Topological Analysis of SARS CoV-2 Main Protease

Ernesto Estrada

April 3, 2020

Institute of Applied Mathematics (IUMA), Universidad de Zaragoza, Pedro Cerbuna 12, E-50009 Zaragoza, Spain; ARAID Foundation, Government of Aragón, 50018 Zaragoza, Spain.

Abstract

There is an urgent necessity of effective medication against SARS CoV-2, which is producing the COVID-19 pandemic across the world. Its main protease (M^{pro}) represents an attractive pharmacological target due to its involvement in essential viral functions. The crystal structure of free M^{pro} shows a large structural resemblance with the main protease of CoV-1. Here we report that CoV-2 M^{pro} is however 300% more sensitive than CoV-1 M^{pro} in transmitting tiny structural changes across the whole protein through long-range interactions. The largest sensitivity of M^{pro} to structural perturbations is located exactly around the catalytic site Cyst-145, and coincides with the binding site of several of its inhibitors. These findings, based on a simplified representation of the protein as a residue network, may help in designing potent inhibitors of CoV-2 M^{pro}.

1 Introduction

Since December 2019 an outbreak of pulmonary disease has been expanding from the city of Wuhan, Hubei province of China [1, 2]. This disease–produced by a new coronavirus named SARS-CoV-2 [3]–has become pandemic in about three months, affecting more than 200 countries around the world. SARS-CoV-2 belongs to the genus Betacoronavirus [4, 5], to which the virus which produced the respiratory epidemic of 2003 (SARS-CoV-1) also belongs to. The new coronavirus shares about 82% of its genome with CoV-1. In spite of this similarity and of the fact that SARS-CoV-1 appeared almost 20 years ago, there are currently no approved specific drugs against SARS-CoV-2 [6, 7, 8, 9]. In consequence, most of the clinical treatment used against the disease is symptomatic in combination with some repurposed drugs, such as the antiviral Remdesivir or the antimalarials chloquine [10] and hidroxychloroquine [11]. This situation

urges the scientific community to search for specific antiviral therapeutics and vaccines against SARS-CoV-2.

Based on previous success in finding antiviral drugs against human immunodeficiency virus (HIV) [12] and SARS CoV-1 [13], the viral protease of SARS-CoV-2 represents an attractive pharmacological target. The main protease (M^{pro}) of CoV-2 is a key enzyme for the virus because it is essential for proteolytic processing of polyproteins [14]. Thus, inhibiting the activity of M^{pro} would result in blocking viral replication. This protein does not exist in humans, which makes it attractive for avoiding possible toxicities of anti-CoV-2 drugs. The three-dimensional structure of CoV-2 M^{pro} has been resolved by Zhang et al. at $1.75\mathring{A}$ of resolution [15]. They also reported the structure of CoV-2 M^{pro} with an α-ketoamide inhibitor in two different space groups. Other structures of CoV-2 M^{pro} complexed with inhibitors have been solved by Masecar [16] and by Jin et al [17].

There are some remarkable characteristics of CoV-2 M^{pro} in relation to the protease of CoV-1. They share 96% of amino acids sequence, i.e., they differ in the amino acids at only 12 out of 303 positions in the sequence. More remarkable the root mean square (r.m.s.) deviation between the two three-dimensional structures is only 0.53 Å for all C_α positions. It seems like the two proteases are almost identical in their three-dimensional structures. The first question that emerges here is whether such similarities are also reflected at the topological structural level of the proteins. By topological we mean here the discrete topology emerging from a network theoretic representation of a protein. In this representation of the protein structure the nodes of the network represent amino acids and the edges connecting them indicate that the corresponding residues are at a distance in which they can interact to each other. Because the Euclidean distance between the amino acids is used to construct the network we more correctly should refer to this framework as topographical more than topological. This network theoretic representation has been previously used to answer several questions related to protein structure and functioning [18, 19, 20, 21, 22]. Among the tools in use, the one of node centrality [23, 24] has played a fundamental role (see for instance [22]. These indices capture the relative importance–both structural and dynamical–of an individual amino acid in the protein.

Here we construct protein residue networks (PRN) for CoV-2 M^{pro} and some of its inhibitors. The PRN of CoV-2 M^{pro} is illustrated in Fig. 1. We then analyze the similarities in the topological structure of CoV-2 M^{pro} with that of CoV-1 for which we also construct the corresponding PRN. We then show that both proteases are very similar in relation to a few topological characteristics which account for a very close environment around the amino acids. That is, when the descriptors used account for the locality of the topological environment of a residue the two proteases do not differ in more than 4%. However, when the descriptors considered account for wider environments around the nodes the difference between the two proteins can increase up to 10%. These descriptors quantify how a perturbation at an amino acid is transmitted through the whole structure to the rest of the residues in the protein. When this transmission is

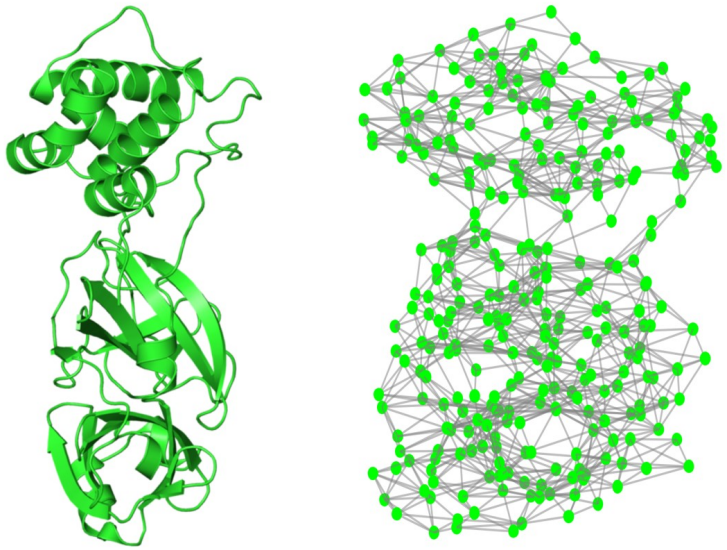


Figure 1: Cartoon representation (left) of the M^{pro} of CoV-2 (PDB=6Y2E) and the corresponding protein residue network (right).

allowed not only between close pairs of amino acids but also between very distant ones, the difference between the two proteases increases up to 300%. That is, CoV-2 M^{pro} is 300% more sensitive to the transmission of perturbations between amino acids through the topological structure of the protein than CoV-1 M^{pro}. We discovered that the residues with this largest sensitivity in CoV-2 M^{pro} are the ones involved in the binding of the three inhibitors studied here. That is, the most central amino acids according to this long-range indices are also the most affected by the interaction with the inhibitors as they are either in the binding site or very close to it. Consequently, we have discovered that the most relevant amino acids from the topological point of view are also the most relevant ones for the binding of some inhibitors to the CoV-2 M^{pro} and should play an important role in the design of drugs inhibiting this protease.

