Protein subunit association: NOT a social network

Mounia Achoch, Giovanni Feverati, Laurent Vuillon, Kavé Salamatian, Claire Lesieur

Protein subunit association: NOT a social network[*]

Mounia Achoch[†]
LISTIC,University of Savoie, Annecy le Vieux, France

Giovanni Feverati[‡]
LAPTH UMR 5108, University of Savoie, CNRS, Annecy le Vieux, France

Laurent Vuillon[§]
LAMA UMR 5127, University of Savoie, CNRS, Le Bourget du Lac, France

Kave Salamatian[¶]
LISTIC, University of Savoie, Annecy le Vieux, France

Claire Lesieur[‖]
AGIM FRE 3405, University of Grenoble Alpes, CNRS, Grenoble, France

4th of April 2014

Abstract

Most proteins cannot function as single unit but associate subunits via the formation of protein interfaces, to be biologically active. How the amino acids involved in subunit association, so-called hot spots, regulate the formation of a protein interface is still an open question. Here, we show how network and graph theories can help addressing the role of hot spots. We built a MatLab code called SpectralPro which identifies hot spots and reconstructs the protein interface as a subnetwork of hot spots in interaction, with the hot spots as nodes and the bonds between hot spots as links. Using as a case study, the cholera toxin B pentamer (five subunits), we investigate if the degree of a node, namely the number of contacts of a hot spot, is important in the formation of an interface. The degree of a node is known to be important in many real networks. For example in social networks, hubs control the communication between most nodes and as such are vulnerable to changes. But our result shows that in the toxin interface sub-graph hub-like nodes are less vulnerable to change than single link node.

[*] Work supported by the Federation de recherche FR2914, MSFI, Modelization, Simulations, Fundamental interactions

[†] e-mail address: Mounia.Achoch@univ-savoie.fr

[‡] e-mail address: feverati@free.fr

[§] e-mail address: laurent.vuillon@univ-savoie.fr

[¶] e-mail address: kave.salamatian@univ-savoie.fr

[‖] e-mail address: claire.lesieur@agim.eu

1. Introduction

Proteins are biological entities made of a chain of amino acids bound to one another in a specific order, called the primary structure or the amino acid sequence of the protein. Based on the sequence and the environment, the protein acquires a tridimensional shape called tertiary structure (3D-structure), suitable for its biological function. The set of reactions leading to the functional 3D-structure is the folding of the protein. It involves the formation of bonds/interactions between atoms of the amino acids of a single chain. These interactions are called intramolecular amino acid interactions. There exist proteins which function as oligomers by associating several copies of the same chains (homo oligomers) or of different chains (hetero oligomers). The association of chains forms the quaternary structure (4D-structure) of the proteins. The zone of contact between two associated chains is called the protein interface. The protein interface involves the formation of interactions/bonds between atoms of the amino acids of adjacent chains. These interactions are called intermolecular amino acid interactions. Among the amino acids involved in intermolecular amino acid interactions, only a subset is important for the formation of the interface, those are called hot spots [1].

Some protein oligomers are involved in diseases as virulence factors, like the notorious cholera toxin responsible for the cholera disease [2]. Understanding and predicting how such proteins assemble into oligomers is essential for designing appropriate inhibitors capable of preventing their pathological assemblies. The design of such inhibitor entails to identify the hot spots and understand their role in the formation of an interface. There are numerous algorithms capable of identifying hot spots from the 3D structure of protein oligomers whose atomic coordinates are available from the Protein Data Base (www.rcsb.org/pdb/). However, these algorithms do not provide means to understand how the hot spots orchestrate the formation of an interface. We propose to consider hot spots as nodes and bonds between hot spots as links, and to build a subgraph or a subnetwork of hot spots in interaction to model the interface. Sub graph because it describes only a local feature of the protein chain, namely the interface and not the entire chain, which would be a graph. The hot spots can be distinguished by network measures and we can look for correlation between the network's measures and the importance of the hot spots in terms of interface formation. A good overview of network measures can be found in [3]. Our case of study is the cholera toxin B subunit pentamer (CtxB$_5$) produced by *Vibrio cholera.* We have written a Matlab code that reasonably identifies the hot spots of the CtxB$_5$'s interface and builds a sub-graph of the toxin's interface based on a matrix of contacts. We look if the degree of the nodes, namely the number of contacts of the hotspots, has any relevance in terms of the formation of the toxin's interface.

