How much Proteins Contact Networks are eccentric? A signaling networks perspective of PCN eccentricity.

Gabriele Oliva, Sergey Kirgizov and Luisa di Paola

Abstract

Protein functionality is based on signal transmissions across the intramolecular interaction networks. This mechanism strictly mirrors the signal transmission efficiency of technological networks, based on shortest paths between network nodes. In this framework, the analysis of protein structures in terms on signaling networks embedded in structures may help to outline key functional features. One of the key descriptors for efficiency of signal transmission in network is the network eccentricity, defining the scale of signal transmission, often declined in a specific definition in terms of network volume. In this work, we present an extensive analysis of protein contact networks in terms of their eccentricity: we made a comparative survey of eccentricity metrics with other structural (volume) and topological (degree, shortest path lengths, betweenness) feature of protein structures. Results point to a strong correlation of eccentricity on signaling network properties (average betweenness and average shortest path length), tracing out a strong similarity between protein contact networks and high-performance technological networks.

Gabriele Oliva
Unit of Automatic, Department of Engineering, University Campus Bio-Medico of Rome, via Ávaro del Portillo 21, 00128 Rome, Italy. e-mail: g.oliva@unicampus.it

Sergey Kirgizov
LE2I UMR6306, CNRS, Arts et Métiers, University Bourgogne Franche-Comté, F-21000 Dijon, France e-mail: sergey.kirgizov@u-bourgogne.fr

Luisa di Paola
Unit of Chemical-physics Fundamentals in Chemical Engineering, Department of Engineering, University Campus Bio-Medico di Roma, via Álvaro del Portillo 21, 00128 Rome, Italy. e-mail: l.dipaola@unicampus.it

1 Introduction

Proteins work as smart, complex systems and their ability to reply to environmental cues is strictly related to their plasticity [1]. Protein functionality relies on complex and systemic regulatory processes. As a consequence of their systemic nature, events are felt far from where they occur (the so called *allosteric effect*, see for instance [2] and references therein). This feature relies on short and efficient signaling pathways, which intervene as well in protein folding. As a result of the folding, protein structure and, in particular, their volume are regarded as key factors to explain protein functionality, although it is not straightforward to define protein volume metrics, which may also include information on the inner structure and compactness [3].

Molecular volume determination usually passes by computing the proteins convex hull (i.e., the smallest convex set containing the protein) [4, 5], or the sum of the volumes of the single atoms, for instance, via a Voronoi Tessellation [6, 7].
An alternative approach is based on the complex networks paradigm, which provides a toolbox to unveil the inner mechanisms of protein structure and function. Specifically, the *Protein Contact Networks* (PCN) proved to be an effective tool to reveal emerging properties of protein structures; such a formalism represents proteins as networks of non-covalent intramolecular interactions [8, 9, 10, 11, 12]. In these works, we linked the properties of PCNs to the structural features of protein structures (e.g., their shape), suggesting that the general shape of proteins does not affect its inner structure, a confirmation of the modular, hierarchical nature of protein molecules, as noted also in [13].

The complex networks methodology allows to describe the reachability and centrality of the nodes. In addition, the analysis of the *shortest paths*, i.e., the set of links connecting pairs of residues with the smallest cardinality, plays a pivotal role in the comprehension of signaling pathways and folding [14]. Several topological descriptors stem from the notion of shortest path, each with its own peculiarities; for instance the *betweenness centrality*, defined for each node as the number of shortest paths passing by it, allows to identify functional residues in protein structures [15, 16].

Eccentricity [17, 18] is yet another fundamental shortest-path-related descriptor of the structure of a network-the eccentricity of a node is defined as the maximal length of the shortest paths between the given node and any other node in the graph. In a wide spectrum of networks, ranging from technological to social networks, eccentricities play a crucial role: for instance in wireless sensor networks it is used to tune network protocols (e.g., setting an adequate time-to-live of packets), to select local coordinators, or to execute distributed algorithms, which typically depend on these parameters (see for instance [19, 20] and references therein); in social networks, instead, eccentricities have been used to understand how people is influenced by their acquaintances (see [21] for a recent analysis on influence over Twitter).