2 Results

2.1 Free protease

The main goal of this section is to analyze a few network theoretic descriptors of the M^{pro} of SARS CoV-2 and compare them with those of the protease of SARS CoV-1. The amino acid sequence of both proteases share 96% of similarity, i.e., only 12 amino acids are different in both proteases of a total of 303. These amino acids are at positions 33, 44, 63, 84 86, 92, 132, 178, 200, 265, 283 and 284. It

descriptor	CoV-1	CoV-2	$\Delta_{rel}(\%)$
δ	0.0260	0.0261	0.38
ρ	0.0151	0.0164	**8.61**
$\langle L \rangle$	6.2722	6.271	-0.02
$\langle C \rangle$	0.539	0.530	-1.59
r	0.398	0.385	-3.36
$\langle BC \rangle$	782.92	792.71	1.26
$\langle EC \rangle$	0.0326	0.0317	-2.76
$\langle SC \rangle$	167.71	177.93	**6.09**
$\langle G_{pq} \rangle$	21.212	23.405	**10.34**
$\langle \theta \rangle$	82.47	82.13	-0.41
$\langle Z_{pp} \rangle$	$1.955 \cdot 10^{17}$	$7.922 \cdot 10^{17}$	**305.44**
$\langle Z_{pq} \rangle$	$6.159 \cdot 10^{16}$	$2.355 \cdot 10^{17}$	**282.39**

Table 1: Global topological properties of the M^{pro} of CoV-1 and CoV-2 as well as the relative difference between them, expressed as percentages of change relative to CoV-1.

has been reported by Zhang et al. that the three-dimensional structures of both proteases are very similar with an r.m.s. deviation of 0.53 Å for all C_α positions using the free enzymes (PDB=6Y2E for CoV-2 and PDB=2BX4 for CoV-1 [25]). Then, it is not surprising that most of the topological characteristics of the first kind of the PRNs of both proteases are very similar with relative differences not bigger than 4% for all the properties analyzed except for the degree heterogeneity. In this case, the relative difference is 8.61% between both proteases, showing that that of CoV-2 is more heterogeneous than the one of CoV-1 (see Table 1). However, the values of this parameter are very close to zero in both cases (the degree heterogeneity index is bounded between zero and one), which indicate that both networks are poorly heterogeneous and look more like regular graphs. In fact, both networks have almost normal degree distributions and the main difference is that for CoV-1 about 27% of the nodes have degree 5, while for CoV-2 about 32% have degree 4.

We then continue the analysis by comparing the topological descriptors of the second kind. We notice that neither the eigenvector centrality, which has been found very useful in previous analysis of PRN [22], nor the communicability angles, which account for the communication efficiency of the PRN, display any significant difference between both proteases. However, there are differences in the mean subgraph centrality of about 6% and of the average communicability between pairs of nodes of more than 10%. In both cases, the indices are significantly larger for the protease of CoV-2 than for that of CoV-1. This means that the 12 punctual mutations that make the difference between the proteases of CoV-1 and CoV-2 increase the capacity of the individual amino acids of feeling a perturbation or thermal oscillation produced in another amino acid of the protein. As we have previously explained these communicability

factors penalizes very heavily any perturbation being transmitted between two amino acids separated by a relatively long distance in the protein. Thus, they can be considered as a indices that account for shorter range interactions than the third kind descriptors considered here.

Both LR subgraph centrality and communicability display dramatic increment in CoV-2 relative to CoV-1. In this case the increase of these indices is more than 280% for the LR communicability and more than 300% for the LR subgraph centrality. In short, this means that the protease of CoV-2 has almost 4 times more capacity of transmitting perturbations between pairs of nodes than the protease of CoV-1. This is equivalent to say that the protease of CoV-2 is significantly much more topologically efficient in transmitting "information" among its amino acids than the protease of CoV-1. And, such efficiency has been obtained only with 12 punctual mutations!

We now proceed to the analysis of the local variation of the subgraph and the LR subgraph centralities for the amino acids of the two M^{pro} (see Fig. 2). In the case of the subgraph centrality the largest change is produced for a few amino acids which increase their centrality in CoV-2 relative to CoV-1. These are the cases of 27, 205, 202, 295, 290, and143. But there are also other amino acids which drop their centrality in CoV-2, such as 172, 225, 234, 171, and 235among others (see Fig. 2(b)). Therefore, the increase of the subgraph centrality of a few amino acids makes that in total the average subgraph centrality increases in CoV-2 in relation to CoV-1. An important characteristic feature of the differences in this centrality between the two proteases is that they are spread across the three domains of the proteases with a large increment in the domains I and III. This is a major difference with the LR subgraph centrality (see Figs. 2(c) and (d)), where the main change is a dramatic increase in the centrality of the nodes in the domains I and II of the CoV-2 protease relative to CoV-1. The changes occurring in the domain III are imperceptible in relation to those of the other two domains.

The distributions of the most central amino acids according to both measures in the three-dimensional structures of the proteases is illustrated in Fig. 3. It can be seen that the largest values of the LR subgraph centrality are concentrated in a relatively small region of the protein structure, while those of the subgraph centrality are more spread across the whole structure. We then inquire about this region of the M^{pro} in CoV-2 which shows the largest change in the LR subgraph centrality relative to its analogue of CoV-1.

The first remarkable observation of the amino acids with the largest change in the LR subgraph centrality is that they are all closely located in the three-dimensional space. For instance, the 22 amino acids displaying the largest change in this centrality form a connected subgraph of the PRN as illustrated in Fig. 4. This subgraph of 22 nodes has 48 connections among these amino acids, which produces an edge density of 0.21, almost 10 times bigger than the total density of the protease. The second remarkable feature of this subgraph is that it contains one of the two catalytic amino acids of the M^{pro} of CoV-2, which is Cys-145. That is, the region with the largest increase in the LR subgraph centrality of the protease of CoV-2 relative to CoV-1 is the one enclosing the