2. Methods

SpectralPro. SpectralPro uses the Cartesian coordinates of the atoms of the 3D-structure of CtxB$_5$ as an input. These coordinates can be extracted from the PDB under the PDB code 1EEI. Each chain of the pentamer is considered as a set of points in the space whose positions are the Cartesian coordinates (x, y, z) of the atoms of the chain. The atoms of the chain 1 constitute the set 1 (S1), the atoms of the chain 2, the set 2 (S2) and the atoms of the chain 5, the set S5. SpectralPro calculates distances between every atom of S1 and every atom of the four other sets (interchain distances) but ignores the distances between atoms of a single set (intrachain distances). It chooses for every atom the 10 closest atoms and among these, it selects the pairs of atoms distant of a maximum of 5 Angstrom. Every atom is involved in a certain number of pairs, namely it has a certain numbers of contacts. SpectralPro builds a N x N matrix with the selected intermolecular atoms as the nodes N and the elements of the matrix as their number of contacts. SpectralPro also builds a coarse-grained matrix where the atoms are replaced by their respective amino acids as nodes. A weightless matrix is produced where the elements of the matrix are one when the amino acids have at least one pair of atoms in contact and zero when they don't. The weightless matrix provides for every amino acid, its number of amino acid contacts.

Fold X. The effect of a local change (amino acid mutation) on the formation of the toxin interface is measured by generating a virtual single point mutation on the toxin PDB with Fold X and by calculating the free energies of interactions at the interface for the non mutated (wild-type) and the mutated proteins [4]. The difference between the two energies measures the effect of the mutation. The amino acid plays a role in the formation of the interface if its mutation leads to a non zero energy difference.

3. Results and discussion

The goal of the investigation is to develop an appropriate tool to reconstruct the CtxB$_5$ interface as a sub-graph of hot spots in interaction, analyze some graph properties to determine their relevancy in terms of the toxin assembly.

3.1. Identification of hotspots

The first step is to test if SpectralPro is capable of identifying hot spots. The details on how SpectralPro detects amino acid in contact is described in the methods. Because SpectralPro reads the atoms following the amino acid sequence of the chain and selects the closest atoms, it retraces a good reading of the geometry of the two surfaces that make the interface compared to a selection based simply on a cut-off distance. The cut-of distance at 5 Angstrom applied subsequently allows to choose the bonds the most chemically probable. It is unlikely that every atom makes ten chemical bonds (ten closest atoms), but the ten links provide a density of interactions instead of evaluating an exact number of interactions. The idea is to obtain an estimate of a probability of interactions of the amino acids. The coarse-grained amino acid sub-graph is built on a square matrix having as rows and columns the amino acids, ordered according to their location along the sequence. The elements of the matrix at position i, j have a one entry if the i-th and j-th amino acids have at least one pair of atoms in interaction (weightless sub-graph).

The sub-graph of the atoms in interaction over the five interfaces of the pentamer has 1498 nodes and 2830 links. In other words, the sub graph is made of 1498 atoms with 2830 closest atoms. The coarse-grained sub-graph of the amino acids in interaction has 283 nodes and also 2830 links (weighted sub-graph). Thus on average every atom has two closest atoms located within 5 Angstrom distance and every amino acid has about five atoms involved in a pairwise interaction. If a single link is counted for every pair of amino acids, the (weightless) sub-graph has 283 nodes and 422 links. To have an idea of the order of magnitude of a protein interface sub-graph, it is interesting to compare with the world wide web which has 200 million nodes (webpages) and 1.5 billion links, links between two pages.