In this chapter we argue that shortest paths and, in particular, eccentricities, are effective tools to assess the volume of a network: if we assume a link connecting two nodes constrains the nodes to be in close spatial proximity, then dense networks will have smaller volume than sparse ones. Based on this intuition, we develop a measure V_{ecc} of volume of a PCN, which amounts to a geometric mean of the eccentricities of its nodes. To validate our metric, we consider a large database of proteins and we compare V_{ecc} and several other topological descriptors with the volume of the actual proteins, calculated via Voronoi tessellation and in terms of its convex hull.

The outline of the chapter is as follows: Section 2 details the materials and methods adopted in this chapter, while Section 3 reports the results of our analysis; finally, we collect some conclusive remarks and future work directions in Section 4.

2 Materials and Methods

2.1 Dataset

We consider a dataset of $M = 2102$ protein structures[1] listed in the *Protein Domain Server* of the Structural Bioinformatics Group at the Imperial College of London, England (see [22, 23] for details on the dataset, the complete list is available online at http://www.sbg.bio.ic.ac.uk/˜domains/). The dataset encompasses proteins with several *domains* (i.e., portions of a protein that can exist independently), from 1 to 5.

2.2 Protein Volume Calculation

To validate the proposed topological measure of volume of a PCN, we consider as reference two different metrics for protein structure volume:

- V_{CH}: for a given protein, we calculate the volume of the convex hull; specifically, we take the coordinates in $\mathbf{R}^3$ associated to each atom in the PDB file (see http://www.rcsb.org for details on PDB files) corresponding to the protein and we select the convex hull of such a set of positions, i.e., the smallest convex region that contains all the points.
- V_V: we calculate the volume in terms of the sum of the volume of the single atoms, where the volume of each atom is approximated by a Voronoi tessellation in 3D [6, 7]. We calculate the tessellation using the pdbremix toolbox (available online at https://github.com/boscoh/pdbremix).

[1] We take into account only the proteins in the dataset whose corresponding PCN is connected (i.e., each node can be reached by each other node via a path involving some of the edges of the graph).

2.3 Graphs

We denote by $G = \{V, E\}$ a graph composed of $n = |V|$ nodes $v_1, \ldots, v_n$ and a set of $e = |E|$ edges $(v_i, v_j) \in E \subseteq V \times V$; in this view the nodes represent a set of *entities*, while the edges represent the existence of a *relation* between pairs of entities.

2.4 Protein Contact Networks

Protein contact networks are constructed starting from the position of atoms in the PDB file. Specifically, we extract the position of the α-carbon within each residue in the PDB file describing the protein, and we calculate the Euclidean distance between pair of residues as the distance between the corresponding α-carbons. We interpret the residues as the nodes in a graph, and we establish a link between two residues if their distance is in the range $[4, 8]\AA$. In this way, only non covalent significant intramolecular contacts are included, which are likely responsible for protein response to environment *stimuli*.

2.5 Topological Descriptors Considered

In this chapter we consider the following topological descriptors for PCNs:

- *number of nodes n*: the number of nodes in the PCN;
- *average degree $\overline{k}$*: the *degree* of a node v_i represents the number of links node v_i participates to; the average degree is the average over the whole set of nodes;
- *average shortest path length $\overline{sp}$*: the shortest path length sp_{ij} is the minimum number of links connecting two nodes; the average shortest path length is averaged over the whole set of node pairs;
- *average betweenness centrality $\overline{btw}$*: the betweenness centrality btw_i of the i-th node is the number of shortest paths passing by it, the average value is averaged over the whole set of nodes;
- *average eccentricity $\overline{ecc}$*: the eccentricity of a i-th node ecc_i is defined as the maximal length of the shortest paths between the given node and any other node in the graph; $\overline{ecc}$ is averaged over the whole set of node pairs.

