Chapter 12

Geometry and Topology in Protein Interfaces -- Some Tools for Investigations

Giovanni Feverati

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/58420

1. Introduction

The present work is motivated by the biological problem of understanding and possibly predicting the assembly of biological molecules, in particular proteins. This is one of the most common processes in living cells thus it is essential to understand its key aspects, especially in relation to the implication in several pathologies, from bacterial infections (cholera, anthrax, ...) to protein misfolding diseases (Alzheimer, Parkinson, ...) [1–4]. The stable association of different subunits requires the formation of specific intermolecular bonds, thus constituting what is called an interface. Unfortunately, in spite of extensive analyses, the identification of the patterns, in the polypeptidic chain, responsible for the establishment of an interface remains difficult.

Geometry has developed the ability to measure and characterize complex shapes but it is not a priori obvious that it may also reveal important aspects of the interactions. To understand this point, let consider the following examples.

The main geometrical elements of the modern Golden Gate Bridge in S. Francisco and the ancient Roman aqueduct bridge Pont du Gard, in the Gard department in France, are arcs. But with a main difference: the three arcs that form the suspension system of the Golden Gate Bridge are concave upward while the many arcs that form the Pont du Gard are concave downward. Indeed, in the first case the arcs resist to longitudinal tension while in the second case resist to longitudinal compression. Stones are unsuited to resist to strong tension, while they perfectly resist to huge compression. Thus, architectural elements that have to undergo strong tensions are made of wood or steel, but not of stone. Notice that the simple observation of the geometrical form, the concave upward or downward aspect of the bridges, has lead us to understand the basic interactions and formulate constraints on the possible choices of materials. This argument can be pushed much forward: the structural analysis in architecture and engineering largely rely on euclidean geometry (diagrams of forces are diagrams of vectors). Even if the elastic properties of construction materials and the action of gravity are important, the main ingredient in studying the equilibrium of forces is geometry.

A second example is taken from Einstein's general relativity, where the notion of gravitational interaction and the space-time geometry are fully identified by the equivalence principle and the Einstein's field equations. The equivalence principle was first formulated by A. Einstein in 1907, when he recognized that the local behaviour of falling bodies is equivalent to the effect of being in an accelerated reference system (this holds for local effects only). The Einstein's field equations (1915) provide with a mathematical formulation of the principle. In summary, spatial and temporal distances fully inform on the gravitational interaction.

This connection between geometry and interactions also works at the atomic scale. The perfectly planar and hexagonal symmetric form of benzene molecules, compared to the non planar and less symmetric cyclohexane molecule, is clear indication of the different nature of the corresponding Carbon-Carbon bonds and of the sp2 or sp3 hybridization respectively.

In the early 50's, F. Crick [5] observed that the formation of the coiled-coil protein interface is due to the appropriate geometrical and chemical complementarity of the two interacting domains, as in a lock and key mechanism. The key has a particular geometrical form combined to some contact points which together provide it the capacity to associate to one lock.

Moving on from all these examples, in our group we have developed tools to investigate the geometry of the interfaces in multi-chain proteins. In essence, we measure shapes and compare measures between different proteins. Our input data are protein atomic positions, as those from the Protein Data Bank (PDB) repository of protein structural data.

In particular, we compare protein interfaces of similar geometrical form. From the previous examples, there is no surprise in claiming that the geometry will provide information on the interactions at the interfaces. In our previous publications [6–8], careful statistical analyses have been performed and have led to formulate constraints on the amino acid sequences and atom pairs that are compatible with a given geometrical form. The long term perspective of our work is to rationalize the interface, namely to establish a clear understanding of its sequence-structure relationship, in order to develop interface prediction tools and help to advance in interface design.

1.1. Basic information on interfaces

The shape and the function of proteins are normally encoded within their sequences, i.e. in their amino acid compositions but it is not yet possible, by simply reading the primary sequence of a protein, to predict its three-dimensional structure or the quaternary organization, in the case of an oligomeric protein. One of the difficulties is the non linear encoding of the information in the sequence, namely the fact that the three-dimensional structure is often generated and stabilized by bonds between residues that are not contiguous or even are very far apart, along the chain. Another difficulty is due to the degeneracy between sequences and structures, consisting in the observation that several sequences can code for the same shape, that indicates a versatile role of the amino acids. The secondary structures of proteins which are mainly composed of α helices, β structures and loops are partially understood, and several prediction programs are now available. Prediction of 3D structures is mainly based on homology, namely comparison of sequences that have similar three-dimensional elements. A rich collection of prediction tools is available on [9]. In some

http://dx.doi.org/10.5772/58420

cases, the prediction also takes into account the known geometrical constraints present in amino acids.

Oligomeric proteins associate by forming an interface. Various descriptions of the interface have been proposed.

The simplest definition of interface between two adjacent polypeptidic chains A and B is provided by selecting the set of pairs of atoms, one from each chain, whose distance is lower than a given cut-off, typically fixed near 0.5 nm (cut-off interface). This definition does not provide any measure to distinguish pairs. As such, it provides little information since firstly physical interactions decrease when distance grows, secondly two interactions of equal strength may not play the same role if they are in different parts of the molecules, inserted in different local atomic environments. In another definition, the interface is identified to the surface buried between the two components A and B, namely to those atoms that belong to the surface of A and B and that loose solvent accessibility once the complex AB is formed [10] This makes use of the Van der Waals atomic radii, and leads to distinguish a rim (exposed to the solvent) from a core (inaccessible to the solvent). The interface can be defined also by constructing the Voronoi α-complex [10], namely the set of Voronoi restricted balls. The construction follows a precise mathematical procedure, and determines the volume in which an atom interacts more than its neighbours.

Both the buried surface and Voronoi restricted balls methods focus on the volume of the atoms and the importance of the specific chemical properties of each atom. They make use of a cut off and describe the interactions by using the Van der Waals radius. Differently from these descriptions of the interface, we felt the need to develop a stronger analysis of the structural organization of a protein interface, in order to evaluate the specific role of each residue and the rules of pairings.

In [11], we shown that many aspects of the structural organization of a protein interface can be effectively described by a graph, namely the ensemble of nodes and edges, constructed following the precise geometrical analysis of the three-dimensional structure of the interface known as symmetrization or symmetric minimization of distances. The algorithm and the graph theory terms are described in Methods. The graph describes how the different atoms are connected. In fact, among its edges one recognizes the known hydrogen bonds present at the interface, that are obtained as a bonus, because the symmetric minimization does not make use of them (see Methods). An example of interaction graph is given in Figure 1.

Statistical analyses have been performed on the case of the β interfaces, that are formed by two adjacent β strands, one from each subunit [6–8]. In [8] the analysis has been extended to a dataset of 755 proteins. It is known that there are three possible orientations of the adjacent β strands: they can be anti-parallel (by far, this is the most common case), parallel or oblique. The latter actually includes all the cases that do not enter into the previous ones, for example perpendicular or oblique β strands. The most significant results of these statistical analyses are summarized here (please refer to the Methods for the precise definition of the motifs).

- Two typical interaction graphs have been observed, one for the parallel and one for the anti-parallel orientation. The anti-parallel case shows a BB graph of type ladder, were rungs are typically spaced of 2 amino acids. The parallel case shows a BB graph of type zigzag, in which one recognizes a separation of 2 amino acids in each oscillation of the

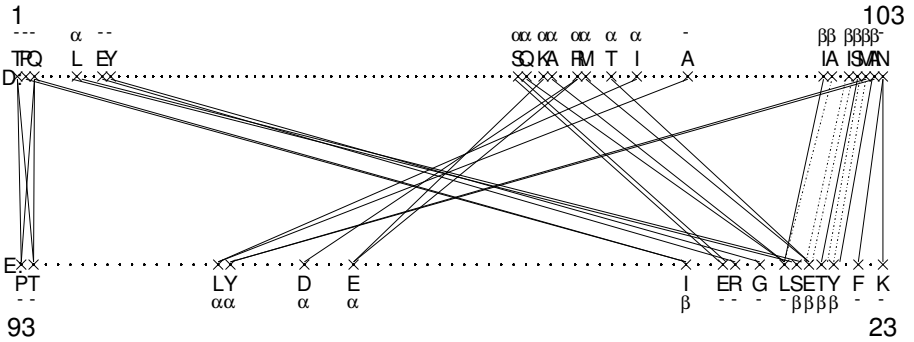

Figure 1. Full interaction graph of the level 0 of the protein 1EEI interface (see Methods). The upper horizontal line represents the sub-unit D, the lower one the sub-unit E. With the crosses we indicate the residues that participate to the interaction graph; their name and membership to specific secondary structures is indicated, when available. The dots represent the residues that do not participate to the interface. The dotted-dashed lines represent pairs of atoms both from the backbone of the residues (BB graph). The solid lines indicate that at least one atom of the pair is from the side chain (SC graph).

zigzag. In some cases, the zigzag topology reduces to a simple vertex V (defined in Methods).

