SYMMETRIZATION: RANKING AND CLUSTERING IN PROTEIN INTERFACES

Giovanni Feverati, Claire Lesieur, Laurent Vuillon

Abstract: *Purely geometric arguments are used to extract information from three-dimensional structures of oligomeric proteins, that are very common biological entities stably made of several polypeptidic chains. They are characterized by the presence of an interface between adjacent amino acid chains and can be investigated with the approach proposed here. We introduce a method, called symmetrization, that allows one to rank interface interactions on the basis of inter-atomic distances and of the local geometry. The lowest level of the ranking has been used previously with interesting results. Now, we need to complete this picture with a careful analysis of the higher ranks, that are for the first time introduced here, in a proper mathematical set up. The interface finds a very nice mathematical abstraction by the notion of weighted bipartite graph, where the inter-atomic distance provides the weight. Thus, our approach is partially inspired to graph theory decomposition methods but with an emphasis to "locality", namely the idea that structures constructed by the symmetrization adapt to the local scales of the problem. This is an important issue as the known interfaces may present major differences in relation to their size, their composition and the local geometry. Thus, we looked for a local method, that can autonomously detect the local structure. The physical neighborhood is introduced by the concept of cluster of interactions. We discuss the biological applications of this ranking and our previous fruitful experience with the lowest symmetrized level. An example is given, using the prototypical cholera toxin.*

Keywords: *symmetrization, protein interfaces, oligomeric proteins, graphs, bonds ranking, interaction clusters.*

ACM Classification Keywords*:*
J.2 Physical sciences - Mathematics and statistics
J.3 Life and medical sciences - Biology and genetics

Conference topic*: MDA 2012 - "Mathematics of Distances and Applications" : Distances in Natural Sciences*

PACS*: 87.15.bk Structure of aggregates, 87.15.km Protein-protein interactions, 87.15.hg Dynamics of intermolecular interactions*

MSC*: 52-XX Convex and discrete geometry, 52C99 None of the above, but in this section, 92-XX Biology and other natural sciences, 92B99 None of the above, but in this section*

Introduction

The present work is motivated by the biological problem of understanding and possibly predicting the assembly of biological molecules. This is one of the most common processes in living cells: proteins and, more generally, biological molecules, interact with each other by temporary or permanently associating into a unit that realises some biological function. Actually, the majority of proteins are permanently formed of several subunits organized into small polymers known as oligomeric proteins [Goodsell, 2000]. Understanding the mechanisms of their formation is particularly important due to their implication in several pathologies, from bacterial infections to protein misfolding diseases (Alzheimer, Parkinson, ...) [Iacovache, 2008; Lesieur, 1997; Kirkitadze, 2002; Harrison, 2007]. The 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.

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 perform interface prediction as well as interface design. The shape

and the function of proteins are encoded within their sequence, i.e. in their amino acid composition. But it is not yet possible, by simply reading the primary sequence of a protein, to know its three-dimensional structure or the quaternary organization, in the case of a protein oligomer. One of the difficulties is the non linear encoding of the information in the sequence, that requires three-dimensional analysis. 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.

One possible strategy to decipher the sequence-structure relationship of protein interfaces is through the analysis of the interface geometry. This is motivated by the observation, made in the early 50's by F. Crick [Crick, 1953], that the formation of the coiled-coil 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 geometry combined to some contact points which together provide it the capacity to associate to one lock. Moreover, he also observed that protein interfaces of similar geometry have similar chemistry, namely similar contact points.

Here we follow the same line of thinking, moving from the geometry toward the chemistry. We develop the "symmetrization", a method to decipher interface properties through the analysis of its geometry and interactions. Their strength is measured by some "distance". Indeed, we rely on the notion that physical interactions decrease with increasing distance. With this, the symmetrization produces a hierarchical ranking of the interactions. The notion of physical neighborhood appears by a concept of "clusters of interactions" (see later).

In [Feverati, 2010] we succeeded in showing that it is possible to encode some information of the interface in a graph (interaction network). Comparative statistics on many protein graphs has been a fruitful method to extract useful features for the intermolecular beta sheet interface geometry [Feverati, 2012]. These results were based on the lowest level of ranking, called S_0. All the higher levels were neglected: possibly a 90% of the actual interactions was ignored, focusing on the strongest only. It is important now to explore the structure of the neglected interactions and estimate the information they may contain. To assess this, we need a careful mathematical understanding of the symmetrization, that will be revisited from a more mathematical point of view, in order to develop its full potential.

Methods and results

The symmetrization, first introduced in [Feverati, 2010] and fully developed here for the first time, extract information on the interfaces from the three-dimensional PDB protein structures. The term information is understood in a very wide sense: which atoms form intermolecular bonds, which is the role of the different amino acids in the interface, how the charge or the hydrophobic character of some atomic groups in the amino acids matters in forming or stabilizing the interface, and so on.

In an interface, the typical distances between atoms of two adjacent chains are of the order of 2-5 Angström. Thus it is important to work at the atomic scale, even if the ultimate information must be expressed at the amino acid scale because the formation of proteins always passes through one or few amino acid sequences.

Let A, B the set of atoms of the first, second subunit, respectively. We indicate with $d(a,b)$ a distance between an atom $a \in A$ and an atom $b \in B$. Notice that the notation itself indicates the set to which the atoms belong, namely the first parameter of $d(\ ,\ )$ must be in A, the second in B.