The amino acids selected as in interaction by SpectralPro are compared to the detection of hot spots by three other available programs (not shown). SpectralPro identifies 283 amino acid contacts over 5 interfaces, with an average of 57 ±1 hot spots per chain. If we consider the set S5, namely the chain E, SpectralPro identifies 56 hot spots against 39, 57 and 54 for Gemini, PSIBASE and SCOWLP, respectively. Gemini detects hot spots by selecting the mutually closest atoms yielding a more stringent selection than SpectralPro and less hot spots identified [5]. All hot spots detected by Gemini are identified by SpectralPro. PSIBASE as SpectralPro calculates the Euclidean distance to determine pairs of interactions [6]. SpectralPro identifies all the hot spots identified by PSIBASE expect three, making about 5 % false negative. Only one amino acid detected by SpectralPro is not detected by PSIBASE, making less than 2 % false positive. On average in PSIBASE, every hot spot has 5 atoms involved in a pairwise interaction as observed for SpectralPro. SpectralPro identifies all the hot spots identified by SCOWLP expect one, making less than 2 % false negative. There are three amino acids detected by SpectralPro but not by SCOWLP, making about 5 % false positive. SCOWLP identifies pairwise interactions using Eucledian distances and shape-based algorithms [7]. Globally the amino acids selected as hot spots by SpectralPro are consistent with those identified by other programs, supporting that SpetralPro detects hot spots reasonably.

3.2. The degree measure

On a previous study on a large dataset of 1048 interfaces involving the interactions between two beta -strands, we had measured the degree of the nodes of the sub-graph interfaces and looked at the degree distributions [8]. The sub-graphs were built with a different algorithm, called Gemini which selects only a framework of interactions, as mentioned above. The result indicates an exponential degree distribution, no hubs and many nodes with one to three contacts. We have determined statistically that the only amino acids with more than three contacts are R, Y, L and W.

Now we look whether this result is confirmed using SpectralPro which sets less stringency on the selection of hot spots and the number of contacts. The average number of contacts $\overline{k}$ over the five CtxB$_5$ interfaces is 3.1 ±1.8. Thus even with SpectralPro, the average number of contacts per residues remains around three.

The degree distribution *P(k)* is the number of hot spots with k degree plotted against the degree k. *P(k)* is calculated for each of the five interfaces of CtxB$_5$ and the average degree distribution and standard deviation is plotted against the degree (Figure 1).

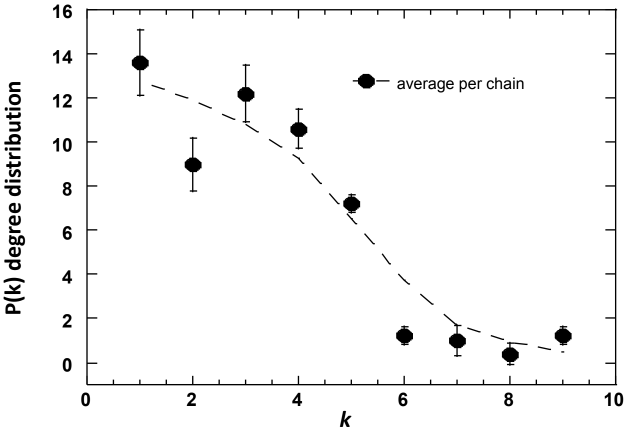

Figure 1: Degree distribution

P(k) for the sub-graph of the CtxB$_5$ interface follows a bell like shape which corresponds to a random network with no hubs but nodes with few links. Again this confirms the observation made on the dataset using Gemini that interface subnetworks do not follow power law degree distribution and have no hubs.