Both these specific topologies identified in the BB graph correspond to the known hydrogen bond graph between backbone atoms. While it is not surprising to find them in the BB graph, the surprise comes from the fact that the position of hydrogen atoms has not been used as input by our algorithms (in most cases it is not even given in the PDB files). Thus, the symmetric minimization algorithm is able to reconstruct the backbone hydrogen bonds network by the unique input of the backbone atoms N, C_α, C, O coordinates, with an accuracy of 90%. In other words, the backbone hydrogen bonds satisfy the mathematical property of symmetrically minimized distances and the backbone non-hydrogen atoms that engage in BB hydrogen bonds are reciprocal nearest neighbours, with the indicated accuracy. This identification has been obtained using the web server RING (see Methods) to calculate the hydrogen bonds and compare with the graphs (see also [6]). Among the 120 pairs of residues in the graphs (anti-parallel and parallel cases only), 108 are recognized by RING as hydrogen bonds and 12 are not. Thus, the symmetric minimization recognizes hydrogen bonds with accuracy of 90%.

- Interfaces of the oblique family have very small or absent BB graph, thus the two BB graphs are rather specific to the respective anti-parallel and parallel cases. It is also rare to find BB graphs in non β interfaces.

  In fact, in these cases the BB graph is intra-chain, namely it develops between atoms of the same sub-unit and is almost absent in the interface.

- The amino acids are not randomly paired. Rather, the frequency with which residues are connected in the interaction graph clearly deviates from the expected frequency calculated from the average frequency of residues in the interface. This indicates that the edges in the interaction graph form in order to provide to the interface very specific features.

http://dx.doi.org/10.5772/58420

2. Methods

2.1. Symmetric minimization of distances

The description of the interface, that has been developed starting with [11] and that will be used in this paper, was introduced to help extract the structural organization of the interface. It focuses on the way atoms pair, keeping into account the local connectivity, namely the possibility that atoms interact with other atoms, according to their local arrangement. It is based on the notion of symmetric minimization of distance pairs, defined by the symmetric minimization algorithm, presented here in pseudocode. The flow chart is given in Figure 3. The needed mathematical explanations and demonstrations have been provided in [12].

Input: $A \leftarrow \{$ atoms of the first subunit $\}$
$B \leftarrow \{$ atoms of the second subunit $\}$
$d(a,b) \leftarrow$ A metric ; typically, the inter-atomic distance is used

Start: $R_0 \leftarrow \{(a,b) : a \in A, b \in B\}$
$i \leftarrow 0$
while $R_i \neq \{\ \}$ **do:**
$\min_A(a) \leftarrow \min\left\{d(a,b') : (a,b') \in R_i \text{ and } b' \in B\right\}$
$\min_B(b) \leftarrow \min\left\{d(a',b) : (a',b) \in R_i \text{ and } a' \in A\right\}$
$L_A \leftarrow \left\{(a,b) \in R_i : d(a,b) = \min_A(a)\right\}$
$L_B \leftarrow \left\{(a,b) \in R_i : d(a,b) = \min_B(b)\right\}$
$S_i \leftarrow L_A \cap L_B$
$R_{i+1} \leftarrow R_i - S_i$
$i \leftarrow i + 1$
end while

Output: $S_0, \quad S_1, \quad S_2, \quad \ldots$

The sets R_0, R_i are sets of edges. The empty set is indicated with $\{\ \}$. The symbols $-$ and $\cap$ indicate set difference and set intersection respectively. The symbols $\min_A(a)$, $\min_B(b)$ indicate the functions that give the shortest distances in the neighbourhood of the indicated point. The symbols L_A, L_B indicate the shortest edges relative to the given iteration R_i. The reciprocal shortest edge sets are indicated with S_i, and called symmetrized levels i.

In summary, the symmetric minimization is a recursive method that, at the first iteration, defines the lowest level of the interface as the pairs of atoms that are reciprocal nearest neighbours. The nearest neighbour condition must be verified for both the two subunit atoms, as the name itself suggests. These pairs form the lowest level S_0, called symmetrized interface. Please see the caption of Figure 2.

The algorithm can be repeated on all the edges that have not been retained at the lowest level. This produces a new set of reciprocally shortest edges, not contained in S_0, that forms the symmetrized set S_1, named level 1. The repetition of the algorithm up to exhaustion of all the atom pairs, provides a hierarchy of levels

$$S_0, \quad S_1, \quad S_2, \quad \ldots, \quad S_M \tag{1}$$

that define the symmetrized interface (SI). Thus, inter-atomic distances between the two sub-units are ranked according to levels. Notice that, as proteins are of finite and not of infinite size, there must be a maximum level. Also, the algorithm is free of ambiguities, even in case of regular structures. In Figure 1 and in previous papers the lowest level only was analysed, while it is a purpose of this paper to start the investigation of the higher levels.

The symmetric minimization of distances is a variant of the case $k = 1$ of the known k-reciprocal nearest neighbour method (kRNN), discussed in [13] and used now in the domain of hierarchical classification and object retrieval in images. A modern presentation is in [14]. Actually, the lowest level S_0 could be equivalently obtained with both methods but at higher levels the equivalence breaks down. The purpose of kRNN is to assert the relative proximity of several images containing the same object appearing in different scenes. Each image is a point in some very high-dimensional space. Using an appropriate metric, the closest images are found and agglomerated in a cluster. At the new iteration, new images will join the cluster. The method kRNN compares high-dimensional vectors. It can be applied to atomic coordinates. Once two atoms are found to be reciprocal nearest neighbours at some iteration, they are removed from the pool (and put in a cluster) before the next iteration could start. On the other hand, in the symmetric minimization, two atoms that are reciprocal nearest neighbours are not removed from the pool: the edge they form is removed but the atoms remain. This is the role of the sets R_i in the algorithm: at the next, the same atoms may be nearest neighbours with others. In kRNN, the sets of edges R_i would simply be replaced by some A_i, B_i where A_i or B_i would be obtained by removing from the initial A and B the atoms found at each iteration. The choice of working with edges comes from the goal of describing the interactions from a geometrical point of view. Moreover, the binding energy in a protein interface accumulates between all pairs of atoms, at least in a suitable range of distances, no matter if they are nearest neighbours or not. Using edges, the information about all the neighbours of an atom is recorded and used at one or other of the levels. Using atoms, part of the edges are not evaluated and some information seems lost, at least for purposes related to protein structure.

The classification of pairs into levels reminds one of perturbative calculations, very common in physics, where the lowest order contains the strongest interaction and the higher orders introduce weaker and weaker terms. The ranking, and the set of levels in equation (1), have been computed on the basis of inter-atomic distances: the higher the level, the larger the distance between atoms. Physically speaking, moving to higher distances implies a tendency to move to weaker interactions. Here, force fields and types of atoms have not been used in the symmetric minimization, thus rising to higher rank only indicates a tendency to weaker interactions and does not hold in a strict sense.