The mathematical development that will be presented in this section is actually independent of the explicit distance function adopted and of the space dimension. Thus, the presentation holds in a generic metric space (or even less than it because we don't actually need the triangular inequality, for the derivation. We need it for the interpretation, in the Discussion section). In the Discussion section, we will present an example based on the Euclidean distance in $\mathbb{R}^3$.

We call raw interface the following set of ordered pairs

$$R_0 = \{(a,b) \in A \times B, \text{ such that } d(a,b) \leq d_0\}, \quad d_0 \geq 0 \quad \text{some cut-off} \tag{1}$$

namely pairs of atoms at distance lower than the cut-off are in the interface. In the present Methods section, the actual cut-off value is immaterial, namely the construction holds whatever its value is. For the purposes of discussing intermolecular interactions, it is often fixed to

$$d_0 = 5 \text{ Angström} \tag{2}$$

This definition of the interface does not provide any measure to distinguish pairs whose atoms are at different distances or in different locations. On the contrary, 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 atomic environments.

This means that besides the distances, we need to rank the interactions within an interface, with a ranking sensitive to the local conditions.

The lowest level

We will make large use of the minimum function

$$\min \mathcal{S}\,, \quad \text{where} \quad \mathcal{S} = \text{ some set of reals} \tag{3}$$

to extract the minimum distance in the set. As we have in mind applications, we will use finite sets. Thus, the minimum will always be realized. Let define a first subset of the raw interface by

$$L_A = \Big\{(a,b) \in R_0 \ : \ d(a,b) = \min\big\{d(a,c) : c \in B \text{ and } (a,c) \in R_0\big\}\Big\} \tag{4}$$

This set associates to every atom $a \in A$ its closest neighbour on B (or neighbours, if equidistant), namely it minimizes the distances by respect to the atoms of the first subunit. Similarly, we consider the set of pairs that minimises distances by respect to the atoms of the second subunit

$$L_B = \Big\{(a,b) \in R_0 \ : \ d(a,b) = \min\big\{d(c,b) : c \in A \text{ and } (c,b) \in R_0\big\}\Big\} \tag{5}$$

The symmetrized interface is the subset of the raw interface defined as the intersection of these two sets

$$S_0 = L_A \cap L_B\,, \quad S_0 \subseteq R_0 \tag{6}$$

A restatement of the definition is that a necessary and sufficient condition for a pair $(a,b) \in R_0$ to be in S_0 is to satisfy the system of equations

$$\begin{cases} d(a,b) = \min\Big\{d(a,c) \ : \ c \in B \text{ and } (a,c) \in R_0\Big\} \\ d(a,b) = \min\Big\{d(c,b) \ : \ c \in A \text{ and } (c,b) \in R_0\Big\} \end{cases} \tag{7}$$

Notice that the unknowns are not numbers but pairs of R_0. In practice the logic is the following. Let fix an arbitrary pair (a,b). Is that pair the shortest (or one of the shortest if several have equal length) of all the bonds coming out of a (like rays of a star)? If not, then (a,b) cannot solve the system. If yes, then is (a,b) the shortest (or one of the shortest if several have equal length) of all the bonds coming out of b? If yes, then (a,b) solves the system and belongs to S_0. In other words, when both equations are satisfied, the bond under examination minimizes two distinct sets of bonds, one at each of its extremes: it is a local minimum. This is our notion of symmetrization of a set of bonds.

Theorem 1. *If $R_0 \neq \{\}$, then S_0 cannot be empty.*

Proof. Let consider the pair $(\bar{a}, \bar{b})$ that minimizes the whole set of distances of R_0 (there must be at least one such pair)

$$d(\bar{a}, \bar{b}) = \min\{d(c_1, c_2) \ : \ c_1 \in A \, , \ c_2 \in B \, , \ (c_1, c_2) \in R_0\} \tag{8}$$

It is a global minimum. This implies that the set appearing on the right hand side of the following equation is also minimized

$$d(\bar{a}, \bar{b}) = \min\{d(c_1, \bar{b}) \ : \ c_1 \in A \, , \ (c_1, b) \in R_0\} \tag{9}$$

Similarly, (8) implies also that the following expression is verified

$$d(\bar{a}, \bar{b}) = \min\{d(\bar{a}, c_2) \ : \ c_2 \in B \, , \ (a, c_2) \in R_0\} \tag{10}$$

These last two equations actually constitute the system in (7), thus the pair $(\bar{a}, \bar{b})$ belongs to S_0 which, in that way, cannot be empty. From now on, we will always assume that R_0 is not empty. ■

Corollary 1. *If an atom a appears more than once in S_0, then it appears with equidistant pairs*

$$d(a, b_1) = d(a, b_2) \quad \textit{where} \quad (a, b_1), (a, b_2) \in S_0 \tag{11}$$

Proof. Indeed, multiple minima are allowed. ■

The higher levels

In this subsection we will prove that the procedure of symmetrization can be iterated. We start by defining a new set of bonds

$$R_i = R_{i-1} - S_{i-1} \, , \qquad i = 1, 2, \ldots \tag{12}$$

As in (7), we look for pairs $(a, b) \in R_i$ that solve the following system

$$\begin{cases} d(a, b) = \min\left\{d(a, c) \ : \ c \in B \text{ and } (a, c) \in R_i\right\} \\ d(a, b) = \min\left\{d(c, b) \ : \ c \in A \text{ and } (c, b) \in R_i\right\} \end{cases} \tag{13}$$