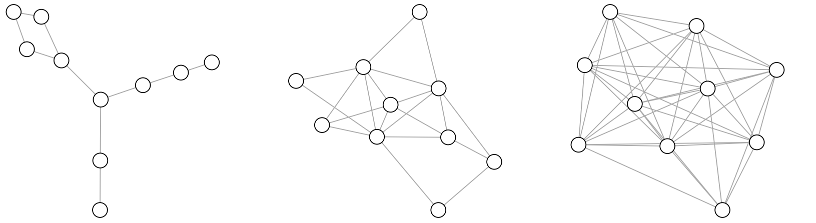
2.6 Volume of a PCN based on Eccentricities

Further to the above topological descriptors, we introduce a novel measure of the volume of a PCN, based on its topological structure and, specifically, on the eccentricities of the n nodes that compose it. We refer to such an indicator as the *eccentricity network volume* V_{ecc}, which we define as the geometric average of the eccentricities of the nodes, i.e.,

$$V_{ecc} = \sqrt[n]{\prod_{i=1}^{n} ecc_i}.$$

The above indicator has several points of contact with the average eccentricity $\overline{ecc}$. In fact, V_{ecc} can be regarded as a geometric mean of the eccentricities. The average value is always greater or equal than the geometric mean, so $\overline{ecc}$ is always greater or equal than the V_{ecc}, with the equality verified *iff* all eccentricities are equal.

When we densify our network (for example, by adding several new links) the volume of the corresponding protein should decrease. And, we may expect the similar behavior of the above PCN descriptors. This properly hold for the average eccentricity $\overline{ecc}$, for eccentricity network volume V_{ecc} (see also Figure 1), and for the average shortest path length $\overline{sp}$, but not for the average degree $\bar{k}$. The situation with the average betweenness $\overline{btw}$ is slightly more complicated: it may increase when we add links to our network, but usually it decreases. An empirical study on the relations between average betweenness and network density is conducted in [24].



(a) $V_{ecc} \approx 4.61$ (b) $V_{ecc} \approx 2.45$ (c) $V_{ecc} \approx 1.62$

Fig. 1: Eccentricity network volume V_{ecc} increases when the network density grows. Networks of 10 nodes with different number of links (from left to right: 10, 20, and 40).

2.7 Statistical Analysis

To analyze the mutual dependence of the above metrics and indicators, we applied the following statistical tools: simple correlation, partial correlation and Principal Component Analysis (PCA). Simple correlation analysis corresponds to the computation of the Pearson correlation coefficient between pairs of variables. Partial correlation, in turn, amounts to the degree of mutual correlation between two variables, removing the effect of one another variable.

Principal Component Analysis (PCA) is a statistical method that, given a set X of observations $x \in \mathbf{R}^m$ returns a set of *principal components*, i.e, mutually orthogonal vectors

$$PC = \{PC_1, PC_2, \ldots, PC_k\}$$

each in $\mathbf{R}^m$ and with $k \leq n$. The vectors are ranked in order of explained variance of the original data set, so the smaller is the index of a principal component, the more descriptive it is of the variability of the dataset. In particular, we are interested in the cumulative variance CSVAR explained by the first j principal components, for each $j = 1, \ldots, k$ and in the *loadings* associated to each principal component– a loading is the correlation between the original variable and the principal component; in other words, loadings measure the influence of the original variables in a given principal component.

3 Results and Discussion

Table 1 reports the Pearson correlation coefficients for each pair of variables, while Figure 2 shows the correlations between V_{ch} (horizontal axis) and all other descriptors.

Table 1: Correlations: significative correlations - higher in module than 0.5 - appears in bold .

Correlations	**n**	$domains$	$\overline{k}$	V_{CH}	V_V	$\overline{ecc}$	V_{ecc}	$\overline{sp}$	$\overline{btw}$
n	1.0000	0.2779	0.4680	**0.9777**	**0.9828**	**0.8366**	**0.8388**	**0.8555**	**0.9579**
$domains$	-	1.0000	0.2898	0.2686	0.2626	0.3401	0.3397	0.3401	0.2328
$\overline{k}$	-	-	1.0000	0.3868	0.4562	0.3212	0.3232	0.3501	0.3511
V_{CH}	-	-	-	1.0000	**0.9656**	**0.8817**	**0.8832**	**0.8928**	**0.9822**
V_V	-	-	-	-	1.0000	**0.8127**	**0.8148**	**0.8316**	**0.9399**
$\overline{ecc}$	-	-	-	-	-	1.0000	**1.0000**	**0.9913**	**0.8534**
V_{ecc}	-	-	-	-	-	-	1.0000	**0.9917**	**0.8546**
$\overline{sp}$	-	-	-	-	-	-	-	1.0000	**0.8641**
$\overline{btw}$	-	-	-	-	-	-	-	-	1.0000