Imagining a mechanical model of balls and sticks, and a quantity of glue, can one mount a human size model of the protein interface? Yes, if the sequence in equation (1) is followed.

http://dx.doi.org/10.5772/58420

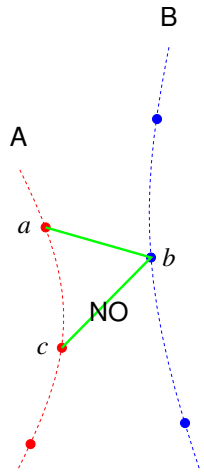

Figure 2. Example of symmetric minimization. The distance (ab) is the smallest for the atom a and, at the same time, is the smallest for b. Thus, the edge (ab) is retained in the symmetrized interface. The distance (bc) is the smallest for (c) but not for (b) thus it is not retained.

Otherwise said, the geometrical action of ranking the pairs of an interface corresponds to the sequence of actions that one has to follow to construct the interface with minimum amount of steric effects. The sequence read in the opposite sense, from the maximum level S_M to S_0, indicates the steps to access the closest atoms of the interface by removing one layer after the other.

It is important to stress that each edge of S_0 behaves as a nucleation centre, or a bud, because every edge that appears at level 1 is attached to a level 0 edge; thus, level 0 edges act as the starting points of a growth process where, at each higher level, new edges attach to the bud. The growth occurs simultaneously from the various level 0 edges. Each bud is technically a cluster. At some level, an edge appears that joins two different buds; the edge length is the resolution threshold for the two buds. Certain edges do not start a growth process; this occurs when the atoms joined by these edges are not further connected to the rest of the interface. Also, a bud must be present at level 0 and cannot appear at higher level: S_0 already contains all the buds. In other words, the various parts of the interface are already contained in S_0 and a set smaller than S_0 is possibly insufficient to reconstruct the full interface. This provides the mathematical justification for calling S_0 a framework of the interface. Indeed, as we have empirically observed in our previous publications, S_0 is the smallest set that can describe the interface. This is the most important point of the whole construction and is based on the theorems proved in [12] but is published here for the first time.

In summary, the symmetric minimization has been introduced to responds to the following needs and possesses the following features [12].

Scalable. The symmetric minimization may be applied to proteins or objects of any size, without size limits. Moreover, it can be applied also to objects at the human or interstellar scale.

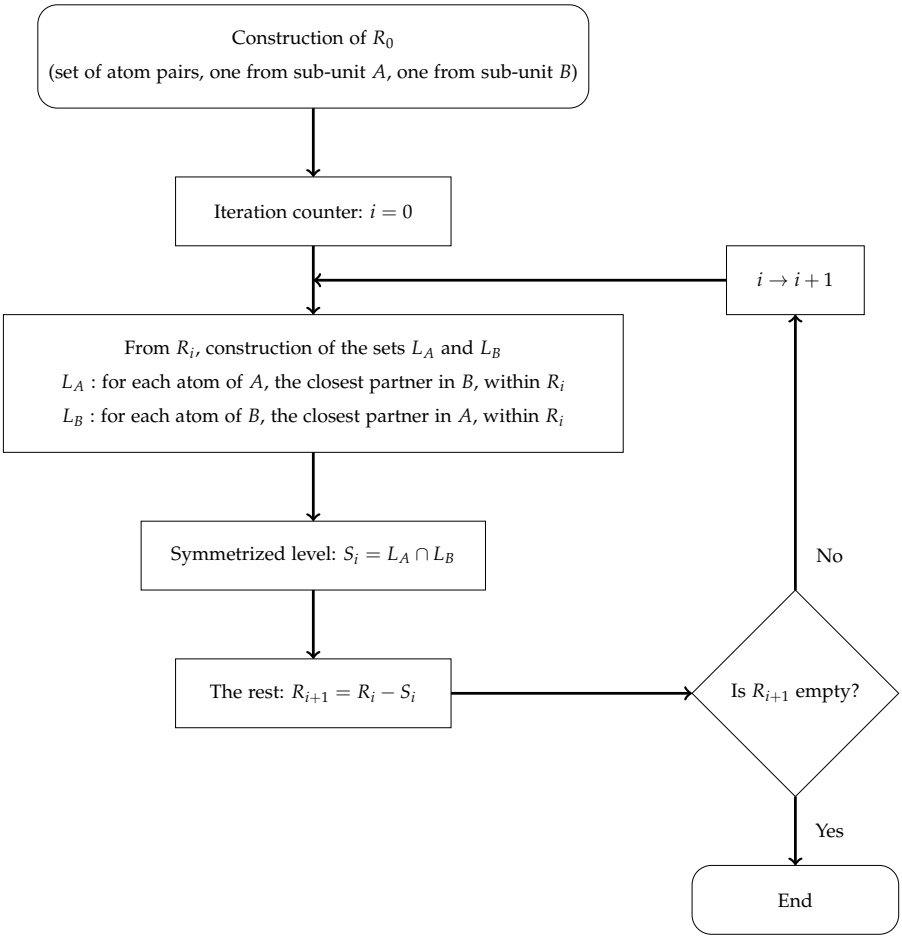

Figure 3. Flow chart of the symmetric minimization algorithm.

Local. It is based on the local arrangement of atoms (or points) and not on global features. So, it captures the differences occurring in situations like a dense atomic packing or a dilute packing.

Scale-free. No length scale has been imposed from outside.

Intrinsic scales. It defines a set of characteristic scales, intrinsic to the interface itself. Indeed, the symmetric minimization allows to divide the interface in clusters (the buds previously presented). The first edge that joins two different clusters (two buds) is a characteristic scale for the interface.

Metrics independent. It is independent on the explicit distance function adopted. Namely, the actual distance used can be different from Euclidean distance, non euclidean geometries being allowed if they use positive definite metrics (a distance).

http://dx.doi.org/10.5772/58420

2.2. Interaction graphs

The previous subsection demonstrates how the data are analysed. By construction, each set in equation (1) is a collection of edges, the extrema of which are points in some metric space. Mathematically, this corresponds to the notion of graph, namely a set of nodes (here the points in the metric space) and a set of edges joining some of the nodes. Thus, from now on each set representing a level will be naturally identified with a graph. In the present case it is important to stress that the graph has been obtained by evaluating distances, thus each edge is associated to the physical distance of the endpoints. Mathematically speaking, this is a weighted graph.

As the graph nodes were initially atoms of one or the other of two adjacent subunits, the graph is automatically bipartite: each edge has an endpoint on one subunit and the other on the second subunit. Edges between atoms of the same subunit are not considered here, as the analysis focuses on the interface.

By construction, a pair of atoms can be connected by a single edge only, as there is just one distance value between them. Thus parallel edges are absent (edges with the same end nodes) in this description that can be called full atoms.

Since the first paper on the subject [11], the levels S_i have been coarse-grained in order to facilitate the human interpretation of the data and to transfer the information to the residue scale. This is actually the scale used by biological entities to store and transfer information: DNA and RNA code for amino acids, not for atoms, and proteins form from residue chains and not from individual atoms.

In the coarse-grained representation the information appears at the amino acid resolution: all the atoms of a residue are identified and represented with the residue name itself. All edges ending on the atoms of the residue are now referred to the residue itself. In graph theory this procedure is called fusion. If $(a,b),(a',b') \in S_i$ are edges of one of the graphs, let's call F (fusion) the equivalence relation such that

$$(a,b)F(a',b') \iff a \text{ and } a' \text{ belong to the same residue and } b,b' \text{ belong to the same residue} \tag{2}$$

Then, the following quotient set defines equivalent classes

$$S_i^{aa} = \frac{S_i}{F} \tag{3}$$

whole elements are all the edges that start and end on the same residues (aa means amino acid resolution). By consequence, the graphs S_i^{aa} could have parallel edges; this happens when two residues are connected by more than one pair of atoms.

The subsequent analyses will always refer to this coarse-grained graphs, while the initial symmetric minimization has always been done with atoms. The graphs S_i^{aa} can be represented in a very effective way: the first subunit residues are represented as an horizontal line of equispaced dots; similarly, the second subunit residues appear as dots in an horizontal line below the previous one. Clearly, the two horizontal lines represent the chains of residues connected by the peptide bonds. Being the graph bipartite, edges can only join a node from one horizontal line to a node of the other horizontal line. In Figure 1 there is an example.