The subset that solves this system of equations defines the symmetrized set at level i

$$S_i = \{(a, b) \in R_i \ : \ \text{it solves (13)}\} \, , \qquad S_i \subseteq R_i \tag{14}$$

The proof of theorem 1 can now be repeated here, after replacing R_0 with R_i. This shows that, as long as $R_i \neq \{\}$, its subset S_i cannot be empty as it must contain at least the global minimum, namely a pair $(\bar{a}, \bar{b})$ that minimizes all distances of R_i

$$d(\bar{a}, \bar{b}) = \min\{d(c_1, c_2) \ : \ (c_1, c_2) \in R_i\} \tag{15}$$

All what is stated from (12) to here can be repeated at all orders. Considering that we have a finite number of atoms and that the symmetrized sets are not empty, at some point the recursion (12) will stop with an empty set

$$R_M = \{\} \quad \text{with} \quad R_{M-1} \neq \{\} \, , \quad M = \text{number of levels} \tag{16}$$

By the definition of the recursion (12), it is obvious that a level cannot intersect the previous ones thus we state the following theorem.

Theorem 2. *Different levels are separated:* $S_i \cap S_{i-1} = \{\}$

The initial set of bonds R_0 is now the union of disjoint sets

$$R_0 = \bigcup_{0 \le i < M} S_i \tag{17}$$

Thus the initial set of bonds has been sliced in levels, with each bond uniquely classified into a level. An important theorem that we will use later is now stated.

Theorem 3. *At least one atom in a pair must appear in one pair of the previous symmetrized set, namely*

$$(a_1, b_1) \in S_i \Longrightarrow (a_1, c) \in S_{i-1} \text{ or } (c', b_1) \in S_{i-1} \text{ for some } c, c' \tag{18}$$

(both are possible). Equivalently, one can say that each bond of S_i contains at least one atom of S_{i-1}.

Proof. To show this, let suppose by absurd that they both do not appear. In that case, for the pair (a_1, b_1) it would be equivalent to replace R_i by R_{i-1} in (13), because no record of this pair is present in S_{i-1}. But then, the system of equations that defines the level i (13) would look as the definition of the previous level $i-1$, indicating that the pair belongs to S_{i-1}, that is absurd because of (12). The second formulation re-writes the first one. The following corollary is obvious. ■

Corollary 2. *With respect to S_i, a new pair (a, b), namely one with both new atoms, cannot be found before the level $i+2$, namely in S_{i+2}.*

The flow chart of the whole symmetrization procedure is represented in Figure 1.

Towers and clusters

The pairs at level zero of the symmetrization can be used as the starting point of towers of bonds. The lowest parts of the towers define non intersecting clusters of bonds. Each pair at level zero defines the level zero (ground) of each tower by

$$T_0(a, b) = \{(a, b)\} \quad \forall \quad (a, b) \in S_0 \tag{19}$$

Each tower level higher that the ground is recursively defined by choosing all bonds of S_i that have one atom in common with the lower tower level

$$T_i(a, b) = \{(a', b') \in S_i \ : \ \text{either } (a', c) \in T_{i-1}(a, b) \text{ or } (c, b') \in T_{i-1}(a, b) \text{ for some } c\} \tag{20}$$

The full tower will be the union of all levels

$$T(a, b) = \bigcup_{i \ge 0} T_i(a, b) \tag{21}$$

Notice that the tower and its levels are always labelled by the initial ground pair, namely $(a, b) \in S_0$.

Theorem 4. *At each level, the union of all towers exhausts the corresponding symmetrized set:*

$$\bigcup_{(a,b) \in S_0} T_i(a, b) = S_i \tag{22}$$

Proof. First, this property holds at level 0 (19). For higher i, we prove it by induction, supposing that it holds at $i-1$ and showing that this implies its validity at i. So, at this level we have

$$T = \bigcup_{(a,b) \in S_0} T_i(a, b) =$$

$$\bigcup_{(a,b) \in S_0} \{(a', b') \in S_i \ : \ \text{either } (a', c) \in T_{i-1}(a, b) \text{ or } (c, b') \in T_{i-1}(a, b) \text{ for some } c\}$$