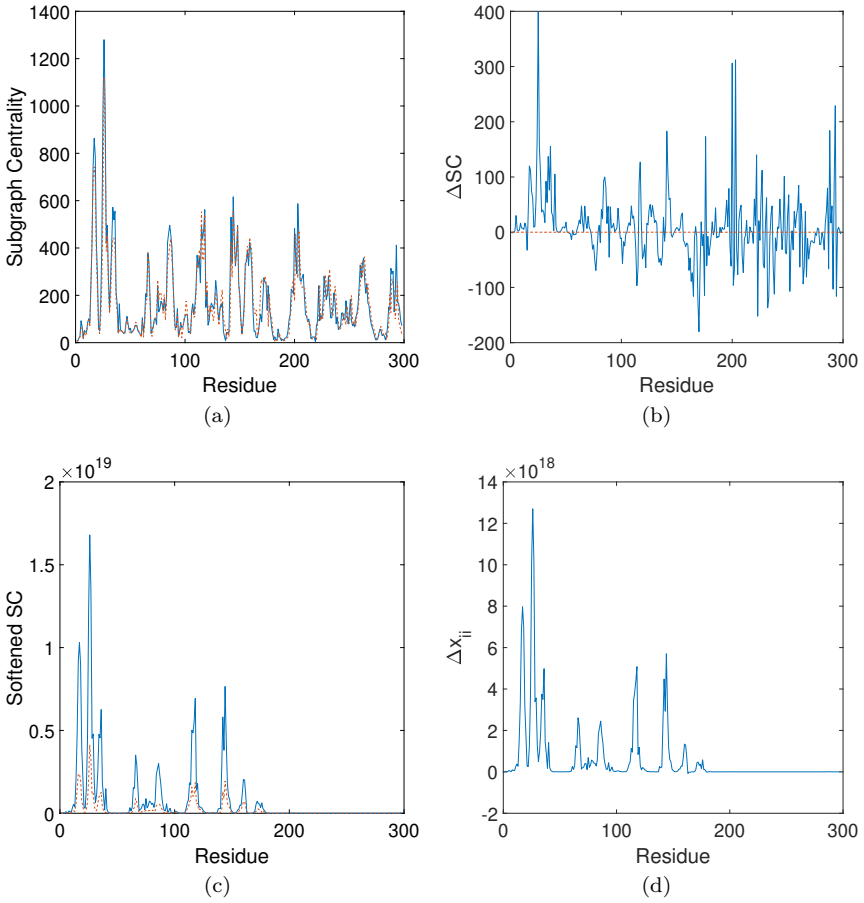


Figure 2: Plot of topological properties of the amino acid residues for the M^{pro} of CoV-1 (broken red line) and of CoV-2 (solid blue lines). (a) subgraph centrality; (b) LR subgraph centrality.

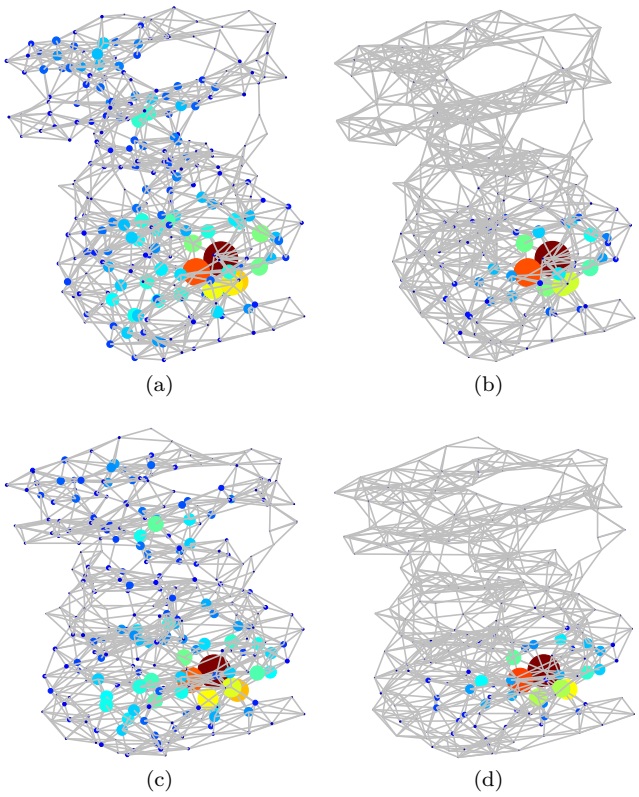


Figure 3: Illustration of the subgraph (a), (c) and LR subgraph (b), (d) centralities of the amino acid residues of the chain A of CoV-1 M^{pro} of (top), and of CoV-2 (bottom).

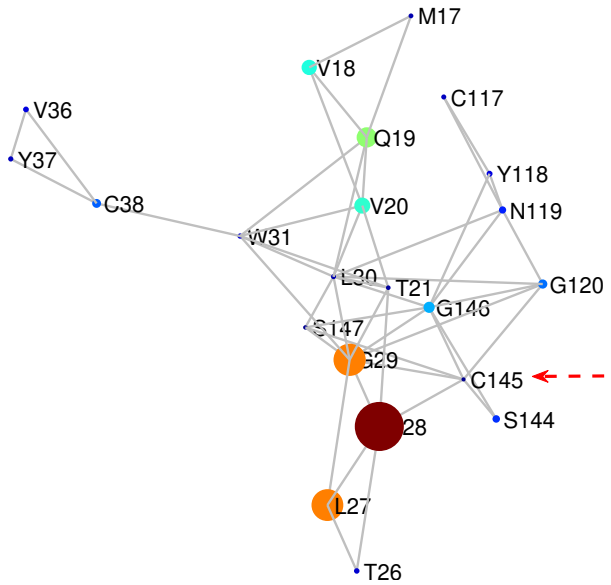


Figure 4: Illustration of the 22 amino acids which display the largest difference in the LR subgraph centrality in the M^{pro} of CoV-2 in relation to that of CoV-1. The radius of the nodes is proportional to the difference in the LR subgraph centrality between the two proteases. The catalytic site Cys-145 is pointed to with an arrow.

catalytic binding site of amino acid Cys-145. It is also remarkable that this region of large increment in the LR subgraph centrality contains some amino acids which are located in the binding site of the M^{pro} to α-ketoamide inhibitors as well as other kind of inhibitors, as we will analyze further in this work. This is the case of the residue 144-147, other amino acids in this binding site like residues 162, 163 also display large increment in the LR subgraph centrality. The last remarkable observation is that the domain III displays small change in relation to the changes of domains I and II in this topological parameter. However, as we will see in the next paragraphs this domain (residues 198-303) which is formed by 5 helices and is involved in the dimerization of the M^{pro}, also increases significantly the LR communicability in relation to CoV-1.

A better picture of the changes in the different regions of the M^{pro} of CoV-2 relative to CoV-1 are obtained by analyzing the differences between the communicabilities and LR-communicabilities between every pair of amino acids in CoV-2 in relation to CoV-1. In Fig. **??** we illustrate the difference matrices for both kinds of communicabilities. In the first case it can be observed that the communicability between all pairs of residues in the domain I (residues 10-99) mainly increase in CoV-2 relative to CoV-1, with an increase of 20.7% relative to CoV-1. However, in the domain II (residues 100-182) there is mainly a drop of the communicability between the residues in the domain, which decrease 6.6%,

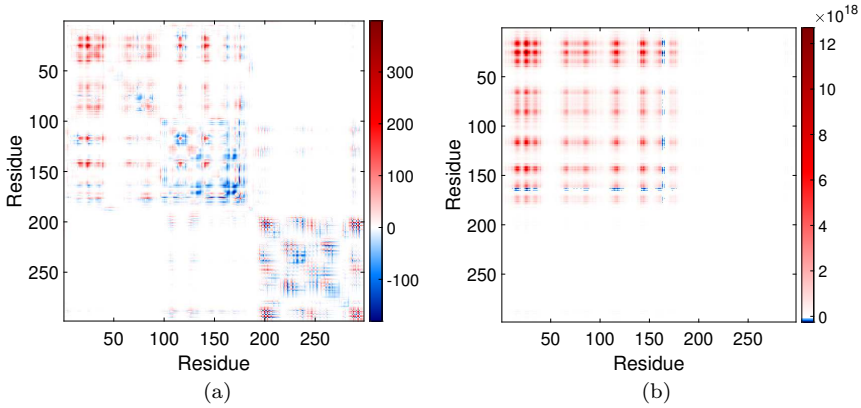


Figure 5: Difference between communicabilities (a),and LR communicability (b) between pairs of amino acids in the M^{pro} of CoV-2 in relation to that of CoV-1.

but there is an increase of 19.2% in the trade off between domains I and II. The domain III shows a mixed behavior with some pairs of residues increasing and other decreasing their communicability, but the main result is an increase of 3.28% relative to CoV-1. Although the absolute values of the communicability between this domain and the other two are very small, there are significant increments in relation to that in CoV-1. For instance, the communicability between domains I and III increases in 17.3% and that between domains II and III in 10.5% relative to the same in CoV-1.