In fact we found it very convenient to distinguish two sub-graphs. The dotted-dashed lines represent pairs of atoms both from the backbone of the residues (BB graph). The solid lines indicate that at least one atom of the pair is from the side chain (SC graph). Mathematically speaking, this is called an edge-labelled graph, the two possible labels being BB and SC. In principle, there is a third edge label; indeed, the nodes of the same sub-unit form a sequence connected by the peptide bond. Albeit this type of edge is not explicitly shown in the graphs, it is implicitly present and motivates the choice of organizing the residues along straight lines.

Notice that while the full atom graphs are a fortiori different, $S_i \cap S_j = \{\}$, this is no longer true for the fused graphs: it is not a priori granted that they are different.

In summary, the sets in equation (3) characterize the interaction between subunits; each set is a bipartite and edge-labelled graph where parallel edges may occur.

The amino acid resolution graph of the interaction between subunits has shown to be extremely effective in interpreting data and designing statistical analyses. Notice that the choice of marking non interacting residues with a dot and interacting residues with a cross is purely conventional and does not add information.

2.3. Topological analysis of the graphs

The analysis will proceed with the inspection and comparison of the interaction graphs of the level 1 of the interface, with some additional information from the level 2. The focus will be on the topology, namely on the organisation of edges in the graphs, whose importance has already been shown in the previous publications [6, 11]. As already stated, the amino acid resolution will be systematically adopted.

The interface graphs will be analysed using the following motifs.

Zigzag: it is a path that alternates from one to the other of the two sub-units, namely from one to the other of the two nodes subsets of the bipartite graph, like in a zigzag seam. In principle, the smallest recognizable zigzag visits three residues; the experience with data has clearly shown that it is not useful to consider such a path in the zigzag family but it is better to classify it in the vertex family (see later). Thus, the smallest zigzag seam considered is the one with 4 residues visited, zigzag4. It is visible in Figure 5, involving the residues A, I, E, A (BB graph). Zigzag paths with 5 or more residues are comprehensively indicated as zigzag5. See Figure 27, residues T, E, L, V, A.

Vertex: it is a residue connected with two or more different amino acids. In the smallest case, the residue is connected to two other residues and indicated V, as in Figure 4, residues F, L, K. The case of more than two residues connected with the same vertex will be comprehensively indicated with V3. Three exemplars are present in Figure 28. Especially, the residue I on chain B has four connections, three of type BB and one of type SC.

Ladder: This motif occurs when 2 (or more) amino acids at separation δ on the first subunit, namely at positions $M, M+\delta$, are connected with 2 (or more) amino acids $N, N-\delta$ on the second sub-unit, respectively. Notice that the second sub-unit runs oppositely to the first. The two groups of amino acids are the ladder rails and the edges are the rungs, from which the name was adopted. In Figure 3, residues E, Y and E, W form a ladder with 2

http://dx.doi.org/10.5772/58420

rungs and separation $\delta = 2$. In Figure 22 the residues C, S, F (47, 49, 51) of chain A and the residues Y, E, E (110, 108, 106) of chain B form a BB ladder of separation $\delta = 2$. The two most common separations are $\delta = 1$ or 2, ladder1 and ladder2 respectively. Cases with more than 3 rungs are known. In Figure 14, the residues from 98 to 102 (of the chain D) with the residues from 29 to 25 (chain E) form a ladder1 with 5 rungs.

Multiple edges: presence of parallel edges in the graph (edges with the same endpoints). The term multiple edges seems more adequate to the present case, given the use of the word parallel to indicate the orientation of the β-strands in the interface. In Figure 8, the residues E and F are connected by 4 edges, three of type SC and one of type BB. In Figure 22, the residues S and E are connected by three BB edges.

It is important to look for these motifs starting from the largest elements, to avoid useless multiple countings. Especially, it is obvious that a zigzag of 8 nodes contains all the shortest sizes, from 4 to 7. There is no need to record all of them, the largest one being the most informative. Also, a V3 or zigzag motif automatically contains the simplest vertex V, and often more than once. Thus, the vertex V is not counted when it appears inserted in a V3 or a zigzag. In Figure 33, the zigzag5 (residues V, M, I, V, A) contains three times a V. In Figure 30, the vertex V3 (residue H) contains three V motifs.

These topological elements may appear alone or in combination and have been chosen, as in other publications, because they represent the most common motifs in interface graphs. It may happen that a graph contains more than one motif, in which case all of them will be recorded. For example, in a common situation one of the nodes of a zigzag5 is also a vertex V3. In that case, both motifs are registered. The main reason is that a systematic classification of graphs needs to be accurate and free of ambiguous search criteria, thus it is definitely preferable to accept a redundancy in the identification of motifs than introducing untested criteria. For example, often a ladder has separation of 2 in one part of the graph and separation of 1 in another part, with a edge (a rung) in common between the two parts. Thus, it is recorded as part of both a ladder of 2 and ladder of 1. Indeed, so far, no acceptable criterion has been found to discriminate if a edge must be considered part of a ladder with separation of 1 or of 2. In Figure 28, there is a BB ladder formed by M, I, K, I (chain A) and V, L, I, E (chain B). Residues M and I are separated of 2 while residues I and K, K and I are separated of 1.

2.4. The dataset

In [6] a dataset of 39 oligomeric proteins was chosen on the basis of the presence, in the three-dimensional structure, of a well recognizable β interface. Here the same set will be used. It is listed in Table 1 with the indication of the chains and the intervals participating to the β interfaces. All proteins are homomeric of stoichiometry from 3 to 8. This set is characterized by the absence of sequence homology, structural homology or functional homology. Viral and membrane proteins are absent, given their specificities. Thus, the set can be considered representative of the general behaviour of β interfaces, without reference to specific classes of proteins. In the dataset there are anti-parallel, parallel and oblique orientations of the adjacent β strands. Actually, the oblique family will be considered for comparison only and not deeply investigated. In fact its graphs are less structured and it is difficult to find similarities between proteins.

2.5. Residue interaction network generator (RING)

It is a web server [15] with software for rendering a protein structure into a network of interactions. Nodes represent single amino acids and edges represent the non-covalent bonding interactions that exist between them. In particular, this web server has been used to calculate the hydrogen bonds for the proteins of the dataset. The hydrogen bonds have not been supplied to the symmetric minimization method but just used to compare some results.

3. Results

The dataset has been described in Methods and its proteins are listed in Table 1.

The graphs of the level 0 were provided as Supporting Information to the publication [6] and are downloadable in open access, as well as the publication itself.

In this section the description focuses on the level 1 and, to a minor extent, on the level 2. The graphs of the level 1 are provided in Figures 3 to 26, and 27 to 34. Given their minor role, the graphs of the level 2 will be provided on demand.

In the following description, the results are grouped and numbered from Result 1 (R1) to Result 5 (R5).

R1. BB graph at level 1. The level 1 graphs are not identical to the level 0 ones.

In all the anti-parallel cases (Figures 3 to 26), the presence of a level 1 BB graph is observed, of size comparable with the one observed at level 0, namely with a similar number of edges. It is very small in just one case (2ojw level 1), where a single edge is present.

Similarly, in all the parallel cases (Figures 27 to 34), there is a level 1 graph, with a number of edges similar to the one found at level 0.

R2. BB graph structure at level 1. In the anti-parallel orientation, the graphs present a ladder structure in 21 out of 24 cases. Its separation is of 2 amino acids in 20 cases out of 24 and of 1 amino acid in 15 out of 24 cases[1]. The zigzag connection is absent. The vertex V is present in 13 graphs on 24. The V3 motif is present in 9 out of 24 cases. The zigzag4 is counted 12 out of 24 times. Just 4 multiple edges are observed. In all cases at least one ladder motif shows up at level 0 or at level 1.

In the parallel orientation, the graphs shows a common presence of the zigzag5 topological element, in 4 out of 8 graphs, often accompanied by one or more V3, in 5 out of 8 graphs. The zigzag5 motifs are probably not completely independent from those that appear at level 0, as they always show up together. This aspect needs a larger statistics to be confirmed or disproved. The zigzag opening is of 1 or 2 amino acids. There are also 4 ladders of separation 1. Multiple bonds appear once out of 8 cases[2].

