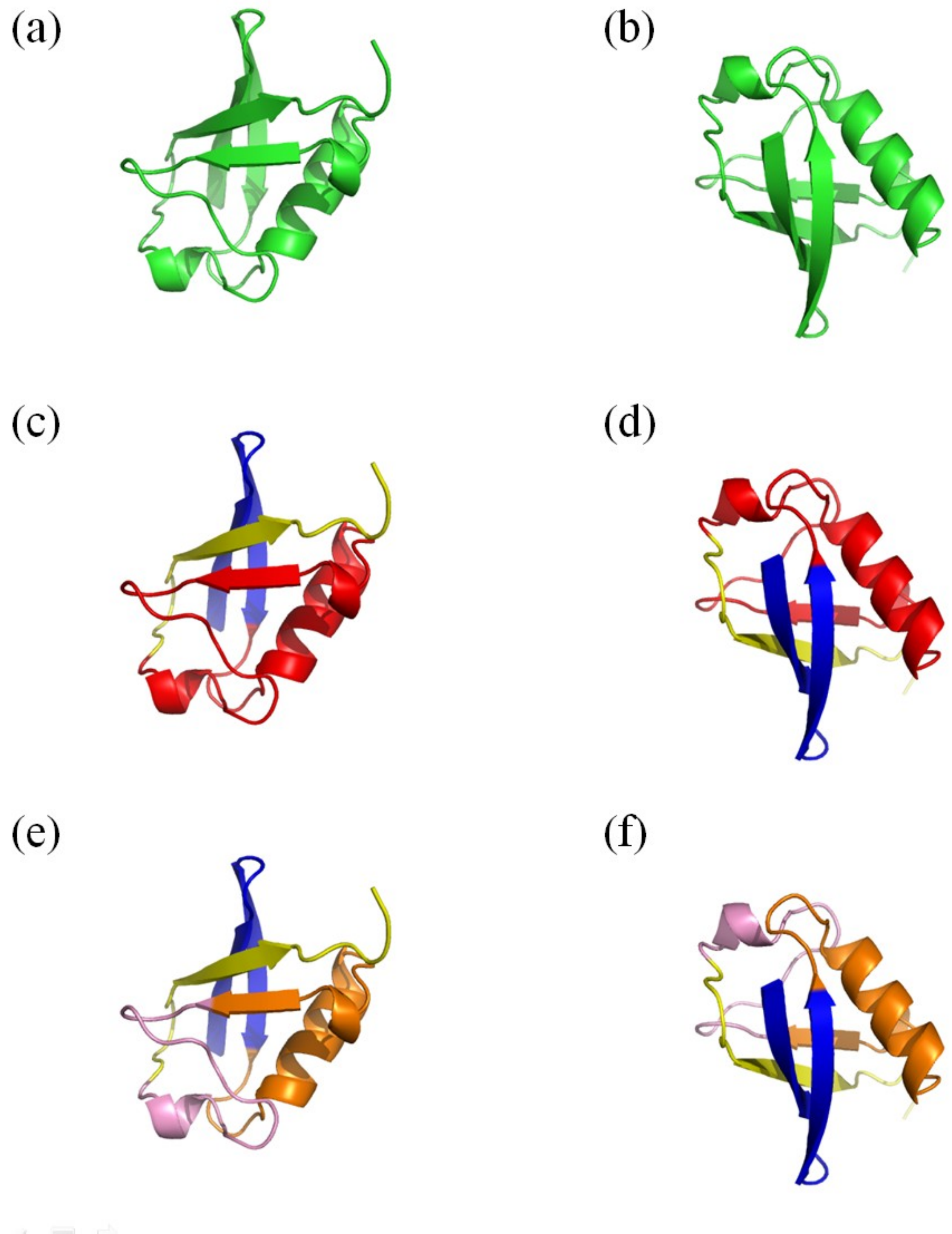

Figure 7. *Protein Domains and Protein Units of ubiquitin (PDB code 1ubi [111]).* In every case the protein is shown with two orientations. (a-b) the protein fold considered as only one domain by SCOP, CATH, DALI and DDOMAIN. Protein Peeling cuts into (c-d) 3 Protein Units for a high *R* value of 80 [1-16, 17-60, 61-76] and (e-f) 4 Protein Units for a *R* value of 20 [1-16, 17-60, 61-76].