We now move to the analysis of the LR communicability between the different domains of the CoV-2 protease. Here the changes are dramatic and in all cases the communicability in the CoV-2 protease is higher than that in CoV-1. For instance, the average communicability between pairs of nodes in the domain I is 360.9% higher in CoV-2 than in CoV-1. This percentage of increment are 196.1% in the domain II and 125.7% in domain III. As we have mentioned before, although the increase in the communicability in the domain III is is one third of that in domain I and one half of the one in domain II, it is still significant when compared to that in the CoV-1 protease. The inter-domain communicability also increases very significantly with increment of 269.6% (domains I-II), 239.4% (domains (I-III) and 169.8% (domains II-III). In closing, the mutations in 12 amino acids of the M^{pro} of CoV-1 has produced a dramatic impact in the LR communicability between residues in the protease of CoV-2 with huge improvement in long-range communication between residues mainly in domain I and II and between domains I and II as well as domains I and III.

2.2 CoV-2 Protease bounded to an inhibitor

We turn now our attention to the analysis of the M^{pro} of CoV-2 complexed with an inhibitor for which the crystal structure was resolved by Zhang et al. in two

descriptor	6Y2F	6Y2G
δ	-1.85	-2.08
ρ	1.85	**10.85**
$\langle L \rangle$	0.68	2.04
$\langle C \rangle$	0.32	1.79
r	5.90	-2.45
$\langle BC \rangle$	2.50	2.42
$\langle EC \rangle$	-1.65	0.00
$\langle SC \rangle$	2.33	**-11.70**
$\langle G_{pq} \rangle$	2.15	**-13.85**
$\langle \theta \rangle$	0.06	0.40
$\langle Z_{ii} \rangle$	**16.39**	**-83.85**
$\langle Z_{ij} \rangle$	**18.57**	**-82.73**

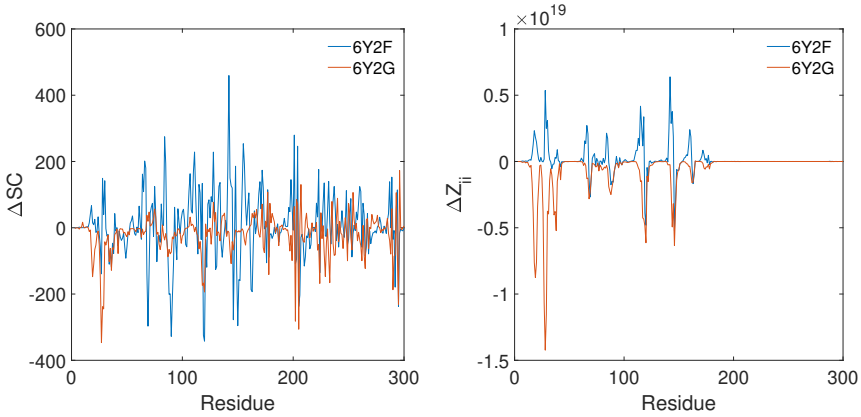
Table 2: Relative differences in percentage of global topological properties of the M^{pro} of CoV-2 complexed to an inhibitor in relation to free one. The PDB of the complexes between the M^{pro} of CoV-2 with an inhibitor correspond to 6Y2F (space group $C2$), and 6Y2G (space group $P2_12_12_1$).

different space groups, space group, PDB C26Y2F and space group $P2_12_12_1$, PDB 6Y2G [15].

As for the case of the free protease we start the analysis by considering a few topological descriptors of the whole PRNs for the proteins with PBD codes 6Y2F and 6Y2G [15]. In this case we consider the relative difference of the topological descriptors for these proteins minus that of the free protease, which has the PDB code 6Y2E. In Table 2 we resume these results. It can be seen that here again the topological descriptors of the first class display relatively little variation for the two complexed proteases relative to the free one. The only exception is again the degree heterogeneity for the case of 6Y2G which shows a relative variation of almost 11%.

Although the variation of the degree heterogeneity for 6Y2G is significant relative to the free protease, the value of this parameter is still very close to zero for all proteins considered here. For instance, it is 0.0172 for 6Y2F and 0.0182 for 6Y2G, which are in the range of that of the free protease which is 0.0164. These values of the heterogeneity close to zero correspond to networks which have scarce degree heterogeneity like almost regular graphs or networks with normal-like degree distributions.

We then move to the analysis of the descriptors of second and third type. As can be seen in Table 2 there are significant changes in the subgraph centrality and the communicability in 6Y2G but not in 6Y2F. In the case of 6Y2G the main change corresponds to a drop in both parameters after the binding with the inhibitor. The plot of the differences in the subgraph centrality for every node of the protease bounded to the inhibitor in both space groups reveals the most important characteristic of the change in this topological parameter. As can be



(a) (b)

Figure 6: Change of the subgraph centrality (a) and its LR version (b) for the protease of CoV-2 bounded to an inhibitor with two different space groups.

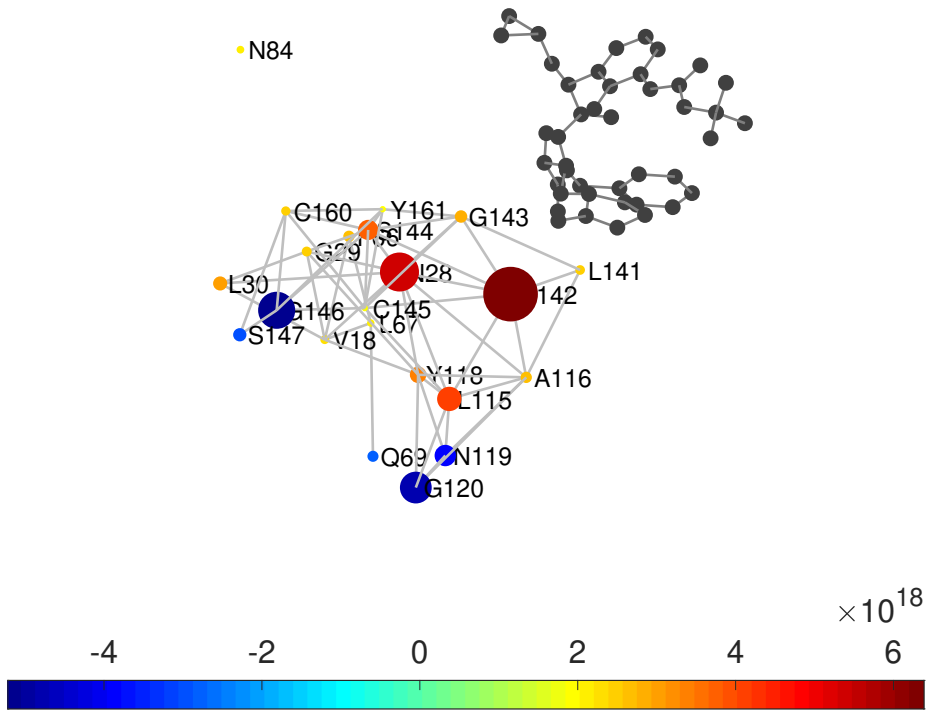
seen in Fig. 6 (a) the change in the subgraph centrality in both structures (6Y2F and 6Y2G) is spread across the protein chain, indicating a global response of the protease to the binding of the inhibitor. The situation is contrasting with that observed for the LR subgraph centrality and communicability (third class of descriptors). In these cases the change relative to the unbounded protease are significant for both 6Y2F and 6Y2G as can be seen in Table 2. More importantly, the change of the LR subgraph centrality is located at very specific regions of the protease as can be seen in Fig. 6 (b). In this case, it is clear that although 6Y2F displays mainly increase of the LR subgraph centrality, and 6Y2G display decrease of this parameter, in both cases the changes are located at approximately the same positions of the protease chain.