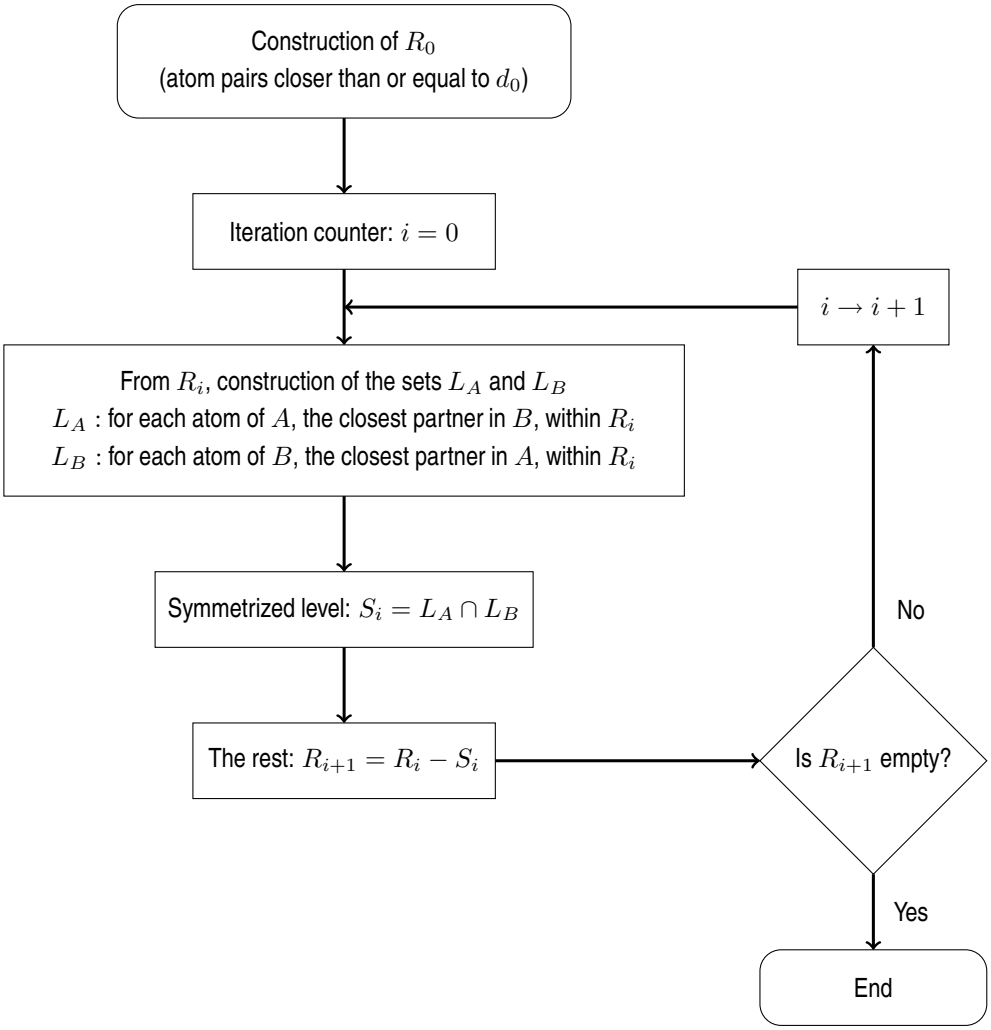


Figure 1. Flow chart of the symmetrization algorithm.

The symmetrized set S_i does not depend on which ground pair (a, b) is considered so the union operation goes inside the braces producing (in two positions) the term

$$\bigcup_{(a,b)\in S_0} T_{i-1}(a, b) = S_{i-1} \tag{23}$$

We have supposed that this holds true so we can write

$$T = \big\{(a', b') \in S_i \ : \ \text{either } (a', c) \in S_{i-1} \text{ or } (c, b') \in S_{i-1} \text{ for some } c\big\} \tag{24}$$

The condition inside the braces is actually the statement of theorem 3 so it is satisfied by all the pairs of S_i, that is precisely the thesis. ■

Corollary 3. *The union of all towers covers R_0:*

$$\bigcup_{(a,b)\in S_0} T(a, b) = R_0 \tag{25}$$

Proof. It is a direct consequence of (22) and (17). ■

Notice that the definition of tower (20) allows the case of bonds belonging to a symmetrized set higher than the ground to be in more than one tower. Thus, two towers can have nonvanishing intersection and the union in (25) redundantly covers the initial set of bonds. On the other hand, it is very easy to provide examples of non intersecting towers: just consider two groups of atoms whose inter-group distances are all larger than the cut-off. The meeting level of two towers is the lowest integer at which they share one or more bonds:

$$T_\ell(a,b) \cap T_\ell(c,d) \neq \{\}\,, \qquad 0 \leq \ell < M\,, \qquad (a,b),\ (c,d) \in S_0 \tag{26}$$

The towers will provide useful information to interpret the bonds organisation of the interface. We define the following relation $\mathcal{R}$ between members of R_0 by saying that two bonds are in relation if one can walk from the first to the last by always traversing intersecting towers:

$$\begin{aligned}(a,b)\mathcal{R}(c,d) \quad \Leftrightarrow \quad & \exists \quad \big\{(e_i,f_i) \in R_0\,, \quad i=1,\ldots,N\big\} \quad \text{for some } N \text{ such that} \\ & \hat{T} \cap T(e_1,f_1) \neq \{\}\,, \quad T(e_1,f_1) \cap T(e_2,f_2) \neq \{\}\,, \quad T(e_2,f_2) \cap \ldots \neq \{\}\,, \\ & \ldots \cap T(e_N,f_N) \neq \{\}\,, \quad T(e_N,f_N) \cap \tilde{T} \neq \{\}\end{aligned} \tag{27}$$

where $\hat{T}$ is a tower that contains (a,b) and $\tilde{T}$ is one that contains (c,d).