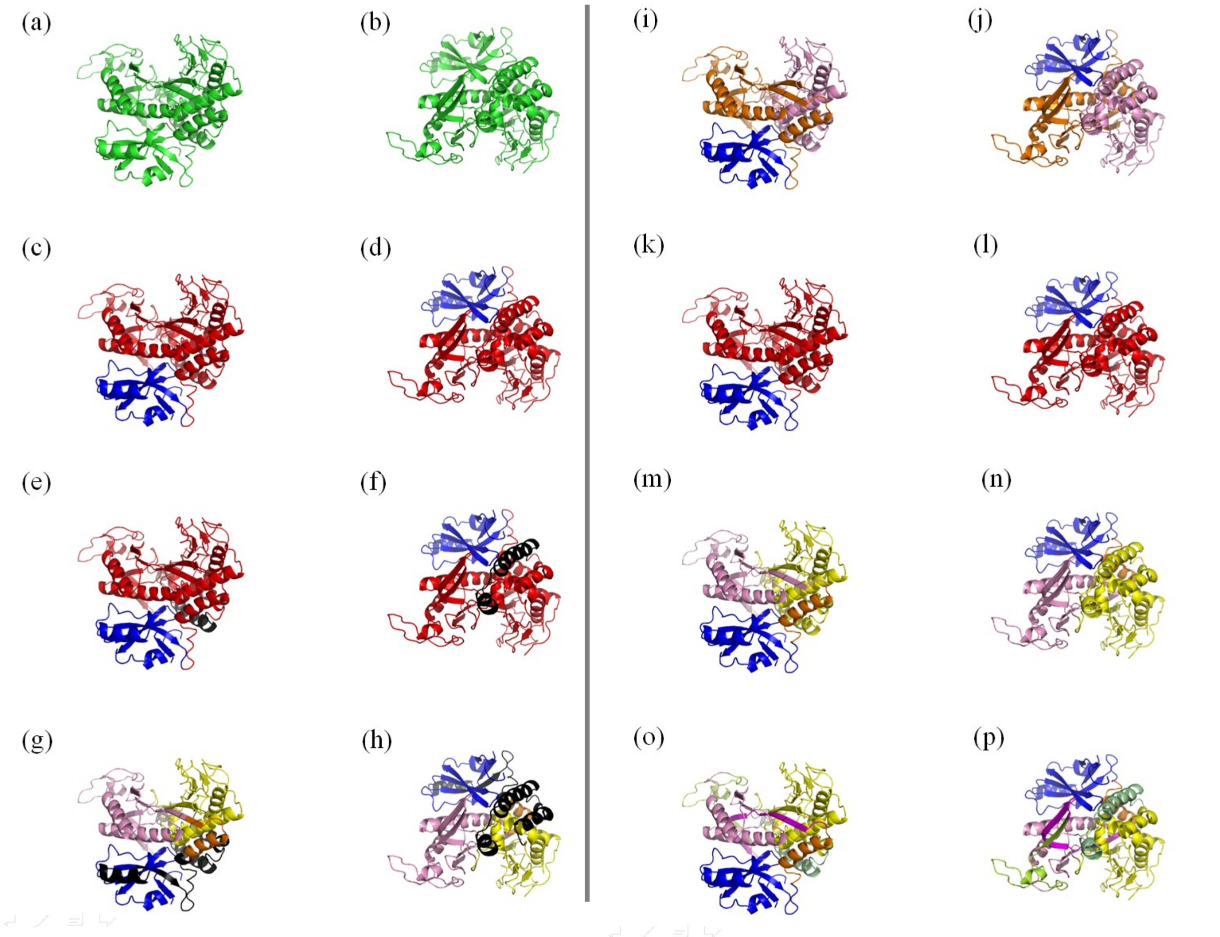

Figure 8. *Protein Domains and Protein Units of glutamine synthetase from Salmonella typhimurium (PDB code 1lgr [112]).* In every case the protein is shown with two orientations. (a-b) the protein fold. (c-d) SCOP assignment in two domains [1-94, 95-445]. (e-f) CATH assignment in two domains [1-97, 98-433]. (g-h) DALI assignment in four domains [12-86, 107-124, 125-261, 261-411]. In black are represented the structure not assigned by CATH [433-445] or DALI [1-11, 87-106, 411-445]. (i-j) DDOMAIN assignment in three domains [1-94, 95-275, 276-445]. Protein Peeling cuts into (k-l) 2 Protein Units for a low *R* value of 20 [1-99, 100-445], (m-n) 4 Protein Units for a *R* value of 80 [1-99, 100-119, 120-260, 261-445] and (m-n) 7 Protein Units for a *R* value of 90 [1-99, 100-119, 120-146, 147-190, 181-260, 261-428, 429-445].