With the goal of disentangling the information contained in the changes produced at the LR subgraph centrality of the bounded protease we study it in more detail here. For this, we consider the amino acids displaying the largest change in this topological parameter for both PDB structures. In the case of 6Y2F the 22 amino acids displaying the largest variability in the LR subgraph centrality forms a connected subgraph with the exception of N84, which is isolated. This subgraph has 54 edges, which means that it is significantly more dense than the whole protein, i.e., its density is 0.23 vs. 0.026. As can be seen in Fig. 7 (a) there are a few residues in this subgraph which increase the LR subgraph centrality, such as N142, N28, L115, S144, while others like G146, G120, N119 decrease it. In Fig. 7 (a) we have included the inhibitor to show that this group of amino acids displaying the largest variation in the LR subgraph centrality are just the ones closely interacting with the inhibitor. Indeed, amino acids F140, G143, S144, and C145 are in the binding site of the protease with this inhibitor. These amino acids are also among the ones that display the largest variation in

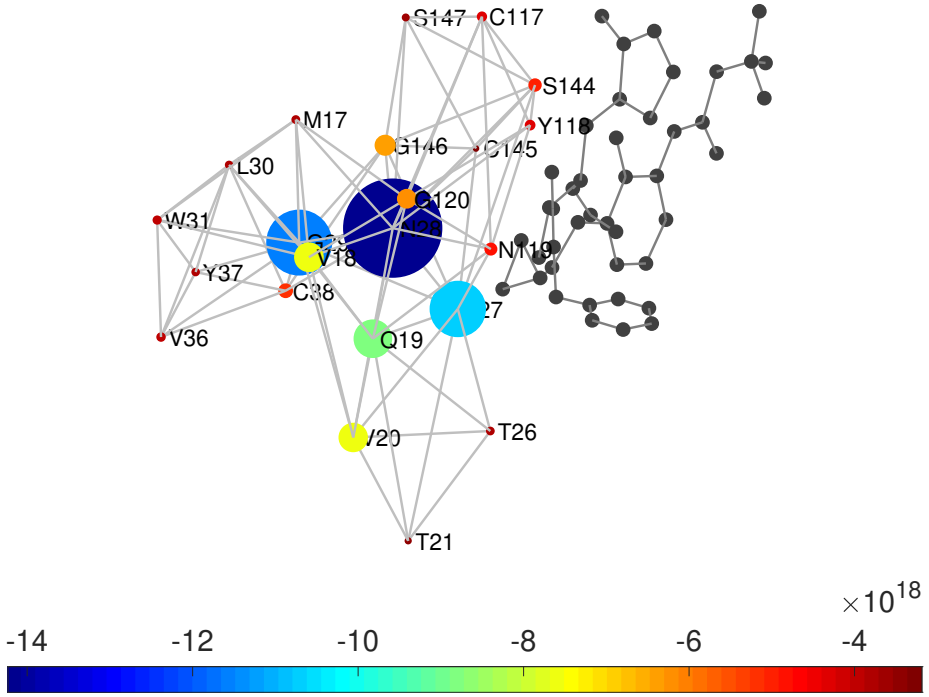
this parameter for the free protease of CoV-2 relative to the one of CoV-1. In fact, 11 out of 22 residues with the largest change in this parameter after the binding of the protease with the inhibitor coincide with those having the largest variation in the free protease relative to CoV-1.

We now consider the amino acids with the largest variation of the LR subgraph centrality in 6Y2G. In this case the 22 residues with the largest variation in this parameter form a connected subgraph having 76 edges, which produces an edge density of 0.33, which is very high in relation to that of the whole protein, i.e., 0.026. In Fig. 7 (b) we illustrate this subgraph together with the inhibitor at its original position in the protein, which clearly indicates that the region with the largest variation of the LR subgraph centrality is the one involved with the binding of the inhibitor. In this case all the amino acids display a significant drop of this topological parameter, indicating that the binding of the inhibitor drops significantly the capacity of this region of the protein, and of the whole protein, to transmit perturbations in a long-range basis. The most remarkable observation of this subgraph is the fact that 100% of the amino acids involved in it are the ones that previously displayed the largest increase in this parameter in the protease of CoV-2 relative to CoV-1. Indeed, they are also the residues with the largest absolute value of the LR subgraph centrality in the free protease. This means that the LR subgraph centrality is a topological parameter which is clearly related with the most important region of the CoV-2 protease for the design of molecules which inhibit this protein.

To complete our analysis we investigate other structures of the CoV-2 M^{pro} bounded to two different inhibitors. They corresponds to the structure with PDB code 6W63 determined by Masecar at $2.1\mathring{A}$ of resolution [16] and the one corresponding to the PDB code 6LU7 determined by Jin et al. at $2.16\mathring{A}$ of resolution [17]. Here we focus only on the difference between the LR subgraph centrality at the amino acids of these proteases relative to that of the free one, structure 6Y2E. In Fig. 8 we illustrate the subgraphs having the 22 most central amino acids according to this topological feature. As can be seen both subgraphs are connected and have edge densities 0.33 (6W63) and 0.31 (6LU7). In the case of 6W63 21 out of 22 amino acids are the same found as the most central ones in 6Y2E according to this index. For 6LU7 16 out of 22 of the amino acids with the largest deviations of the LR subgraph centrality coincide with the most central ones in the free CoV-2 M^{pro}. In both cases, these residues are in the binding site of CoV-2 M^{pro} or very close to it, indicating that the functional importance of them coincides with their topological relevance in the PRN. An interesting difference between the responses of the CoV-2 M^{pro} to the inhibitors in 6W63 and 6LU7 is that in the first the centrality of the nodes after binding is smaller than in the free protease. Although the free protease 6Y2E and the one bounded to inhibitor 6W63 are solved at different resolutions we point out that this behavior is similar to the one observed for 6Y2G. However, in the case of 6LU7 the response of the amino acids after binding with inhibitor is an increment of their centrality, which is equivalent to a more compactness of this region of the protein and of the whole protein itself. It is possible that these differences are due to the main structural variations between the inhibitor