Theorem 5. *The binary relation $\mathcal{R}$ is an equivalence relation.*

Proof. First, the relation is reflexive $(a,b)\mathcal{R}(a,b)$, as a tower intersects itself. Second, it is symmetric

$$(a,b)\mathcal{R}(c,d) = (c,d)\mathcal{R}(a,b) \tag{28}$$

because the set intersection is commutative. Third, it is transitive

$$(a,b)\mathcal{R}(c,d) \text{ and } (c,d)\mathcal{R}(e,f) \Rightarrow (a,b)\mathcal{R}(e,f) \tag{29}$$

To see this, one can simply walk back one of the paths of the left hand side, join it with the other and use the joined path to connect the two bonds on the right hand side. ■

This equivalence relation partitions the set R_0 into equivalence classes, members of the quotient set

$$Q = R_0/\mathcal{R} = \{[(a,b)] : (a,b) \in R_0\} \tag{30}$$

Two different equivalent classes represent two groups of bonds that cannot be connected by a bond in R_0. By (1), this means that the respective atoms are separated from each other of more than the cut-off distance (1), thus forming two separated domains of points. They are also called patches, or regions of the interface. By the definition itself, it is obvious that a tower belongs to a single patch. Vice versa, a patch in general contains several towers; actually it contains all the towers that have a (pairwise) nonvanishing intersection.

Two towers at a level lower than the meeting one (26) do not intersect. Thus the bonds belonging to levels lower than the intersection one are well separated and form clusters around the two level 0 bonds that generate the towers. Let ℓ the lowest meeting level for a given tower, namely the lowest integer that satisfies (26) for the tower $T(a,b)$. Then, all its bonds that are ranked at a level lower than ℓ belong to the cluster $\mathcal{C}(a,b)$ that emanates from (a,b)

$$\mathcal{C}(a,b) = \{(e,f) \in T(a,b) : (e,f) \in S_i\,,\ 0 \leq i < \ell\} \tag{31}$$

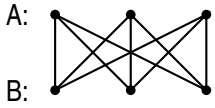
The bonds at the meeting level or higher are not attributed to this cluster, because they are equally well described as members of the nearby cluster $\mathcal{C}(c,d)$. One can go further. Following (26), one can look for the level ℓ_1 at which a third tower is met, thus form a larger cluster. The construction can go on and will stop when the full patch will be included in the cluster. This hierarchical division into clusters is important because it provides a notion of vicinity between bonds, and thus between atoms, based on the ranks and on the meeting levels.

Complexity of the symmetrization algorithm

For simplicity, we suppose that the two subunits A and B have the same number of atoms n. We compute the complexity of the algorithm by counting the operations needed for the construction of the sets R_i and S_i with $i = 0, 1, \cdots, M-1$.

For the set R_0, the expression (1) considers the pairs of atoms at a distance lower than a given cut-off. Thus, for each atom $a \in A$ we must evaluate the distance to each atom $b \in B$ and make the choice to include or not the pair (a, b) in R_0 by comparing with the cut-off. This construction of R_0 requires n^2 operations. Particularly, in the worst case the cardinality of R_0 is quadratic and of the order of n^2, $O(n^2)$.

The worst case arises when the two subunits are strongly connected; in graph theory such event is called a complete bipartite graph, that occurs when each atom of A is connected to each atom of B, as in the following example



When all the distances between pairs of atoms are less than the cut-off, R_0 turns out to be a complete bipartite graph with exactly n^2 pairs of atoms.

Now, the symmetrization uses the sets defined in (4) and (5). In order to construct L_A, for each atom a we consider all the atoms b such that the pair (a, b) is in R_0. In the worst case we need to consider all the pairs between atoms of A and atoms of B. Thus the construction of L_A uses n^2 operations. The same argument holds for the construction of L_B. The intersection of L_A and L_B requires once more n^2 operations, in the worst case. Thus, at worst, the symmetrization at level 0 needs $3n^2$ operations, that is $O(n^2)$ operations. Notice that the cardinality of L_A and L_B can be estimated to be of order n. Actually, it is precisely n if the distances are all different. The size of S_0 cannot be greater that the one of L_A, L_B; it is expected to be of order n or lower.

A new level is built with the subtraction indicated in (12). This operation is estimated to remove a linear number of pairs of atoms (that is $O(n)$ pairs of atoms). As the removal of a linear set from a quadratic set doesn't affect the order, the cardinality of R_1 turns out to be quadratic. We will use this counting at all levels, claiming that the size of all the R_i is quadratic in n and their symmetrization needs $O(n^2)$ operations. Clearly, this is an overestimation, because we know that the cardinality of R_i decreases, with i, down to zero (16) thus the number of operations needed by its symmetrization decreases accordingly. A more careful counting is possible but has no effect on the final order.

The maximum value of the number of levels in (16) is $M = n^2$, corresponding to a case where at each level we remove only one pair of atoms from R_0, precisely the one that attains the global minimum (15). In practice, the actual number of levels is often less than n (see the example given later).

The total number of operations is evaluated to $n^2 + Mn^2$, that amounts to $n^2 + n^2 n^2$, given the maximum value of M. This means $O(n^4)$ operations for the whole construction.

Comparison with connected components decomposition

It is very natural to express the notion of interface by the form of a graph. Its set of vertices is

$$V = A \cup B \tag{32}$$

The set of edges (bonds) is R_0 itself. As we only consider bonds between A and B, the graph is automatically bipartite. We do not need to consider loops (bonds joining a vertex with itself). Also, the graphs are simple, as we do not attach any meaning to parallel edges.

In [Feverati, 2010] and in later publications, we have systematically represented the level S_0 of interface proteins by undirected bipartite graphs (Gemini graphs) and carefully investigated their properties.

The connected components are a global decomposition of a graph into subgraphs, where the reference length used to separate the different components is the same everywhere. The symmetrization, instead, is a local decomposition of the graph into subgraphs because the ranking is decided by the local features of the graph.