Table 1. *Analysis of inter-PU contacts*. Are given the amino acid relative contact frequencies (*rf*) between the Protein Units of the contacts that have a *rf* variations higher than 0.1. The initial *rf* values for $C\alpha^8$ is given followed by *rf* computed for R = 20 and R = 90.

	$C\alpha^8$	*R* 20	*R* 90
[C→C]	6.14	5.36	5.55
[K→C]	1.28	0.97	1.10
[S→C]	1.48	1.15	1.40
[V→M]	1.12	1.32	1.20
[H→G]	1.19	1.39	1.28
[W→G]	1.05	1.25	1.20
[M→C]	1.55	1.35	1.67
[W→W]	1.50	1.39	1.30
[A→C]	1.32	1.13	1.27
[H→C]	1.57	1.38	1.45
[N→Y]	1.12	1.01	1.01
[D→V]	1.09	0.95	0.98
[R→G]	1.12	1.24	1.23
[D→H]	1.28	1.46	1.38
[C→K]	0.60	0.48	0.49
[W→I]	1.19	1.04	1.07
[Y→P]	1.03	1.14	1.14
[W→N]	0.86	1.00	0.94
[R→I]	1.12	0.95	1.03
[H→I]	1.09	0.92	0.99
[A→V]	1.53	1.38	1.48
[R→V]	1.25	1.09	1.16
[N→H]	1.08	1.18	1.14
[Q→P]	1.11	1.25	1.16
[E→H]	1.14	1.26	1.23
[E→K]	1.11	1.00	1.03
[E→V]	1.30	1.14	1.22
[G→H]	1.10	1.25	1.18
[I→V]	1.67	1.51	1.62
[L→V]	1.57	1.43	1.51
[F→I]	1.36	1.22	1.29
[S→P]	1.02	1.18	1.10
[T→I]	1.24	1.07	1.17
[T→V]	1.34	1.22	1.25

Table 2. *Analysis of intra-PU contacts of α+β class*. Are given the amino acid relative contact frequencies (*rf*) inside the Protein Units (of α+β SCOP class) of the contacts that have a *rf* variations higher than 0.1. The initial *rf* values for Cα[8] is given followed by *rf* computed for R = 20 and R = 90.

	Cα[8]	R20	R90
[N→N]	0.93	1.54	1.50
[C→C]	5.73	4.78	4.97
[Q→W]	1.69	1.16	1.11
[W→C]	1.90	1.34	1.38
[C→W]	1.37	0.90	0.92
[H→H]	1.70	1.25	1.11
[W→W]	0.98	1.48	1.60
[R→M]	1.11	0.90	0.83
[D→H]	1.55	1.22	1.25
[Q→C]	1.30	1.57	1.54
[E→S]	0.88	1.11	1.09
[G→H]	1.30	1.07	1.02
[H→W]	1.22	0.91	0.89
[H→Y]	1.27	1.01	1.02
[H→V]	1.05	1.27	1.33
[K→C]	1.23	1.47	1.46
[K→I]	1.40	1.19	1.18
[M→C]	1.28	1.63	1.49
[M→P]	0.72	0.97	0.92
[P→Y]	1.41	1.18	1.14
[W→Q]	1.16	0.81	0.79
[W→H]	1.17	0.89	0.85
[W→M]	0.98	1.24	1.21
[V→W]	1.09	0.88	0.89
[F→W]	1.33	1.12	1.11
[P→M]	0.86	1.10	1.03
[D→P]	0.92	1.13	1.07
[R→C]	1.33	1.13	1.16
[C→M]	0.94	1.18	1.02
[D→N]	1.16	0.97	0.95
[H→P]	0.89	1.10	1.05
[M→W]	0.96	1.16	1.18
[P→C]	1.30	1.49	1.50
[R→N]	0.68	0.86	0.89
[N→C]	1.58	1.39	1.36
[L→M]	1.25	1.08	1.05
[K→W]	1.09	0.90	0.94
[M→R]	0.83	0.64	0.63
[Y→H]	1.09	0.91	0.88

References

[1] T.L. Blundell, B.L. Sibanda, R.W. Montalvao, S. Brewerton, V. Chelliah, C.L. Worth, N.J. Harmer, O. Davies, D. Burke, Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery, Philos Trans R Soc Lond B Biol Sci 361 (2006) 413-423.