(a)



(b)

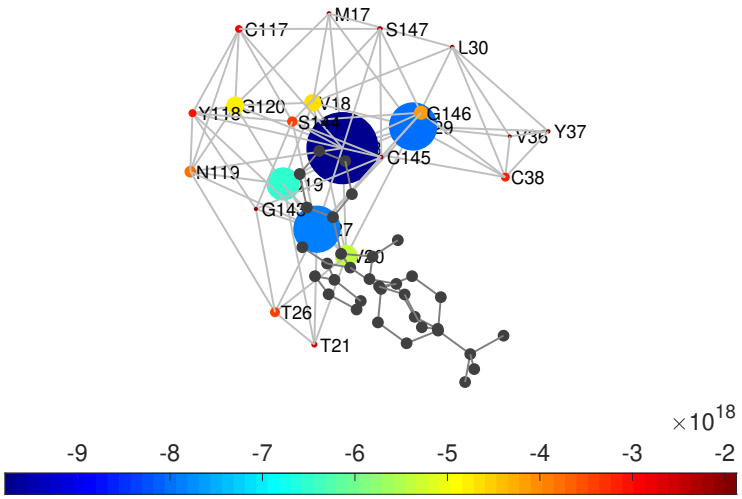
Figure 7: Illustration of the 22 amino acids with the largest variation of the LR subgraph centrality in 6Y2F (a) and 6Y2G (b) relative to 6Y2E. The residues are connected if they are at no more than 7$\AA$. The color bar corresponds indicates the values of ΔZ_{ii} and the radius of the nodes is proportional to the absolute value of this difference.

in 6LU7 and the other two.

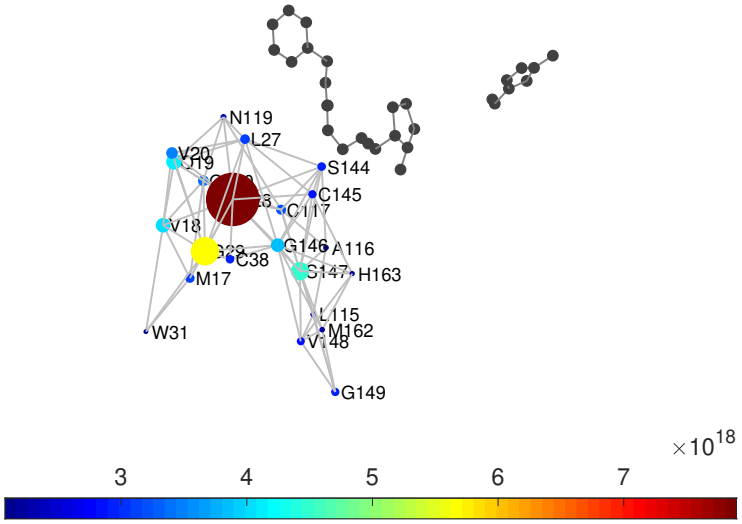
3 Discussion

We present an analysis of some of the most relevant topological properties of the main protease of the SARS CoV-2. Our approach is based on the representation of the three-dimensional structure of the protein as a residue network in which C_α of every amino acid is represented by a node of the network and two nodes are connected if the corresponding C_α are at no more than 7.0 Å. CoV-2 M^{pro} differs only in 12 punctual mutations from CoV-1 M^{pro} and the superposition of their three-dimensional skeletons gives an r. m. s. of only 0.53 Å. Then, it comes at no surprise that the difference between most of the topological properties of the PRNs representing both proteases differ in less than 5%. If we exclude from the analysis the LR descriptors, then 70% of the topological descriptors shows only a small variation between the two proteases. In this situation it is certainly remarkable that there are topological descriptors which change in about 300% from one protease to the other. These are the cases of the LR subgraph centrality of the amino acids and of the LR communicability between pairs of them. The increase of these parameters in about 300% for CoV-2 M^{pro} relative to CoV-1 M^{pro} means that the 12 punctual mutations that differentiate both proteases have created a huge increment in the efficiency of CoV-2 M^{pro} in transmitting perturbations of any kind between the amino acids of the protein using all available routes of connection and allowing for long-distance transmission. To make clearer what this sensitivity means we are going to use a simple example. Let us consider a tiny perturbation on the structure of the proteases which prevent the interaction between the amino acids P9 and G11, which have been selected at random. In CoV-1 M^{pro} these amino acids are at 5.69 Å and in CoV-2 M^{pro} they are 6.48 Å apart. Thus, in both cases they are connected in the corresponding PRN. Let us consider that the perturbation remove this edge from the PRN of both proteases. The relative decrement of the average path length in CoV-2 M^{pro} relative to CoV-1 M^{pro} is almost imperceptible, i.e., 5.7%. In the case of the subgraph centrality it is of the same order, i.e., 3.4%. This means that according to these parameters CoV-2 M^{pro} is as sensitive as CoV-1 M^{pro} to perceive a structural change in its structure produced by a given perturbation. However, when we consider the LR subgraph centrality this relative change is 316.8%. That is, according to this topological parameter which takes into account long-range interactions, CoV-2 M^{pro} is more than three times more sensitive to a tiny structural change than CoV-1 M^{pro}. This remarkable finding indicates that the 12 mutations produced in CoV-1 M^{pro} makes the resulting CoV-2 M^{pro} much more efficient in transmitting "information" through the protein skeleton using short and long-range routes.

The second remarkable finding of the current work is that the largest changes in the LR subgraph centrality occurring in CoV-2 M^{pro} relative to CoV-1 M^{pro} do not spread equally across the whole structure of the protease. Instead, they



(a)



(b)

Figure 8: Illustration of the 22 amino acids with the largest variation of the LR subgraph centrality in 6W63 (a) and 6LU7 (b) relative to 6Y2E. The residues are connected if they are at no more than 7$\mathring{A}$. The color bar indicates the values of ΔZ_{ii} and the radius of the nodes is proportional to the absolute value of this difference.

are concentrated around a geometrical region which includes most of the amino acids involved in the binding site of the protease to inhibitors or close to it. One of the amino acids which has increased more dramatically its sensitivity to long-range transmission of information in CoV-2 M^{pro} is Cys-145, which is one of the two catalytic sites of the protease, and the one involved in interactions with the inhibitors, such as the ones analyzed here. We have analyzed here three different inhibitors of CoV-2 M^{pro}. In the three cases we have observed a significant variation in the LR subgraph centrality of the amino acids which were previously observed to have increased their LR sensitivity in the free protease. Therefore, these amino acids corresponds to those involved in the binding of these three inhibitors, showing that their increased topological role in the CoV-2 M^{pro} also may play an important functional role in it.

The analysis of PRN is easier than the study of the whole protein structure. In this sense the PRN represents a simplified model of the three-dimensional structure of the protein. Typically, such simplification in the complexity of the representation of systems convey a loss in the structural information which is represented by the global system. In this case, however, we have shown that the use of a network representation of the proteins reveals some hidden patterns in their structure that were escaping to the analysis by using the global structure. To detect such important structural factors it is necessary to account for long-range interactions among the amino acids of the proteases, which are the ones revealing the their most important characteristics in terms of their sensitivity to tiny structural changes produced by local or global perturbations to the system. Such LR interactions revealed here the main differences between the proteases of CoV-1 and CoV-2, as well as the most important amino acids for the interaction with inhibitors, which may produce therapeutic candidates against COVID-19.