In connected component analysis, a "running" cut-off is used in order to evaluate how far two (or more) parts of the interface are. In the case of an interface, we could introduce a cut-off f with $0 < f < d_0$ in order to consider only pairs of atoms with distance less than f. This construction helps to investigate the structure of the interface and can be used to propose pairs of atoms considered crucial to maintain the interface. On the other hand, it is somehow artificial because the cut-off is not aware of the local "geometry". For example, a very densely populated region and a sparse one are treated on an equal ground by a global cut-off while they are ranked differently by the symmetrization.

Thus, the stratification of R_0 into levels, produced by the symmetrization, is another way to investigate the pertinence of the interface. In particular, there are atoms which are involved at many levels in the towers; these atoms could be important in the construction of the interface itself. We will use all this information in future publications to investigate a new measure of the stability of the interface and to focus on pairs of amino acids with atoms appearing at many different levels of the towers.

Discussion

Example

The subunit B of the cholera toxin protein (CtxB) is a pentamer made of identical chains of 103 amino acids, showing the cyclic group C_5 symmetry. We use the Protein Data Bank structure 1EEI, where the chains are enumerated from D to H. From now on, unless explicitly specified, all the results and the discussion are based on the Euclidean distance in $\mathbb{R}^3$. The raw interface R_0 (1) contains 740 bonds of length smaller that 5 Angström. This interface decomposes in 30 levels S_i, $i = 0, 1, \ldots, 29$ (17), as shown in table 1.

The lowest levels of the towers are indicated in Figure 2 and 3.

At the level 0, one recognizes 10 black straight line segments, namely 10 clusters. Six of them join at level 2, by a purple. Two join at level 3, by a red segment. Two clusters are still disjoint and will join at some higher level.

Table 1. Decomposition of the interface in levels, for the CtxB protein (1EEI). In average, each symmetrized level has 24.7 bonds. The total interface R_0 has 740 members.

level	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
bonds in the level	33	39	34	32	33	37	32	29	30	32	34	38	38	35	25
level	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
bonds in the level	29	33	36	25	27	24	17	13	8	7	6	4	4	4	2

Table 2. Effect of the variation of the cut-off in the range [5,25] Angström. The table indicates the number of pairs in S_0, at three different values of the cut-off d_0. The test has been performed on 40 proteins of stoichiometries from 3 to 8. We show one example per stoichiometry.

$d_0 \downarrow$ name $\rightarrow$	1PM4	1J8D	1EEI	1U1S	1HX5	1Q3S
5 Angström	15	36	33	22	18	72
6.5 Angström	16	36	33	23	19	75
25 Angström	18	36	33	23	19	79

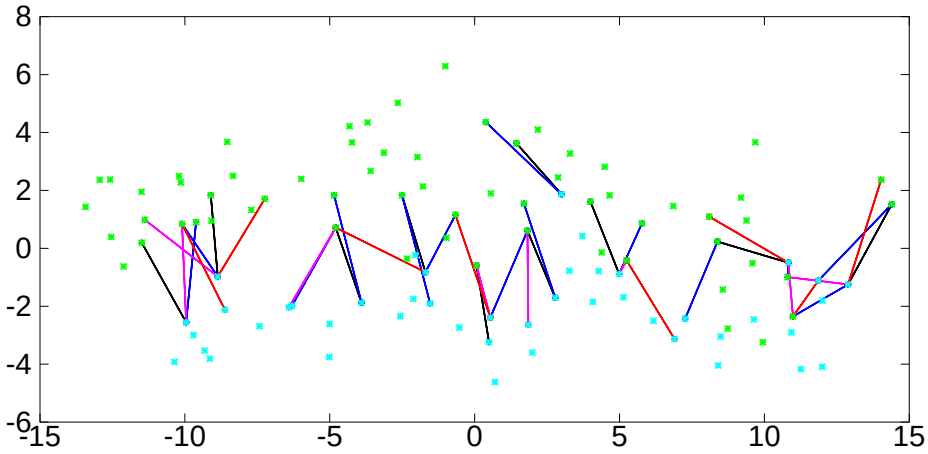


Figure 2. A planar projection of the interface of the CtxB protein is shown; axis coordinates are indicated in Angström. The projection has been chosen to maximise the visibility of the interface. Only atoms belonging to the amino acids in the range $[96, 103]$ in chain D and $[23, 31]$ in chain E are shown. We restrict to it in order to have an image of reasonable size. The four lower levels are shown in colors: S_0 black, S_1 blue, S_2 purple, S_3 red. The stars represent atoms: cyan for the subunit D, green for the subunit E. The unconnected atoms will all join at some level higher than level 3. Here and in the next figure, please notice that the apparent crossing of the straight line segments is an artifact of the projection; in the three dimensional space they actually avoid each other (in general).