[2] O. Doppelt, F. Moriaud, A. Bornot, A.G. De Brevern, Functional annotation strategy for protein structures, Bioinformation 1 (2007) 357-359.

[3] O. Doppelt-Azeroual, F. Moriaud, F. Delfaud, A.G. de Brevern, Analysis of HSP90 related folds with MED-SuMo classification approach, Drug Design, Development and Therapy 3 (2009) 59-72.

[4] C.B. Anfinsen, The formation and stabilization of protein structure, Biochem J 128 (1972) 737-749.

[5] A.C. Clark, Protein folding: Are we there yet?, Archives of Biochemistry and Biophysics 469 (2008) 1-3.

[6] R. Santucci, F. Sinibaldi, L. Fiorucci, Protein folding, unfolding and misfolding: role played by intermediate States, Mini Rev Med Chem 8 (2008) 57-62.

[7] O.B. Ptitsyn, A.A. Rashin, A model of myoglobin self-organization, Biophys Chem 3 (1975) 1-20.

[8] J.B. Udgaonkar, R.L. Baldwin, NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A, Nature 335 (1988) 694-699.

[9] M. Karplus, D.L. Weaver, Protein folding dynamics: the diffusion-collision model and experimental data, Protein Sci 3 (1994) 650-668.

[10] S. Rackovsky, H.A. Scheraga, Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins, Proc Natl Acad Sci U S A 74 (1977) 5248-5251.

[11] A. Fersht, Nucleation mechanism in protein folding, Curr. Opin. Struct. Biol 7 (1997) 3-9.

[12] G.D. Rose, Hierarchic organization of domains in globular proteins, J Mol Biol 134 (1979) 447-470.

[13] R.L. Baldwin, G.D. Rose, Is protein folding hierarchic? I. Local structure and peptide folding, Trends Biochem Sci 24 (1999) 26-33.

[14] R.L. Baldwin, G.D. Rose, Is protein folding hierarchic? II. Folding intermediates and transition states, Trends Biochem Sci 24 (1999) 77-83.

[15] A.M. Lesk, G.D. Rose, Folding units in globular proteins, Proc Natl Acad Sci U S A 78 (1981) 4304-4308.

[16] N. Haspel, C.J. Tsai, H. Wolfson, R. Nussinov, Hierarchical protein folding pathways: a computational study of protein fragments, Proteins 51 (2003) 203-215.

[17] N. Haspel, C.J. Tsai, H. Wolfson, R. Nussinov, Reducing the computational complexity of protein folding via fragment folding and assembly, Protein Sci 12 (2003) 1177-1187.

[18] K.A. Dill, Theory for the folding and stability of globular proteins, Biochemistry 24 (1985) 1501-1509.

[19] K.A. Dill, H.S. Chan, From Levinthal to pathways to funnels, Nat Struct Biol 4 (1997) 10-19.

[20] R. Srinivasan, P.J. Fleming, G.D. Rose, Ab initio protein folding using LINUS, Methods Enzymol 383 (2004) 48-66.

[21] R. Srinivasan, G.D. Rose, LINUS: a hierarchic procedure to predict the fold of a protein, Proteins 22 (1995) 81-99.

[22] R. Srinivasan, G.D. Rose, A physical basis for protein secondary structure, Proc Natl Acad Sci U S A 96 (1999) 14258-14263.

[23] V. Daggett, Protein folding-simulation, Chem Rev 106 (2006) 1898-1916.

[24] R. Day, V. Daggett, All-atom simulations of protein folding and unfolding, Adv Protein Chem 66 (2003) 373-403.

[25] R. Day, V. Daggett, Ensemble versus single-molecule protein unfolding, Proc Natl Acad Sci U S A 102 (2005) 13445-13450.