R3. BB graph structure at level 2. A level 2 BB graph is observed in all cases, often less populated than the lower levels graphs.

In the anti-parallel orientation, there are 13 ladders out of 24. The other cases show a V or zigzag4 motif, and in one case a zigzag5 is found.

[1] At level 0 the ladder separation is of 2 residues (23 cases out of 24). There are multiple edges in 21 cases out of 24.
[2] At level 0 there are 5 zigzag5 out of 8. After, one finds 2 V and 1 zigzag4. Multiple bonds are absent.

http://dx.doi.org/10.5772/58420

In the parallel orientation, the zigzag topology and the vertices V3 are absent at this level. On the contrary, these are the most common topological elements at the lower levels. At this level we just find the ladder motif with separation of 1 or 2 amino acids. In some cases the BB graph reduces to one edge.

R4. Other orientation. In interfaces other that the anti-parallel or parallel β, it was systematically observed in previous publications and is confirmed here that the BB graphs are rare and poorly structured, as composed of one or two edges. This obviously holds true also at levels 1 and 2.

R5. SC graph. These graphs are more elaborated and very rich of a variety of elements. One may recognize that the motif of the ladder is present in nearly half the graphs, at all the various levels examined (0, 1, 2). The V, V3 and zigzag elements are very common in all the three orientations. A more complete analysis of the SC graphs will be realized in future publications.

4. Discussion

In Methods, four definitions of interface have been introduced.

The participating amino acids are actually very similar: in [11], paper fully dedicated to compare the level 0, S_0^{aa}, with published interface data more than 85% of similarity has been detected. At the level 0, the difference is more in terms of the description that emerges. In the buried surface or Voronoi cells based interface approach, is very natural to distinguish between rim and core: a solvent molecule that tries to penetrate the interface first has to visit the rim, and after may force into the core. The symmetrized levels instead present the growth of the interface, as in a budding process, from the set S_0. The hierarchy in (1) indicates the dynamical sequence of events that may construct the interface with minimal steric effects. Somehow, this introduces the notion of time in an otherwise static view. The distinction between rim and core is possibly given by the number of the connections of a residue: a highly connected residue must be in the core and cannot be in the rim. Vice versa, a poorly connected residue is in the rim. Indeed, the part of a residue surface that is exposed to the solvent will not be connected to other atoms while an atom that is completely buried will have connections with all its neighbours. This indicates that a minimal distinction between rim and core may appear if one compares few close levels, like S_0, S_1 and S_2. The point is that the more an atom is present at different levels, the more is connected thus the rim must correspond to those atoms that appear few times across the levels, the others being in the core.

The importance of how the information circulates in an interface has been first shown in [8]. The graphs S_i^{aa} show the presence of long range correlations that are not easily detected with approaches based on the contact surface (buried surface or Voronoi cells based interface). The motifs that allow this transfer are the zigzags, especially the long ones, and the vertices V, V3. On the contrary, a pure ladder topology, in BB or SC or both, does not have ways to correlate far apart atoms. The frequent presence of zigzag, V and V3 motifs implies the existence of constraints on the positions or the physico-chemical properties of non neighbouring atoms, and sometimes of very far apart atoms. Indeed, in Figure 27, the mere establishment of the path formed by the residues S2, E123, L5, L124, I7, V127 (where BB and SC are both present)

requires to satisfy physical conditions due to the volume of the atoms and the distribution of electric charge. Each residue has its internal constraints, among which the fixed length of covalent bonds and the planarity of the surface $C_\alpha, C, 0, N$ that contains the peptide bond. The sequence of edges transfers constraints and establishes a correlation between the outer amino acids S2 and V127 even if they are not in physical contact. The exact microscopic description of the constrains is not easy to find, although. Notice that this example of information transfer along the interface has been detected thanks to a description guided by graph theory and based on edges.

In summary, the main difference that appears comparing buried surface and Voronoi cells based interface with SI is that the first two definitions focus on the spatial organization of the interface while the latter may suggest a temporal organization and allows to evaluate how the information circulate in the interface.

The result R1 clearly states that the level 1 graphs can have a BB graph of size comparable to that of level 0, thus still informative. At level 2 a smaller BB graph is observed. Preliminary results on levels higher than 2 indicate that the BB graphs are also present.

The result R2 indicates that the level 1 graphs have a structure similar to the one found at level 0, in other words that these two levels present several common elements. Instead, from R3 the level 2 graphs seem organized in a different way. The main structure of the level 2 graphs is the ladder one, in both the anti-parallel and parallel orientations, that indicates that these orientations are not distinguishable at this level, and possibly above. Preliminary results on levels higher than 2 indicate that the main distinctive motif of the parallel orientation, namely the zigzag5, is quite rare. Thus, it seems that level 0 and 1 are those that contain the most useful geometrical and topological information.

In [8] we have already used the properties in R2, R3, by implementing algorithms that, from the PDB structure, are able to characterize an interface and tell if it has a β structure, and which is its orientation. These algorithms are based on the level 0 only[3]. The analysis of level 1 graphs confirms and expands this possibility, because the information from both the levels can now be combined for a more accurate recognition.

The BB graphs at level 0 have been previously associated to structural hydrogen bonds that are present in the anti-parallel and in the parallel orientations of β strands.

It is possible that level 1 (for both the BB and SC graphs) doesn't describe proper chemical bonds but weaker dipole-dipole or Van der Waals interactions . This comparison for the level 1 has not yet been explored.

4.1. Counting the degrees of freedhom

The question that we address now is to evaluate if the description provided by the graphs is enough to reconstruct the shape of the interface or not.

To reconstruct a shape in three dimensions one needs 3 coordinates for each point: $3N$, where N is the number of points. Actually, the absolute position of the centre of mass and the spatial orientation of the object in the space are totally irrelevant thus three overall translations and

[3] Other interface arrangements are not yet recognized.

http://dx.doi.org/10.5772/58420

three rotations can be removed from the counting, that reduces the number of needed relative positions or distances to

$$\text{degrees of freedhom} = 3N - 6 \tag{4}$$

As an example, consider the graph of 2ojw in Figure 19; there are $N = 7$ amino acids (even the amino acids that do not participate to the graph but are included within the considered regions must be counted) thus one needs $3N - 6 = 15$ relative positions. The distances between two consecutive C_α is fixed in all polypeptides (as the distance TG in the graph). In the present case there are 5 of them. The distance between the first and the third of three consecutive C_α (distances TT and GI in the graph) is fixed by the general properties of polypeptides, that makes 3 distances. The graph has 3 edges, namely 3 other distances, thus one remains with $15 - 5 - 3 - 3 = 4$ more distances to be fixed. This indicates that this graph is insufficient to reconstruct the shape. A more general counting is possible. In [8] the average number of amino acids (18) and of edges (12) in a β interface have been evaluated. Their ratio is very close to $3/2 = 1.5$, thus it is reasonable to assume that if there are N residues in the interface, there will be nearly $2N/3$ edges (actually, multiple edges should not be counted, here; this may further reduce the number of known edges). Also, we expects $N - 2$ consecutive C_α and $N - 4$ groups of three consecutive C_α namely we have

$$\text{known distances} = (N - 2) + (N - 4) + \frac{2}{3}N = \frac{8}{3}N - 6 \tag{5}$$

We subtract the number of known distances to the number of degrees of freedhom and we are left with the number of distances that are needed to fix the shape

$$\text{unknown distances} = (3N - 6) - \left(\frac{8}{3}N - 6\right) = \frac{1}{3}N \tag{6}$$

Thus a single level does not provide enough data but two levels provide sufficient information to fix the shape of the interface.

Of course, a full evaluation of the interface degrees of freedom needs a much more complex calculation with atoms but the present counting suggests that few lowest levels should be enough to provide an accurate description of the interface shape and the position of most of the atoms.