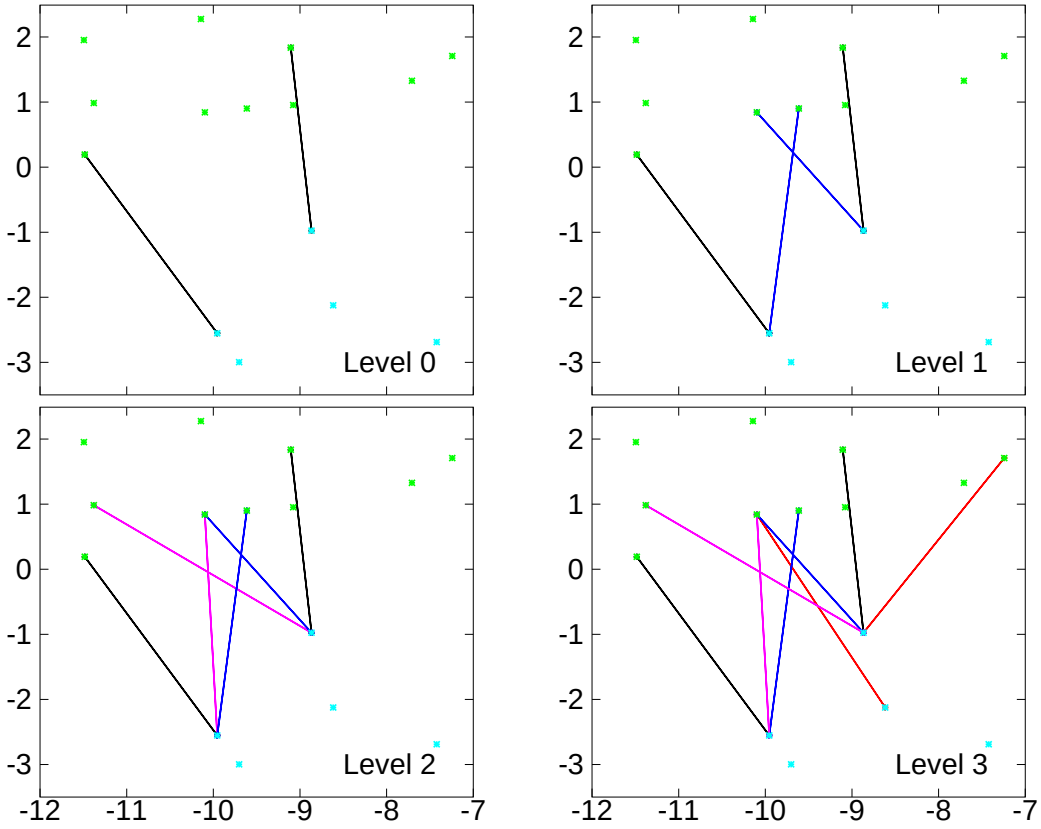


Figure 3. Zoom on the left part of Figure 2. Each image adds one level on top of the previous ones. Two bonds appear at level zero and form the basis of two towers. They meet at level 2, thanks to the purple segment closest to “vertical”, that joins one atom of each tower.

Features of the symmetrization

We present now the features of the symmetrization that we find important in studying the protein interfaces. We also suggest that the symmetrization could help to treat problems issued of different domains, like other biological interfaces.

1. The whole construction is independent on the explicit distance function. In relation to protein interfaces, this means that one could replace the three-dimensional Euclidean distance by some other distance that knows about the actual interaction energy of the pair. This would allow to interpret the ranking in a strict sense: the higher the rank the weaker the strength[1]. Unfortunately, this goal is not easily realized, the major problem being that part of the interaction energy between subunits is accumulated as torsion or bending of covalent bonds. This requires a description based on three or four atoms while the distance is a function of two atoms. Said otherwise, the correct interaction energy is not just the sum of pairwise interactions but also contains three and four atom terms. While here we remain with the Euclidean distance, we plan to perform, in a future publication, several tests in order to appreciate the relative importance of two, three and four atom interaction terms and present a distance expression that (somehow) simulates their effects.

2. The symmetrization self-adapts to the size and packing of the interface. To clarify this point, imagine we define an interface by taking (a) its ten shortest distances, or (b) the three shortest distances for each atom. In (a), the number of representatives would be totally uncorrelated to the actual interface size, while in (b) the local arrangement of points and the interaction strengths would play no role. In particular, an atom connected with three others or a second atom connected with ten others would be treated in the same way. We have observed an enormous variability in the type, number and local organization of atoms or amino acids present in an interface so we mandatorily need to avoid situations as those described in (a) or (b).

3. The symmetrization makes the set S_0, the lower levels content and the clusters (31) weakly dependent upon the cut-off value. Indeed, if $d_0 \to \infty$ (or simply higher that the largest distance in the data set), the full construction becomes cut-off and scale independent. Decreasing it down to a finite value, still sufficiently large by respect to the pairs in S_0, is expected to have a limited influence on the lowest levels, because the whole procedure is based on minimizations while, if we decrease the cut-off, we just slice out widely separated pairs. The validity of this argument strongly depends on the actual data set. In the domain of protein interfaces, it is strongly supported by the data in table 2.

   This indicates that a good strategy to choose the cut-off consists in fixing it to a value such that the effect of a variation on the lowest levels (before meeting) is negligible. Physically, this represents the very common situation where the interaction of sufficiently far apart objects is screened by the interposed ones and the contribution to the bond energy becomes negligible. This is why we choose $d_0 = 5$ Angström in order to capture intermolecular interactions at the interfaces.

4. The symmetrization applies to all scales, namely one could replace the atoms by some interacting entity at some distance scale. Indeed, the only scale length in the problem is the cut-off, that can be removed by simply sending it to infinity. This can be particularly interesting in analysing problems where two or more different scale lengths appear.