[26] K.A. Scott, D.O. Alonso, Y. Pan, V. Daggett, Importance of context in protein folding: secondary

structural propensities versus tertiary contact-assisted secondary structure formation, Biochemistry 45 (2006) 4153-4163.

[27] K.W. Plaxco, K.T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins, J Mol Biol 277 (1998) 985-994.

[28] Y. Chen, F. Ding, H. Nie, A.W. Serohijos, S. Sharma, K.C. Wilcox, S. Yin, N.V. Dokholyan, Protein folding: Then and now Archives of Biochemistry and Biophysics 469 (2008) 4-19

[29] C.P. Ponting, R.R. Russell, The natural history of protein domains, Annu Rev Biophys Biomol Struct 31 (2002) 45-71.

[30] L. Holm, C. Sander, Parser for protein folding units, Proteins 19 (1994) 256-268.

[31] A.S. Siddiqui, G.J. Barton, Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions, Protein Sci 4 (1995) 872-884.

[32] U. Dengler, A.S. Siddiqui, G.J. Barton, Protein structural domains: analysis of the 3Dee domains database, Proteins 42 (2001) 332-344.

[33] A.S. Siddiqui, U. Dengler, G.J. Barton, 3Dee: a database of protein structural domains, Bioinformatics 17 (2001) 200-201.

[34] M.B. Swindells, A procedure for detecting structural domains in proteins, Protein Sci 4 (1995) 103-112.

[35] L. Holm, C. Sander, Dictionary of recurrent domains in protein structures, Proteins 33 (1998) 88-96.

[36] L. Wernisch, M. Hunting, S.J. Wodak, Identification of structural domains in proteins by a graph heuristic, Proteins 35 (1999) 338-352.

[37] Y. Xu, D. Xu, H.N. Gabow, Protein domain decomposition using a graph-theoretic approach, Bioinformatics 16 (2000) 1091-1104.

[38] J.T. Guo, D. Xu, D. Kim, Y. Xu, Improving the performance of DomainParser for structural domain partition using neural network, Nucleic Acids Res 31 (2003) 944-952.

[39] N. Alexandrov, I. Shindyalov, PDP: protein domain parser, Bioinformatics 19 (2003) 429-430.

[40] H. Zhou, B. Xue, Y. Zhou, DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile, Protein Sci 16 (2007) 947-955.

[41] F. Emmert-Streib, A. Mushegian, A topological algorithm for identification of structural domains of proteins, BMC Bioinformatics 8 (2007) 237.

[42] C. Anselmi, G. Bocchinfuso, A. Scipioni, P. De Santis, Identification of protein domains on topological basis, Biopolymers 58 (2001) 218-229.

[43] R.R. Joshi, A Decade of Computing to Traverse the Labyrinth of Protein Domains, Current Bioinformatics 2 (2007) 113-131.

[44] T.A. Holland, S. Veretnik, I.N. Shindyalov, P.E. Bourne, Partitioning protein structures into domains: why is it so difficult?, J Mol Biol 361 (2006) 562-590.

[45] R. Sowdhamini, T.L. Blundell, An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins, Protein Sci 4 (1995) 506-520.

[46] C.J. Tsai, R. Nussinov, Hydrophobic folding units derived from dissimilar monomer structures and their interactions, Protein Sci 6 (1997) 24-42.

[47] G. Pugalenthi, G. Archunan, R. Sowdhamini, DIAL: a web-based server for the automatic identification of structural domains in proteins, Nucleic Acids Res 33 (2005) W130-132.

[48] S.J. Wodak, J. Janin, Location of structural domains in protein, Biochemistry 20 (1981) 6544-6552.

[49] R. Sowdhamini, S.D. Rufino, T.L. Blundell, A database of globular protein structural domains: clustering of representative family members into similar folds, Fold Des 1 (1996) 209-220.

[50] D.B. Wetlaufer, Nucleation, rapid folding, and globular intrachain regions in proteins, Proc Natl Acad Sci U S A 70 (1973) 697-701.

[51] D.B. Wetlaufer, Folding of protein fragments, Adv Protein Chem 34 (1981) 61-92.

