Biochimie 91 (2009) 876–887

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

# Analysis of protein contacts into Protein Units

Guilhem Faure [a,1,2], Aurélie Bornot [a,b,1], Alexandre G. de Brevern [a,b,*]

[a] 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
[b] 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

ARTICLE INFO

*Article history:*
Received 24 July 2008
Accepted 13 April 2009
Available online 19 April 2009

*keywords:*
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 implications 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

* Corresponding author. INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot - Paris 7, INTS, 6 rue Alexandre Cabanel, 75739 Paris cedex 15, France. Tel.: +33 1 44 49 30 38; fax: +33 1 47 34 74 31.
*E-mail address:* alexandre.debrevern@univ-paris-diderot.fr (A.G. de Brevern).
[1] Both authors contribute equally to this work.
[2] 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.

0300-9084/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2009.04.008

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 maximize 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 Matthews' coefficient correlation (MCC) [53] between contact submatrices, 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

Fig. 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 Fig. 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 analyses 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 1736 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 1230 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 Fig. 2 for an example). A refinement of