A more complete account of the problem of reconstructing the shape of a set of points given an incomplete set of distances is treated in [16].

4.2. Perspectives

The result R5 is clearly indicative of the major complexity of the side chain by respect to the backbone. In the paper [8] the role of the residues with multiple interactions, namely V and V3, has been studied in detail in S_0^{aa} and has been correlated to the length of the side chains. Following this observation, one could introduce a parametrization based on the length of the side chain. The discussion on the information flow in the interface points in the same direction, of dedicating a new publication to the study of the SC graphs.

stoichiometry	PDB name	chains	range on the chains	orientation of the β interface
3	1JN1	AB	120-140,1-20	oblique
3	1PM4	AB	82-95,64-80	anti-parallel
3	1SJN	AB	1-12,118-130	parallel
3	1SNR	AB	109-129,326-342	anti-parallel
3	1T0A	AB	1-15,125-141	oblique
3	1Y13	AB	9-21,161-174	anti-parallel
3	2BAZ	AB	1-8,116-130	parallel
3	2BCM	BA	43-53,17-26	anti-parallel
3	2BT9	AB	44-57,1-16	oblique
3	2GVH	AB	189-202,59-73	anti-parallel
3	2I9D	AB	148-166,17-33	anti-parallel
3	2JCA	AB	1-17,103-124	oblique
3	2P90	AB	71-88,168-180	anti-parallel
4	1J8D	AB	19-29,30-40	anti-parallel
4	1L3A	AD	118-129,88-98	parallel
4	1PVN	AD	489-496,432-438	anti-parallel
4	2A7R	AD	1-16,327-339	parallel
4	2H5X	AD	1-8,18-29	anti-parallel
4	3BF0	AB	445-468,178-197	anti-parallel
5	1B09	AB	197-206,99-112	oblique
5	2XSC	AB	62-69,8-16	oblique
5	1EEI	DE	94-103,21-33	anti-parallel
5	1EFI	DH	23-33,94-103	anti-parallel
5	1FB1	AE	125-138,218-237	anti-parallel
5	1HI9	AB	66-84,175-191	anti-parallel
5	1NQU	AE	1-6,43-54	parallel
5	1WUR	AB	186-197,93-105	anti-parallel
5	2OJW	AB	42-48,188-195	anti-parallel
5	2RCF	AB	72-83,8-21	anti-parallel
6	1U1S	AB	48-60,54-69	anti-parallel
6	2BVC	AF	211-219,33-40	oblique
6	2GJV	AB	43-56,102-112	anti-parallel
6	2Z9H	AB	5-18,77-89	anti-parallel
7	1HX5	AG	3-13,92-99	anti-parallel
7	1OEL	AG	34-43,511-524	parallel
7	1WNR	AG	1-10,87-96	anti-parallel
7	2RAQ	AB	33-46,76-91	anti-parallel
8	1Q3S	AB	46-57,515-527	parallel
8	2V9U	AB	140-148,170-177	parallel

Table 1. Table of the proteins considered in this paper, from [6]. In summary, we have 24 antiparallel, 8 parallel and 7 oblique orientations.

http://dx.doi.org/10.5772/58420

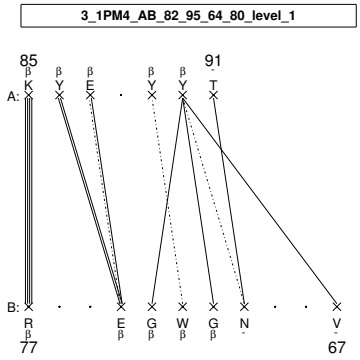

Figure 4. Anti-parallel orientation of the β-strands.

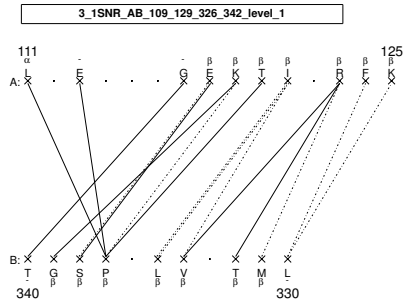

Figure 5. Anti-parallel orientation of the β-strands.

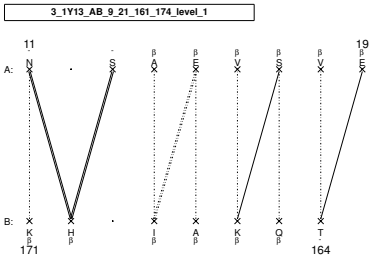

Figure 6. Anti-parallel orientation of the β-strands

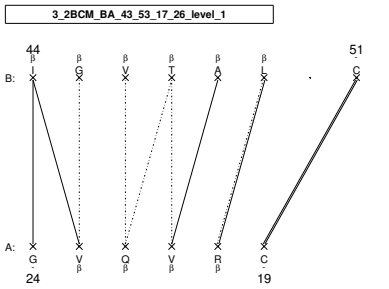

Figure 7. Anti-parallel orientation of the β-strands.

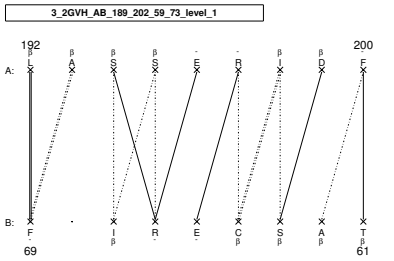

Figure 8. Anti-parallel orientation of the β-strands.

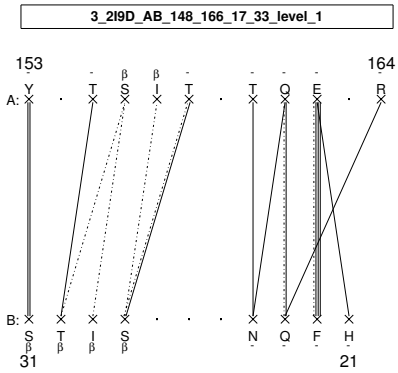

Figure 9. Anti-parallel orientation of the β-strands.

http://dx.doi.org/10.5772/58420

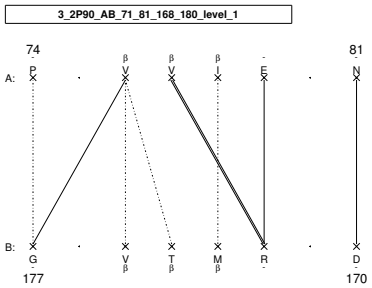

Figure 10. Anti-parallel orientation of the β-strands.

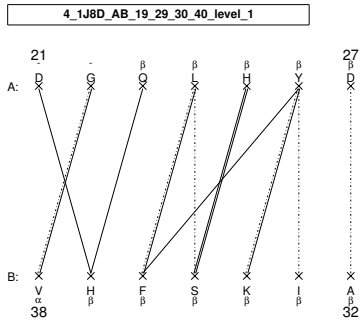

Figure 11. Anti-parallel orientation of the β-strands.

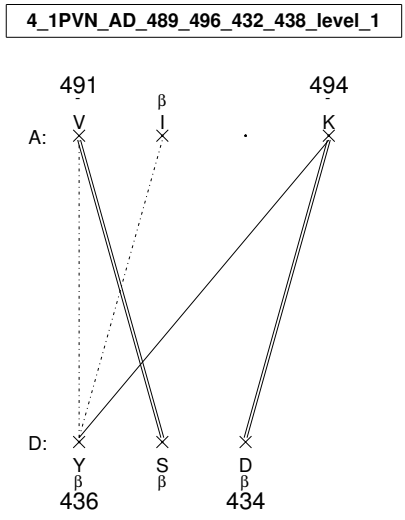

Figure 12. Anti-parallel orientation of the β-strands.

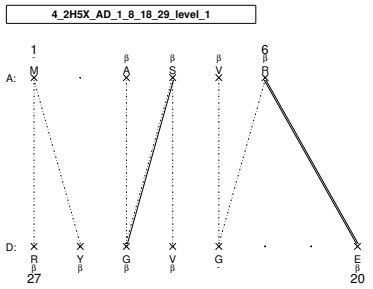

Figure 13. Anti-parallel orientation of the β-strands.