[52] J.C. Gelly, A.G. de Brevern, S. Hazout, 'Protein Peeling': an approach for splitting a 3D protein

structure into compact fragments, Bioinformatics 22 (2006) 129-133.

[53] B.W. Matthews, X-ray crystallographic studies of proteins, Annu. Rev. Phys.Chem 27 (1976) 493-523.

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

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

[56] C. Etchebest, C. Benros, S. Hazout, A.G. de Brevern, A structural alphabet for local protein structures: improved prediction methods, Proteins 59 (2005) 810-827.

[57] S. Hazout, Entropy-derived measures for assessing the accuracy of N-state prediction algorithms., in: de Brevern A.G. (Ed.), Recent Advances in Structural Bioinformatics, Research signpost, Trivandrum, India, 2007, pp. 395-417.

[58] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, V.A. Voelz, The protein folding problem: when will it be solved?, Curr Opin Struct Biol 17 (2007) 342-346.

[59] L. Zhang, J. Skolnick, How do potentials derived from structural databases relate to "true" potentials?, Protein Sci 7 (1998) 112-122.

[60] S. Miyazawa, R.L. Jernigan, Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, J Mol Biol 256 (1996) 623-644.

[61] D. Baker, A surprising simplicity to protein folding, Nature 405 (2000) 39-42.

[62] E. Paci, K. Lindorff-Larsen, C.M. Dobson, M. Karplus, M. Vendruscolo, Transition state contact orders correlate with protein folding rates, J Mol Biol 352 (2005) 495-500.

[63] A.D. Pandit, A. Jha, K.F. Freed, T.R. Sosnick, Small proteins fold through transition states with native-like topologies, J Mol Biol 361 (2006) 755-770.

[64] G. Faure, A. Bornot, A.G. de Brevern, Protein contacts, inter-residue interactions and side-chain modelling, Biochimie 90 (2008) 626-639.

[65] L. Brocchieri, S. Karlin, How are close residues of protein structures distributed in primary sequence?, Proc Natl Acad Sci U S A 92 (1995) 12136-12140.

[66] M.J. Bower, F.E. Cohen, R.L. Dunbrack, Jr., Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool, J Mol Biol 267 (1997) 1268-1282.

[67] R.L. Dunbrack, Jr., M. Karplus, Backbone-dependent rotamer library for proteins. Application to side-chain prediction, J Mol Biol 230 (1993) 543-574.

[68] R.L. Dunbrack, Jr., F.E. Cohen, Bayesian statistical analysis of protein side-chain rotamer preferences, Protein Sci 6 (1997) 1661-1681.

[69] C. Hartmann, I. Antes, T. Lengauer, IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models, Protein Sci 16 (2007) 1294-1307.

[70] E. Eyal, R. Najmanovich, B.J. McConkey, M. Edelman, V. Sobolev, Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins, J Comput Chem 25 (2004) 712-724.

[71] A.A. Canutescu, A.A. Shelenkov, R.L. Dunbrack, Jr., A graph-theory algorithm for rapid protein side-chain prediction, Protein Sci 12 (2003) 2001-2014.

[72] J. Xu, Rapid side-chain prediction via tree decomposition., RECOMB, 2005.

[73] J.C. Gelly, C. Etchebest, S. Hazout, A.G. de Brevern, Protein Peeling 2: a web server to convert protein structures into series of protein units, Nucleic Acids Res 34 (2006) W75-78.

[74] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res 28 (2000) 235-242.

[75] A.G. de Brevern, New assessment of Protein Blocks, In Silico Biology 5 (2005) 283-289.

[76] T. Noguchi, Y. Akiyama, PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003, Nucleic Acids Res 31 (2003) 492-493.

[77] T. Noguchi, H. Matsuda, Y. Akiyama, PDB-REPRDB: a database of representative protein chains from

the Protein Data Bank (PDB), Nucleic Acids Res 29 (2001) 219-220.

[78] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442-451.

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

[80] C. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379-423.

[81] S.J. Hubbard, J.M. Thornton, 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London., 1993.

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

[83] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J Mol Biol 247 (1995) 536-540.

[84] A.D. Michie, C.A. Orengo, J.M. Thornton, Analysis of domain structural class using an automated class assignment protocol, J Mol Biol 262 (1996) 168-185.