Most descriptors strongly correlate with the number of nodes n (i.e., all but the average degree and the number of domains); noticeably, the number of domains does not correlate with any other descriptor, confirming the high modularity of protein structures, whose connectivity features are mostly independent on size or domains number [25]. The average degree $\overline{k}$, as well, shows no significant correlation with any other variable, suggesting that the average connectivity is more or less constant for proteins of different sizes, number of domains and function.

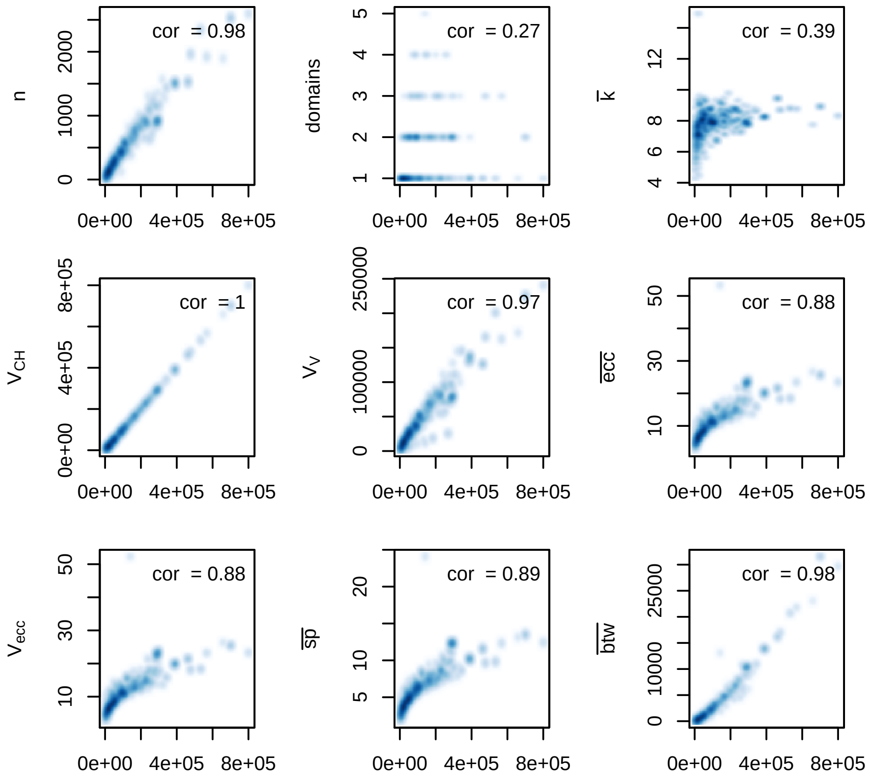


Fig. 2: Correlations between V_{ch} (horizontal axis) and all other descriptors.

Both volume metrics (V_{CH} and V_V) strongly correlate with all descriptors but the number of domains and the average degree, and so do eccentricity-related descriptors ($\overline{ecc}$ and V_{ecc}), the average shortest path $\overline{sp}$ and the average betweenness $\overline{btw}$.

Table 2: Principal Components: significative loadings - higher in module than 0.5 - are reported in bold.