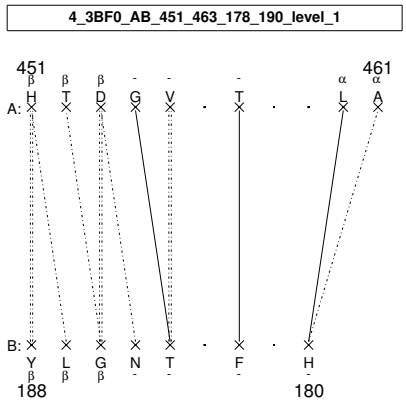

Figure 14. Anti-parallel orientation of the β-strands.

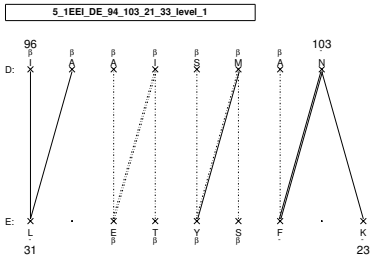

Figure 15. Anti-parallel orientation of the β-strands.

http://dx.doi.org/10.5772/58420

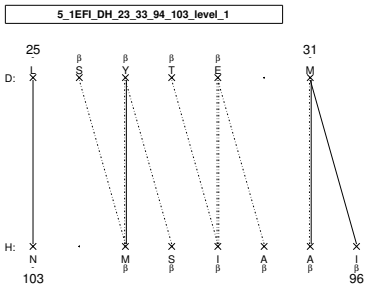

Figure 16. Anti-parallel orientation of the β-strands.

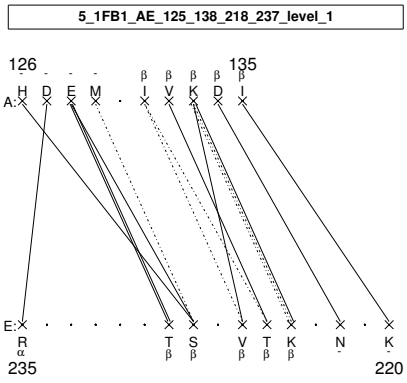

Figure 17. Anti-parallel orientation of the β-strands.

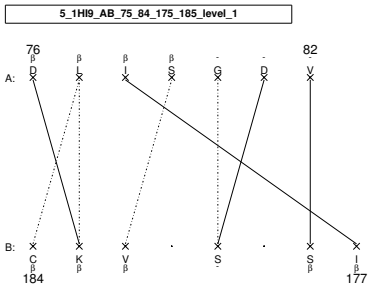

Figure 18. Anti-parallel orientation of the β-strands.

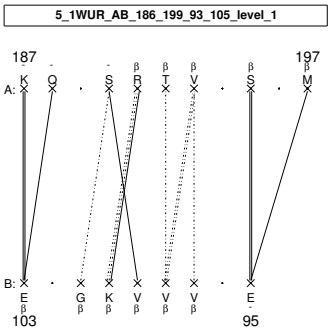

Figure 19. Anti-parallel orientation of the β-strands.

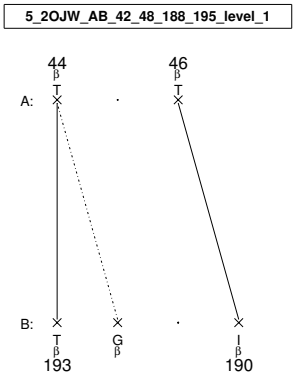

Figure 20. Anti-parallel orientation of the β-strands.

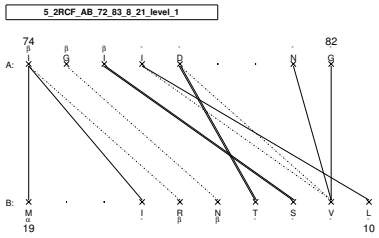

Figure 21. Anti-parallel orientation of the β-strands.

http://dx.doi.org/10.5772/58420

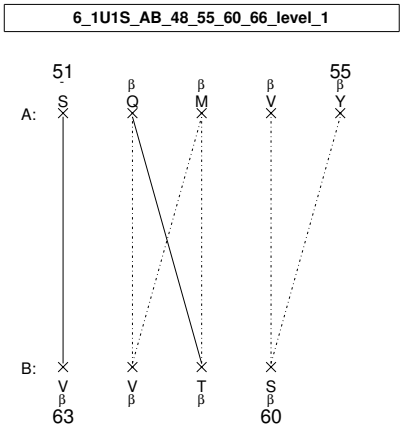

Figure 22. Anti-parallel orientation of the β-strands.

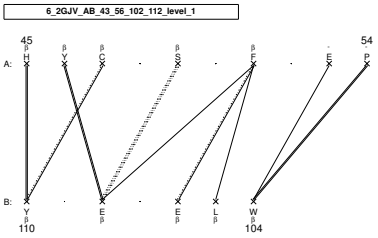

Figure 23. Anti-parallel orientation of the β-strands.

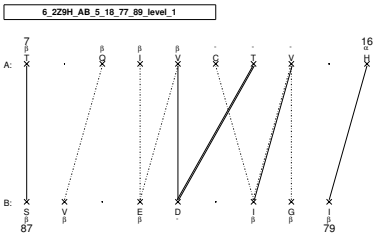

Figure 24. Anti-parallel orientation of the β-strands.

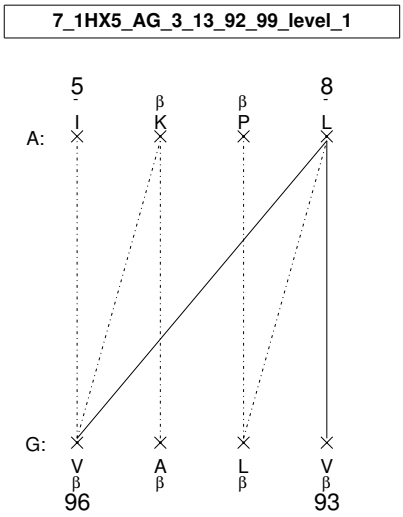

Figure 25. Anti-parallel orientation of the β-strands.

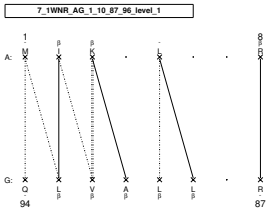

Figure 26. Anti-parallel orientation of the β-strands.

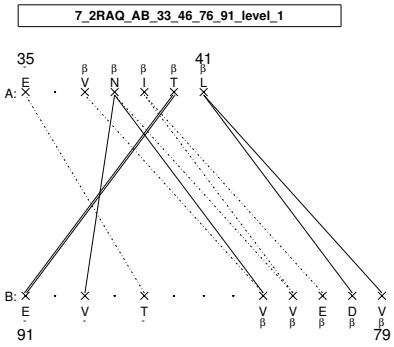

Figure 27. Anti-parallel orientation of the β-strands.

http://dx.doi.org/10.5772/58420

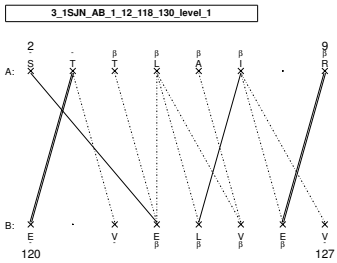

Figure 28. Parallel orientation of the β-strands.

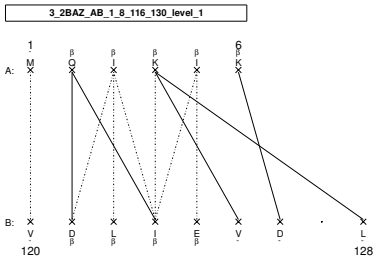

Figure 29. Parallel orientation of the β-strands.

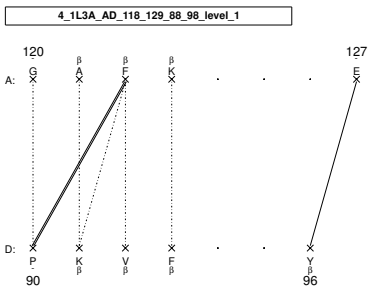

Figure 30. Parallel orientation of the β-strands.