4 Methods

4.1 Construction of the protein residue networks

The protein residue networks (PRN) (see ref. [23] Chapter 14 for details) are built here by using the information reported on the Protein Data Bank [26] for the proteases of CoV-1 and CoV-2 as well as the complexes of the last one with an inhibitor. The nodes of the network represent the α-carbon of the amino acids. Then, we consider cutoff radius r_C, which represents an upper limit for the separation between two residues in contact. The distance r_{ij} between two residues i and j is measured by taking the distance between C_α atoms of both residues. Then, when the inter-residue distance is equal or less than r_C both residues are considered to be interacting and they are connected in the PRN. The adjacency matrix A of the PRN is then built with elements defined by

$$A_{ij} = \begin{cases} H\left(r_C - r_{ij}\right) & i \neq j, \\ 0 & i = j. \end{cases} \tag{1}$$

Here we use the typical interaction distance between two amino acids, which

is equal to 7.0 Å. We have tested distances below and over this threshold obtaining in general networks which are either too sparse or too dense, respectively.

In this work we consider the structures of the M^{pro} of CoV-1 deposited in the PDB with code 2BX4 [25], of CoV-2 with PDB code 6Y2E as well as the one of the CoV-2 with an inhibitor and space groups C_2, PDB code 6Y2F and space group $P2_12_12_1$ with code 6Y2G [15]. In addition we also consider two other structures of the CoV-2 M^{pro} with inhibitors, which corresponds to two other different studies from the ones of Zhang et al. [15]. The PDB codes of these structures are: 6W63 [16] and 6LU7 [17]. The length of the proteases is 306 amino acids. However, the structure of 2BX4 is resolved for amino acids 3 to 300, which gives a length of 298 [25]. Thus, for the sake of homogeneity of the analysis we consider here the same part of the amino acids sequence for all the structures analyzed, i.e., from residue 3 to residue 300. This does not alter the analysis as the two extremes of the protease are disordered as do not participate in important interactions.

4.2 Descriptors

The first category of descriptors correspond to those related to the most local structure around the nodes, such as those based on the degree of the nodes, i.e., the number of connections that a node has (see [23] for details). The degree accounts for the immediate effect of a node to its closest neighborhood. Among these descriptors we use here the edge density, which is defined as $\delta = \dfrac{2m}{n(n-1)}$, where m is the number of edges and n is the number of nodes. Because the average degree $\langle k \rangle = \dfrac{2m}{n}$, the relation with the edge density is clear. Another descriptor related to the degree of the nodes is the degree heterogeneity, $\rho = \sum_{(i,j)\in E} \left(k_i^{-1/2} - k_j^{-1/2}\right)^2$ [27], which represents a measure of how heterogeneous the degrees of the nodes is [28]. A regular network, i.e., a network with all nodes of the same degree, will have $\rho = 0$, it is followed by networks with normal-like degree distributions, then networks with more heterogeneous ones, and will end up with networks with in which the probability $P(k)$ of finding a node of degree k decays like distribution of the form $P(k) \sim k^{-1}$, where $\rho = 1$. The average Watts-Strogatz clustering coefficient [29] is defined as $\langle C \rangle = \dfrac{1}{n}\sum_{i=1}^{n} \dfrac{2t_i}{k_i(k_i-1)}$, where t_i is the number of triangles incident to the vertex. It account for the cliquishness around a node in terms of triangles, that is it account for how crowded the immediate neighborhood of a node is. Another descriptor related to the degree is the degree assortativity coefficient [30], which is Pearson correlation coefficient of the degree-degree correlation. $r > 0$ (degree assortativity) indicates a tendency of high degree nodes to connect to other high degree ones. $r < 0$ (degree disassortativity) indicates the tendency of high degree nodes to be connected to low degree ones. Other descriptors in this class assume that "information" is transmitted in the network through the topological shortest paths. The length of the shortest path is a distance $d(i,j)$

between the corresponding pairs of nodes i and j, and it is known as the shortest path distance. The average path length $\langle L \rangle = \frac{1}{n(n-1)} \sum_{i<j} d(i,j)$ is typically used as a measure of the 'small-worldness' of the network [29]. We also consider the average betweenness centrality [31] $\langle BC \rangle = \frac{1}{n} \sum_{i \neq k \neq j} \frac{\rho_{ikj}}{\rho_{ij}}$,where ρ_{ikj} is the number of shortest paths between the nodes i and j that cross the node k, and ρ_{ij} is the total number of shortest paths that go from i to j. It accounts for the importance of a node in passing information through it to connect other pairs of nodes via shortest path only.

The second category of descriptors is formed by those that account for the transmission of information not only via the shortest paths but by using any available route that connect the corresponding pair of nodes. These descriptors use the concept of walk instead of that of a path. A walk of length k in G is a set of nodes $i_1, i_2, \ldots, i_k, i_{k+1}$ such that for all $1 \leq l \leq k$, $(i_l, i_{l+1}) \in E$. A *closed walk* is a walk for which $i_1 = i_{k+1}$. The number of walks of length k between the nodes i and j in a network is given by $\left(A^k\right)_{ij}$. The first of these descriptors considered here is the eigenvector centrality EC [32], which is the corresponding entry of the eigenvector associated with the largest eigenvalue of A. The relation of this index with walks is given by the following. Let $N_k(i)$ be the number of walks of length k starting at node i and ending elsewhere. Then, if the network is not bipartite, which is the case of the current work, $EC_i = \lim_{k \to \infty} N_k(i) / \sum_{j=1}^{n} N_k(j)$ (see Chapter 5 in [23]). That is, the eigenvector centrality of a node is the ratio of the number of walks of infinite length that start at this node to the whole number of such walks starting elsewhere. Consequently, the average eigenvector centrality $\langle EC \rangle$, accounts for the spread of information from the nodes beyond the nearest neighbors and using any infinite-length walk in the graph. A type of descriptors of the second kind are based on counting all walks of any length, but giving more weight to the shorter than to the longer ones. These descriptors are based on the following matrix function: $G \coloneqq \sum_{k=0}^{\infty} \frac{A^k}{k!} = \exp(A)$, where $\exp(A)$ is the exponential of the matrix. then, we consider the average of the diagonal entries of this matrix, which is known as the average subgraph centrality $\langle SC \rangle = \frac{1}{n} \sum_{p=1}^{n} G_{pp}$ [33], which accounts for the participation of the corresponding node in all subgraphs of the graphs, giving more weight to the shortest than to the longer ones. Such subgraphs include for instance, edges, triangles, wedges, squares, etc. Another descriptor is the average of the non-diagonal entries of $\exp(A)$, which is known as the average communicability of the network, $\langle G_{pq} \rangle = \frac{2}{n(n-1)} \sum_{p,q} G_{pq}$ [34]. It accounts for how much a pair of nodes can communicate to each other by using all potential routes available in the network, but giving more weight to the shortest than to the longer ones. Finally, in this category we include the average communicability angle $\langle \theta \rangle = \frac{2}{n(n-1)} \sum_{p,q} \theta_{pq}$ [35], where the angle

between a pair of nodes is defined as: $\theta_{pq} = \cos^{-1}\left(\frac{G_{pq}}{\sqrt{G_{pp}G_{qq}}}\right)$. The average communicability angle describes how efficiently a network transmit information between its pairs of nodes by using all available routes.