Principal Components	*n*	*domains*	$\overline{k}$	V_{CH}	V_V	$\overline{ecc}$	V_{ecc}	$\overline{sp}$	$\overline{btw}$	*CSVAR*
P1	**0.9628**	0.3668	0.4615	**0.9756**	**0.9466**	**0.9449**	**0.9461**	**0.9542**	**0.9526**	74.7609
P2	-0.0464	**0.7679**	**0.6020**	-0.1151	-0.0561	-0.0705	-0.0747	-0.0619	-0.1560	85.9826
P3	0.1505	-0.4921	**0.6171**	0.0579	0.1656	-0.1963	-0.1935	-0.1617	0.0633	94.6775

To further inspect the dependency of the other descriptors on the number of nodes, we report in Table 2 the results of the PCA analysis, which aims to assess the role of each variable in explaining the variance of the data set. In the table, we report the loadings and the cumulative variance CSVAR. We show, sorted from top to bottom, only the components accounting, individually, for at least 5% of the overall variance, which in our case coincide with the first three principal components. From the results in Table 2 we notice that the loadings of the first principal component PC1 are above 0.9 for all variables but the degree and number of domains, and the component explains about 75% of the total variance; the second component, instead, individually explains around 11% of the total variance and is correlated to the number of domains and the degree; the third component, instead, has a relevant correlation only with the degree and explains an additional contribution worth about 8% of the variance. We notice that eccentricity-related metrics, alone, are not relevant in the cumulative explained variance, as further principal components (not reported in the table) explain a little fraction of the total variance.

As a result of the above PCA analysis, we obtain stronger evidence of the strong dependence of all descriptors on n, which is likely to influence the remaining correlations. Since most of the above descriptors may scale considerably with the number of nodes, and the number of nodes has strong correlations with almost all descriptors, we are tempted to believe that the remarkable correlations in place in Table 1 are strongly dependent on the number of nodes.

To highlight the specific contribution of the descriptors, we calculate the partial correlation between variables, excluding the effect of the number of nodes (Table 3). By removing the effect of n, we observe strong correlations between V_{CH} and V_{ecc} and $\overline{btw}$.

Table 3: Partial correlation w.r.t. number of nodes: significative correlations - higher in module than 0.5 - appears in bold.

Correlation	*domains*	$\bar{k}$	V_{CH}	V_V	$\overline{ecc}$	V_{ecc}	$\overline{sp}$	$\overline{btw}$
domains	1.0000	0.1882	-0.0153	-0.0594	0.2044	0.2038	0.2058	-0.1211
$\bar{k}$	-	1.0000	-0.3818	-0.0234	-0.1454	-0.1443	-0.1099	-0.3834
V_{CH}	-	-	1.0000	0.1195	**0.5545**	**0.5520**	**0.5189**	**0.7569**
V_V	-	-	-	1.0000	-0.0948	-0.0955	-0.0964	-0.0299
$\overline{ecc}$	-	-	-	-	1.0000	**0.9999**	**0.9714**	0.3308
V_{ecc}	-	-	-	-	-	1.0000	**0.9723**	0.3271
$\overline{sp}$	-	-	-	-	-	-	1.0000	0.3001
$\overline{btw}$	-	-	-	-	-	-	-	1.0000

We notice that, getting rid of the effect of the number of nodes via partial correlation, most of the previously significant correlation coefficients become negligible. In the previous table, V_{CH} looses its correlation with V_V, while keeping correlation with $\overline{ecc}$, V_{ecc} and with $\overline{btw}$; this latter is the most significant, albeit strongly reduced with respect to the total correlation (from 0.98 to 0.76). Moreover, V_V loosed all correlations; this is not surprising, since V_V simply accounts for the sum of the volumes of the atoms, so removing the dependence on n unties all other correlations. In other words, V_V strictly describes protein encumbrance, in turn depending on protein size.

On the other hand, V_{CH} describes the envelope of the protein molecule, which depends less markedly on the number of atoms. Such a metric describes better protein compactness, so it is able to catch to some extent the functional aspects of the protein volume. Such functional aspects are also accounted by the betweenness centrality, irrespectively of the protein size.

As for $\overline{ecc}$, V_{ecc} and $\overline{sp}$, they remain strongly correlated with each other and with the average degree, while they loose correlation with the betweenness. Moreover, their correlation with V_{CH} is above 0.5, suggesting a strong relationship with the convex hull of the protein, regardless of the network size.