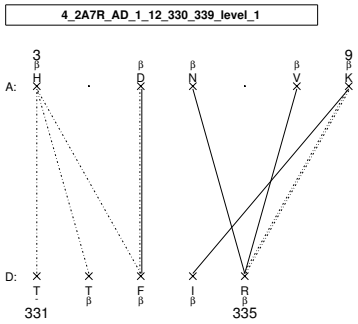

Figure 31. Parallel orientation of the β-strands.

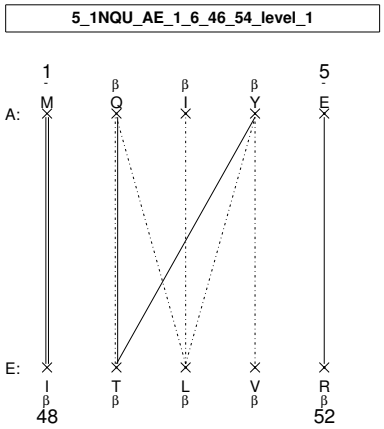

Figure 32. Parallel orientation of the β-strands.

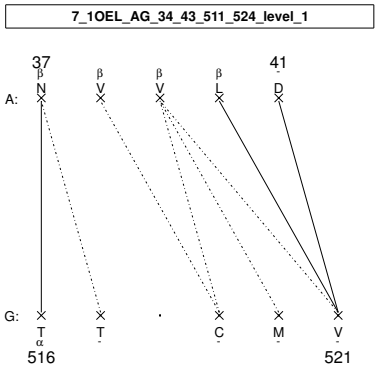

Figure 33. Parallel orientation of the β-strands.

http://dx.doi.org/10.5772/58420

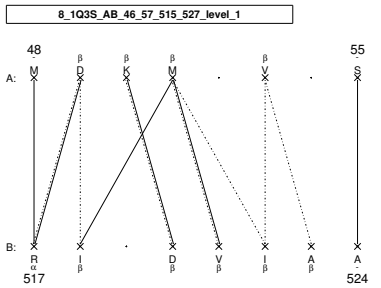

Figure 34. Parallel orientation of the β-strands.

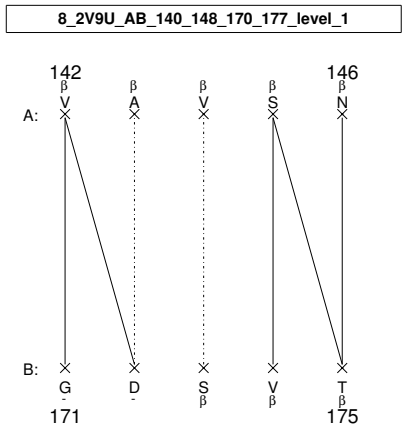

Figure 35. Parallel orientation of the β-strands.

5. Conclusion

The whole analysis presented so far has been triggered by the problem of investigating biological interfaces, namely interfaces that form during the biochemical activity in a cell, between or inside proteins, protein-DNA or protein-RNA complexes and so on.

In the introduction, motivations are given for the possibility of using geometry to understand the interactions. In the Results and Discussion sections, the topological, but intrinsically geometrical, properties of the interaction graphs have been presented with the first goal of learning how to distinguish the two main orientations and the second goal of estimating aspects that have been systematically negletted in all the previous publications.

Both these aspects have been clearly addressed in this paper. The tools to distinguish the two orientations are now more precise.

The results are the consistency of the information from level 1 and level 0 and the limited amount of information from level 2. Both the results suggest to enrich previous statistics and models for prediction with some input from level 1.

These results are confirmed by the naive counting of the degrees of freedhom: at least in a coarse-grained view, the two levels S_0^{aa} and S_1^{aa} provide enough information to reconstruct the shape of the interface (completeness).

In Methods, the notion of S_0^{aa} as the minimal description of the interface, or framework, already used in all our previous publications, has been presented here with solid mathematical arguments: as all buds are present in S_0^{aa} and no bud can appear later, it is legitimate to call S_0^{aa} a framework because a smallest set would miss some buds, namely some parts of the interface. We find that this and the completeness of the interface add important values to the validity of the methods.

Also, it is important to stress the ability of the symmetric minimization to detect the BB hydrogen bonds from the knowledge of the positions of non-hydrogen atoms only: in this case, geometry intrinsically reveals the chemical interactions, without making use of a cut off or other external scales.

Acknowledgements

This work is founded by the region Rhone-Alpes.

It's a pleasure to thanks Claire Lesieur for most valuable comments and suggestions, and Laurent Vuillon, for his help.

With sadness, the author remembers the friend and colleague Laurent Fournier, recently deceased, for his useful suggestions to improve the C++ codes used in the analyses.

Author details

Giovanni Feverati

Fédération de recherche MSIF, University of Savoie and CNRS, Annecy-le-Vieux, France

References

[1] Iacovache I, van der Goot G F, Pernot L. Pore formation: An ancient yet complex form of attack. Biochim Biophys Acta, vol. 1778, num. 7-8, p. 1611-23 (2008).

[2] Lesieur C, Vecsey-Semjen B, Abrami L, Fivaz M, van der Goot G F. Membrane insertion: The strategies of toxins (review). Mol Membr Biol 14: 45-64 (1997).

[3] Kirkitadze M D, Bitan G, Teplow D B. Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies. J Neurosci Res 69: 567-577 (2002).

[4] Harrison R S, Sharpe P C, Singh Y, Fairlie D P. Amyloid peptides and proteins in review. Rev Physiol Biochem Pharmacol 159: 1-77 (2007).

[5] Crick F H C. The packing of alpha-helices: simple coiled-coils. Acta Crystallogr 6: 689-697 (1953).

[6] Feverati G, Achoch M, Zrimi J, Vuillon L, Lesieur C. β-strand interfaces of non-dimeric protein oligomers are characterized by scattered charge residues pattern. PLoS ONE 7(4): e32558 (2012). The article and the Supporting Information are freely downloadable from Plos One: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0032558#s5.

[7] Zrimi J, Ng Ling A, Giri-Rachman Arifint E, Feverati G, Lesieur C. Cholera toxin B subunits assemble into pentamers: proposition of a fly-casting mechanism. Plos One 5(12) e15347 (2010).

[8] Feverati G, Achoch M, Vuillon L, Lesieur C. Residue interaction networks of β-strand interfaces from healthy protein oligomers have few HUBs (multiple contact residues) and low interconnectedness: a potential protection mechanism against network rewiring and chain dissociation. Under revision by Plos One.

[9] ExPASy: Bioinformatics resource portal, http://www.expasy.org/proteomics/protein_structure

[10] Janin J, Bahadur R P, Chakrabarti P. Protein-protein interaction and quaternary structure. Q Rev Biophys 41: 133-180 (2008).

[11] Feverati G, Lesieur C. Oligomeric interfaces under the lens: Gemini. PLoS ONE 5(3): e9897 (2010).

[12] Feverati G, Lesieur C, Vuillon L. Symmetrization: ranking and clustering in protein interfaces. Referred proceeding of the conference "Mathematics of distances and applications", MDA 2012, Varna. Publisher: ITHEA, Sofia. Editors: M. Deza, M. Petitjean, K. Markov.

[13] Lance G N, Williams W T. A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems The Computer Journal (1967) 9 (4): 373-380.

[14] Murtagh F and Contreras P. Methods of Hierarchical Clustering. Data Mining and Knowledge Discovery. Wiley-Interscience, Vol. 2, No. 1, pp. 86âĂŞ97 (2012).

[15] Martin A J M, Vidotto M, Boscariol F, Di Domenico T, Walsh I, Tosatto S C E. Residue Interaction Network Generator (RING). http://protein.bio.unipd.it/ring/

[16] Liberti L, Lavor C, Maculan N, Mucherino A. Euclidean Distance Geometry and Applications SIAM Review 56[1]:3-69. doi 10.1137/120875909