5. As the levels are obtained by minimization, they are "fragile" to perturbations or experimental errors, if these are sufficiently large to swap the relative distances of two atoms. Most probably this is not a serious issue if clusters are considered. A more complete analysis will be performed in later publications.

[1] Notice that the triangular inequality is needed to supports this interpretation.

Connection with biology

Having adopted the Euclidean distance, the symmetrization produces a ranking that is purely geometric, because atomic positions only matter while the atom and interaction types do not play any role. The F. Crick approach cited in the Introduction was also purely geometric. If some "improved" distance based on the actual chemical interactions was implemented, as suggested at point 1 in paragraph Features of the symmetrization, the ranking would become geometrical and chemical. With our purely geometric criterion, in previous papers we have show that the lowest symmetrized level S_0 contains useful information on the interfaces. Its use has been validated in [Feverati, 2010]. It has been employed in [Zrimi, 2010] to help investigating the role of histidines in the protein assembly.

More recently, in [Feverati, 2012] it has been shown that the intermolecular β-sheet interfaces are made of two interdigitated interaction networks, one involving backbone (BB) atoms only, the other involving at least one side-chain (SC) atom. The two networks have slightly different amino acids and bond compositions and some specific pattern of charged and hydrophobic amino acids. The hydrophobic amino acids observed in the BB network are similar to those observed in intramolecular β-sheets. The SC network is slightly enriched in charged amino acids and the spatial distribution of the charged amino acids in the interface has some peculiar pattern specific to the intermolecular case.

The work of J. Janin and collaborators [Janin, 2008] indicates that an interface is composed of two parts, the core and the rim. The core is completely inaccessible to the solvent molecules, typically water, while the rim is exposed to them. The distinction is geometrical, based on the position and on steric hindrance of atoms. The core is more conserved during evolution. The complementarity of the two subunits is more pronounced in the core than in the rim.

On the opposite, the symmetrization, albeit geometrical, does classify according to the strength of the interactions. The two classifications (core/rim versus symmetrization) are "dual", in the sense that they complement each other by giving different information on the same entity. For example, atoms of the core can appear at many different levels thus, in general, it would not be appropriate to claim that some levels describe the core, some others the rim.

A consequence of our choice of using the notion of the interaction strength is that we have instruments to characterize the physical neighborhood: the patches, the clusters and the meeting levels.

If the cut-off is well chosen, patches represent different regions of the interface, namely groups of atoms sufficiently far apart to be considered independent. This phenomenon is common, as quite often an interface is made of several patches. Different patches in the same interface can play different roles during assembly and possibly they have different evolutive histories. The notion of independent region or patch is somehow hard to define thus it is very important to have a procedure to construct it. The present definition is based on atomic interactions while the one we used in [Feverati, 2010] was based on the proximity along the amino acid sequence.

The clusters (31) may help to characterize fundamental blocks of an interface, at least for those clusters whose meeting level is very high. A block can mean some group of atoms that is recurrent, in interfaces of similar geometry, or that is conserved during evolution, or that has some special role in the recognition of the two subunits during the assembly.

The assembly is a complex phenomenon where the two partners, namely the subunits, need to recognize each other, then stabilize the interface. In most cases it is not known how these steps develops. Thus, the levels seem to suggest the sequence of events occurring during the assembly, the lower levels accommodating first, the higher levels later.

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.

We have proposed a new approach to the treatment of protein interfaces, based on the idea that the ranking of interactions helps clarifying their role. Our previous experience with the lowest level S_0 has been so fruitful that we believe the full set-up constructed here will allow to improve the description of interfaces.

The main tool, in our view, is the cluster. Indeed, it identifies the locally strongest bonds and the meeting level indicates how rich are the interconnections within each cluster.

Higher levels, clusters, towers and patches are presented here for the first time, in a proper mathematical development. The hope is that their full biological relevance will appear after a comparative analysis of many interfaces of similar geometry, as it has been the case with the level 0. Also, with the use of S_0 alone, as we did in our previous papers, it was not possible to evaluate the role of the interactions that were neglected. The need to address this issue has strongly motivated the present work.

Acknowledgements

It's a pleasure to thanks Alexander Grossmann for his most welcome comments.

Bibliography

[Goodsell, 2000] D.S. Goodsell, A.J. Olson. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29: 105-153, 2000.

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

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

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

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

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

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

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

[Feverati, 2012] G. Feverati, M. Achoch, J. Zrimi, L. Vuillon, C. Lesieur. β-strand interfaces of non-dimeric protein oligomers are characterized by scattered charge residues pattern. PLoS ONE 7(4): e32558, 2012.

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

Authors' Information

Giovanni Feverati *- LAPTH, University of Savoie, CNRS, BP: 110, Annecy-le-Vieux 74941, France; e-mail: feverati@lapp.in2p3.fr*
Major Fields of Scientific Research: Mathematical physics, theoretical biophysics, models of Darwinian evolution, models of protein assembly.

Claire Lesieur *- AGIM, University of Grenoble, CNRS, Archamps, France;*
e-mail: claire.lesieur@agim.eu
Major Fields of Scientific Research: Biophysics, protein assembly, kinetics, dynamics, interfaces.

Laurent Vuillon *- LAMA, University of Savoie, CNRS, Le Bourget du lac, France;*
e-mail: Laurent.Vuillon@univ-savoie.fr
Major Fields of Scientific Research: discrete mathematics, combinatorics, discrete dynamical systems.