It is interesting to note that, while the betweenness strongly correlates with V_{CH}, it shows a poor correlation with eccentricity-related indicators; this suggests that, although both metrics are are good volume descriptors, and, in particular, they are particularly well descriptive of the bulk/envelop of the folded proteins, they are somehow complementary.

4 Conclusions

Protein structure depends on size, and so do most topological descriptors linked to shortest paths (from eccentricities to centralities). When the effect of nodes is removed, most correlations fall down, and only size-independent links remain, such as that between the convex hull volume and the betweenness centralities. This suggests the existence of a link between volume metrics - accounting for protein compactness - and betweenness, which is also primarily independent on degree (so more compact proteins are not said to be endowed with a high connectivity). This work highlights again the modular architecture of protein structures, which is also related to the signal transmission throughout the protein molecule.

Finally, the definition of a new metrics of network volume V_{ecc} has a potential impact both in the field of PCNs analysis and protein structure prediction, since we demonstrated this descriptor well complies with structural features related to protein functionality (such as protein compactness).

References

1. N. Plattner, F. Noé, Nature communications **6** (2015)
2. Z. Bu, D. Callaway, Adv Protein Chem Struct Biol **83**, 163 (2011)
3. M.B. Enright, D.M. Leitner, Physical Review E **71**(1) (2005)
4. X.S. Zhang, Z.W. Zhan, Y. Wang, L.Y. Wu, Operations Research and Its Applications, Lecture Notes in Operations Research **5**, 276 (2005)
5. Y. Wang, L.Y. Wu, X.S. Zhang, L. Chen, in *Theory and Applications of Models of Computation* (Springer, 2006), pp. 505–514
6. M. Gerstein, J. Tsai, M. Levitt, Journal of molecular biology **249**(5), 955 (1995)
7. Y. Harpaz, M. Gerstein, C. Chothia, Structure **2**(7), 641 (1994)
8. L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, A. Giuliani, Chem Rev **113**(3), 1598 (2013)

9. L. Di Paola, A. Giuliani, Adv Sys Biol **3**(1), 7 (2014)
10. L. Di Paola, A. Giuliani, Curr Opin Struct Biol **31**, 43 (2015). DOI 10.1016/j.sbi.2015.03.001
11. A. Giuliani, L. Di Paola, Curr Protein Pept Sci **17**(1), 3 (2016)
12. R.K. Grewal, S. Roy, Protein Pept Lett **22**(10), 923 (2015)
13. N. Arrigo, The Open Bioinformatics Journal **6**(1), 20 (2012)
14. A.R. Atilgan, P. Akan, C. Baysal, Biophys J **86**(1 Pt 1), 85 (2004). DOI 10.1016/S0006-3495(04)74086-2
15. A. del Sol, H. Fujihashi, D. Amoros, R. Nussinov, Mol Syst Biol **2**, 2006.0019 (2006). DOI 10.1038/msb4100063
16. G. Bagler, S. Sinha, Bioinformatics **23**(14), 1760 (2007)
17. F. Harary. Graph theory. 1969
18. P. Hage, F. Harary, Social networks **17**(1), 57 (1995)
19. N. Mitton, A. Busson, E. Fleury, in *Mediterranean ad hoc Networking Workshop (MedHoc-Net'04).* (2004), p. 0000
20. D. Peleg, L. Roditty, E. Tal, Automata, Languages, and Programming pp. 660–672 (2012)
21. M. Reed, Journal of Choice Modelling **17**, 28 (2015)
22. M. Sternberg, H. Hegyi, S.A. Islam, J. Luo, R.B. Russell, in *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 3 (1994), vol. 3, pp. 376–383
23. S.A. Islam, J. Luo, M.J. Sternberg, Protein Engineering **8**(6), 513 (1995)
24. L. Gulyas, G. Horváth, T. Cséri, Z. Szakolczy, G. Kampis, in *19th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2010)* (2010)
25. S. Tasdighian, L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, P. Palumbo, G. Mei, A. Di Venere, A. Giuliani, J Chem Inf Model **54**(1), 159 (2014)