The third category of descriptors is formed by all-walks indices that penalize less heavily longer walks connecting pairs of nodes in a network. That is, although $G = \exp(A)$ accounts for all walks connecting every pair of nodes, it penalizes very much those walks of relatively large length, then making more emphasis in shorter walks around a given node. In order to include longer walks in the analysis we study the following matrix function [36]: $Z \coloneqq \sum_{k=0}^{\infty} \frac{A^k}{k!!} = \frac{1}{2}\left[\sqrt{2\pi}\mathrm{erf}\left(\frac{A}{\sqrt{2}}\right) + 2I\right]\exp\left(\frac{A^2}{2}\right)$, which penalizes the walks of length k not by $k!$ (simple factorial) but by $k!!$ (double factorial). Then, we will consider here the average of the main diagonal $\langle Z_{ii}\rangle = \frac{1}{n}\sum_{i=1}^{n} Z_{ii}$, which accounts for the participation of the node i in all subgraphs in the graph but including bigger subgraphs than in SC. In a similar way we consider $\langle Z_{ij}\rangle = \frac{2}{n(n-1)}\sum_{i,j} Z_{ij}$, which accounts for the global capacity of the network of transmitting information between pairs of nodes and allowing longer-range transmission than in the case of the communicability. For those reasons we propose to call these indices long-range (LR) subgraph centrality and communicability.

References

[1] Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L. & Chen, H. D. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).

[2] Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y. & Yuan, M. L. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).

[3] Gorbalenya, A., Baker, S. & Baric, R. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses: The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiol.* (2020) 3(04).

[4] King, A. M. Q., Adams, M. J., Carsten, E. B. & Lefkowitz, E. J. (Eds.). Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier, 2012, pp. 806.828.

[5] Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nature Rev. Microbiol.* **17**, 181-192 (2019).

[6] Li, G. & De Clercq, E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nature Rev. Drug Discov.* **19**, (2020) 149-.

[7] Zhang, L. & Liu, Y. Potential interventions for novel coronavirus in China: A systematic review. *J. Med. Virol.* **92**, 479-490 (2020).

[8] Brüssow, H. The Novel Coronavirus–A Snapshot of Current Knowledge. *Microb. Biotech.* (2020).

[9] Cao, B., Wang, Y., Wen, D., Liu, W., Wang, J., Fan, G., Ruan, L., Song, B., Cai, Y., Wei, M. & Li, X. A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. *New England J. Med.* (2020) Mar 18.

[10] Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., Shi, Z., Hu, Z., Zhong, W. & Xiao, G. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* **30**, 269-71 (2020).

[11] Liu, J., Cao, R., Xu, M., Wang, X., Zhang, H., Hu, H., Li, Y., Hu, Z., Zhong, W. & Wang, M. Hydroxychloroquine, a less toxic derivative of chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. *Cell Discov.* **6**, 1-4 (2020) .

[12] Ghosh, A. K., Osswald, H. L. & Prato, G. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. *J. Med. Chem.* **59**, 5172-208 (2016).

[13] Yang, S., Chen, S. J., Hsu, M. F., Wu, J. D., Tseng, C. T., Liu, Y. F., Chen, H. C., Kuo, C. W., Wu, C. S., Chang, L. W. & Chen, W. C. Synthesis, Crystal Structure, Structure− Activity Relationships, and Antiviral Activity of a Potent SARS Coronavirus 3CL Protease Inhibitor. *J. Med. Chem.* **49**, 4971-4980 (2006).

[14] Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300**, 1763-1767 (2003).

[15] Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. & Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* (2020) Mar 20.

[16] Mesecar, A. D. A taxonomically-Driven approach to development of pot broad-Spectrum inhibitors of coronavirus main proteas including sars-Cov-2 (covid-19). To be published,

[17] Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, Zhang B, Li X, Zhang L, Peng C, Duan Y. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. bioRxiv. 2020 Jan 1.

[18] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* **344**, 1135-46 (2004).

[19] Estrada, E. Universality in protein residue networks. *Biophys. J.* **98**, 890-900 (2010).

[20] Karain, W. I. & Qaraeen, N. I. The adaptive nature of protein residue networks. *Proteins: Struct., Funct. Bioinf.* **85**, 917-923 (2017) .

[21] Doshi, U., Holliday, M. J., Eisenmesser, E. Z. & Hamelberg, D. Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proc. Natl. Acad. Sci. USA* **113**, 4735–4740 (2016).

[22] Negre, C. F. , Morzan, U. N., Hendrickson, H. P. , Pal, R., Lisi, G. P., Loria, J. P., Rivalta, I., Ho, J. & Batista, V. S. Eigenvector centrality for characterization of protein allosteric pathways. *Proc. Natl. Acad. Sci. USA* **115**, E12201-8 (2018).

[23] Estrada, E. The structure of complex networks: theory and applications. Oxford University Press, (2012).

[24] Latora, V. & Nicosia, V. & Russo, G. Complex networks: principles, methods and applications. Cambridge University Press, (2017).

[25] Tan, J., Verschueren, K. H., Anand, K., Shen, J., Yang, M., Xu, Y., Rao, Z., Bigalke, J., Heisen, B., Mesters, J. R. & Chen, K. pH-dependent conformational flexibility of the SARS-CoV main proteinase (Mpro) dimer: molecular dynamics simulations and multiple X-ray structure analyses. *J. Mol. Biol.* **354**, 25-40 (2005).

[26] Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520-D528 (2019).

[27] Estrada, E. Quantifying network heterogeneity. *Phys. Rev. E* **82** 066102 (2010).

[28] Estrada, E. Degree heterogeneity of graphs and networks. I. Interpretation and the "heterogeneity paradox". *J. Interdisc. Math.* **22**, 503-529 (2019).

[29] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world'networks. *Nature* **393**, 440-442 (1998).

[30] Newman, M. E. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).

[31] Freeman, L. C. Centrality in social networks: Conceptual clarification. *Social Networks* **1**, 215–239 (1979).

[32] Bonacich, P. Power and centrality: A family of measures. *Am. J. Soc.* **92**, 1170-82 (1987).

[33] Estrada, E. & Rodriguez-Velazquez, J. A. Subgraph centrality in complex networks. *Phys. Rev. E* **71**, 056103 (2005).

[34] Estrada, E. & Hatano, N. Communicability in complex networks. *Phys. Rev. E* **77**, 036111 (2008).

[35] Estrada, E. & Hatano, N. Communicability angle and the spatial efficiency of networks. *SIAM Rev.* **58**, 692-715 (2016).

[36] Estrada, E. & Silver, G. Accounting for the role of long walks on networks via a new matrix function. *J. Math. Anal. Appl.* **449**, 1581-600 (2017) .