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

[86] A.G. de Brevern, H. Valadie, S. Hazout, C. Etchebest, Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship, Protein Sci 11 (2002) 2871-2886.

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

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

[89] C. Etchebest, C. Benros, A. Bornot, A.-C. Camproux, A.G. de Brevern, A reduced amino acid alphabet for understanding and designing protein adaptation to mutation, European Biophysics Journal 36 (2007) 1059-1069.

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

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

[92] T. Kohonen, Self-Organizing Maps (3rd edition), Springer, 2001, 501 p.

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

[94] S. Kullback, R.A. Leibler, On information and sufficiency, Ann Math Stat 22 (1951) 79-86.

[95] R.B. Russell, M.A. Saqi, P.A. Bates, R.A. Sayle, M.J. Sternberg, Recognition of analogous and homologous protein folds--assessment of prediction success and associated alignment accuracy using empirical substitution matrices, Protein Eng 11 (1998) 1-9.

[96] T. Laborde, M. Tomita, A. Krishnan, GANDivAWeb: a web server for detecting early folding units ("foldons") from protein 3D structures, BMC Struct Biol 8 (2008) 15.

[97] J. Sammon, J. W. , A nonlinear mapping for data structure analysis, IEEE Transactions on Computers 18 (1969) 401-409.

[98] E. Jacob, R. Unger, A tale of two tails: why are terminal residues of proteins exposed?, Bioinformatics 23 (2007) e225-230.

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

[100] K. Volz, P. Matsumura, Crystal structure of Escherichia coli CheY refined at 1.7-A resolution, J Biol Chem 266 (1991) 15511-15519.

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

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

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

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

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

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

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

[108] A. Bornot, C. Etchebest, A.G. de Brevern, A new prediction strategy for long local protein structures using an original description, Proteins (2009) in press.

[109] A. Bornot, A.G. de Brevern, Protein beta-turn assignments, Bioinformation 1 (2006) 153-155.

[110] R. Day, D.A. Beck, R.S. Armen, V. Daggett, A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary, Protein Sci 12 (2003) 2150-2160.

[111] R. Ramage, J. Green, T.W. Muir, O.M. Ogunjobi, S. Love, K. Shaw, Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin, Biochem J 299 (Pt 1) (1994) 151-158.

[112] S.H. Liaw, G. Jun, D. Eisenberg, Interactions of nucleotides with fully unadenylylated glutamine synthetase from Salmonella typhimurium, Biochemistry 33 (1994) 11184-11188.

[113] D. Naor, D. Fischer, R.L. Jernigan, H.J. Wolfson, R. Nussinov, Amino acid pair interchanges at spatially conserved locations, J Mol Biol 256 (1996) 924-938.

[114] Z. Dosztanyi, A. Fiser, I. Simon, Stabilization centers in proteins: identification, characterization and predictions, J Mol Biol 272 (1997) 597-612.

[115] I.N. Berezovsky, A.Y. Grosberg, E.N. Trifonov, Closed loops of nearly standard size: common basic element of protein structure, FEBS Lett 466 (2000) 283-286.

[116] I.N. Berezovsky, E.N. Trifonov, Loop fold nature of globular proteins, Protein Eng 14 (2001) 403-407.

[117] M. Lamarine, J.P. Mornon, N. Berezovsky, J. Chomilier, Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding?, Cell Mol Life Sci 58 (2001) 492-498.

[118] N. Papandreou, I.N. Berezovsky, A. Lopes, E. Eliopoulos, J. Chomilier, Universal positions in globular proteins, Eur J Biochem 271 (2004) 4762-4768.

[119] Y. Inbar, H. Benyamini, R. Nussinov, H.J. Wolfson, Protein structure prediction via combinatorial assembly of sub-structural units, Bioinformatics 19 Suppl 1 (2003) i158-168.

[120] W.J. Cook, L.C. Jeffrey, M.L. Sullivan, R.D. Vierstra, Three-dimensional structure of a ubiquitin-conjugating enzyme (E2), J Biol Chem 267 (1992) 15116-15121.

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

[122] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, J Comput Graph Stat 5 (1996) 299-314.

[123] U.B. Ericsson, P. Nordlund, B.M. Hallberg, X-ray structure of tRNA pseudouridine synthase TruD reveals an inserted domain with a novel fold, FEBS Lett 565 (2004) 59-64.

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