At most the hot spots have 9 contacts, and there are only two such nodes, the amino acid arginine 67 (Arg67) and the leucine 31 (Leu31). Thus the bigger ratio between the highest and lowest degree in the sub-graph is 9. On a subgraph of the WWW of 325 729 nodes, which follows a power law degree distribution, the average $\overline{k}$ is 5.46, the ratio between the lowest and highest node degree is 10000. So the hot spots with 9 contacts might be better referred to as hub-like rather than hub. Interestingly, in comparison the average degrees $\overline{k}$ of the two networks appear rather similar, illustrating the difficulty in interpreting average $\overline{k}$ values for different types of degree distribution. This is discussed in [9].

3.3. Influential nodes

We then explore if the degree of the nodes is any relevant to the formation of the toxin interface. For this purpose, a hot spot with a single contact, lysine 69 (Lys69) and a hub-like hot spot, Arg67 are virtually mutated to an asparagine (Asp, N) using Fold X [4]. The free energy of interaction at the interface is calculated for the mutant and the wild type (WT) proteins. The effect of the mutation is measured as the difference between the wild type and mutant free energies of interaction at the interface. Differences not equals to zero indicate that the mutated hot spot is involved in the formation of the interface. Asparagine is chosen because it has "average" amino acid properties, so if a mutation has no effect on the free energy of interaction, it indicates that the mutated hot spot has average property for the formation of the interface and is plastic to mutation. If a mutation has an effect, the mutated hot spot must have an involvement in the formation of the interface above average, this hot spot can be considered more influential for the formation of the interface and less plastic to mutation. The WT, Lys69Asp and Arg67Asp free energies of interaction are -13,35; -19, 65 and -16, 65 kcal/mol, respectively, as determined by Fold X. This shows that the hot spots are not equally important for the formation of the interface, suggesting their different roles. The free energy of the interface has decreased by a factor of 0.4 and 0.2 upon mutation of the Lys69 and Arg67, respectively. The largest mutational effect on the free energy is for the Lys69Asp mutant over mutation of all other amino acids of the toxin (not shown). Thus the mutation of the single link hot spot Lys69 has more effect on the interface than the mutation of the hub-like Arg67. Thus in contrast to social networks and other real networks, in the sub-graph of the toxin interface, the influence of a node is not directly linked to its degree. More precisely, hub-like residues are not more vulnerable to change, namely mutation, than single link node.

4. Conclusion

In conclusion, we can say that protein interface subnetworks have very different scales compared to other real networks, much less links, lower ratio high degree/low degrees, no hub and behave rather like a random network. Thus to infer "biological rules", such as the mechanism of assembly or the formation of interfaces, one cannot simply use the network measures that regulate other real networks (www or social network). Intuitively, we could have expected that hub-like hot spots would have been the most influential for the formation of the interface and highly susceptible to mutation as demonstrated for other real networks [10], but that is not the case . Here the result shows that connected does not imply influential in the case of protein interface networks. It remains to be established what makes a node influential if not its degree and to analyze the effect of the mutation on the network.

References

[1] Clackson T, Wells JA, *Science* 267(5196):383-6 (1995)

[2] Hirst TR, *J. Moss BI, M. vaughan and A. t. Tu, editor. New York: M. Dekker* 123-84 (1995)

[3] Barabasi A.L, Oltvai Z.N, *Nature reviews Genetics* Feb;5(2):101-13 (2004)

[4] Guerois R, Nielsen JE, Serrano L, *Journal of molecular biology* 320(2):369-87 (2002)

[5] Feverati G, Lesieur C, *PloS one* 5(3):e9897 (2010)

[6] Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, et al, *Bioinformatics* May 15;21(10):2541-3 (2005)

[7] Teyra J, Doms A, Schroeder M, Pisabarro MT, *BMC Bioinformatics* 7:104 (2006)

[8] Feverati G, Achoch M, Vuillon L, Lesieur C, *PloS one* in press (2014)

[9] Newman ME, Strogatz SH, Watts DJ, *Phys Rev E Stat Nonlin Soft Matter Phys* Aug;64(2 Pt 2):026118 (2001)

[10] Albert R, Jeong H, Barabasi AL, *Nature* Jul 27;406(6794):378-82 (2000)