scientific reports

Check for updates

OPEN

Computational analysis of the sequence-structure relation in SARS-CoV-2 spike protein using protein contact networks

Pietro Hiram Guzzi[1,5]✉, Luisa di Paola[2,5], Barbara Puccio[1], Ugo Lomoio[1], Alessandro Giuliani[3] & Pierangelo Veltri[1,4]

The structure of proteins impacts directly on the function they perform. Mutations in the primary sequence can provoke structural changes with consequent modification of functional properties. SARS-CoV-2 proteins have been extensively studied during the pandemic. This wide dataset, related to sequence and structure, has enabled joint sequence-structure analysis. In this work, we focus on the SARS-CoV-2 S (Spike) protein and the relations between sequence mutations and structure variations, in order to shed light on the structural changes stemming from the position of mutated amino acid residues in three different SARS-CoV-2 strains. We propose the use of protein contact network (PCN) formalism to: (i) obtain a global metric space and compare various molecular entities, (ii) give a structural explanation of the observed phenotype, and (iii) provide context dependent descriptors of single mutations. PCNs have been used to compare sequence and structure of the Alpha, Delta, and Omicron SARS-CoV-2 variants, and we found that omicron has a unique mutational pattern leading to different structural consequences from mutations of other strains. The non-random distribution of changes in network centrality along the chain has allowed to shed light on the structural (and functional) consequences of mutations.

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has accounted for over 657 million infections and over 6.8 million deaths at the end of the 2022 (Data from https://covid19.who.int)[1]. SARS-CoV-2 is a large enveloped coronavirus (family-Coronaviridae, subfamily-Coronavirinae) with non-segmented, single-stranded, and positive-sense RNA genomes[2]. SARS-CoV-2 is composed of the spike (S), nucleocapsid (N), membrane (M), and envelope (E) proteins, of 16 non-structural proteins (NSP1-NSP16), and six accessory proteins (NS3, NS6, NS7a, NS7b, NS8, and ORF10). The Spike protein (S) infects human cells by binding the ACE2 human receptor[3–5]. During the pandemic, many mutant strains emerged, the vast majority of which had a very short life span and limited spatial distribution. In very few cases, the emerging mutant became the most common strain, with the sudden disappearance of the other variants.

Viruses, with their often small genomes and error-prone replication mechanisms, adapt very rapidly to changing micro-environmental cues. Moreover, the huge number of replications make it possible to observe their 'natural evolution in real time' as they acquire antiviral drug resistance or mediate persistent infection through escape from T and B cell immune system responses to infection[6]. This behaviour can be equated with a sort of Darwinian struggle for life in which 'the fittest strain' is the most efficient in terms of reproductive success. In the case of viruses, reproductive success is strictly related to its infective power, i.e. on the ability to enter into the host cells and reproduce. One consequence of this is the tendency of emerging infections to become milder in terms of their pathological effect (in some cases creating a symbiotic relationship[7,8]) thereby enabling more rapid diffusion across host (in this case human) populations. The above considerations allow to understand why proteins at the interface between viral particle and host cell are the key to understanding the evolution of the viruses.

[1]Department of Surgical and Medical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy. [2]Unit of Chemical-Physics Fundamentals in Chemical Engineering, Department of Engineering, Universita Campus Bio-Medico di Roma, via Alvaro del Portillo 21, 00128 Rome, Italy. [3]Environment and Health Department, Istituto Superiore di Sanita, Rome, Italy. [4]Department of Computer, Modeling, Electronics and System Engineering, University of Calabria, Rende, Italy. [5]These authors contributed equally: Pietro Hiram Guzzi and Luisa di Paola ✉email: hguzzi@unicz.it

The importance of S protein for infectivity and the consequent spread of the SARS-CoV-2 virus, and the fact that vaccines are tailored upon this protein, made S protein the privileged point of departure for the study of natural history of viruses in structural/bioinformatic terms. Sequence analyses of S protein have revealed the emergence of new SARS-CoV-2 mutation hotspots whose random or selection-driven character is hotly debated[1,9–11]. Here, we focus on a number of selected variants: Alpha, Beta, Gamma, Delta, and Omicron. These variants have different transmission rates, evolutionary patterns and levels of vaccine resistance and became predominant one after the other. Alpha prevailed over the original Wuhan strain, then Delta overcame Alpha and Omicron became virtually the only one left. The Omicron variant, identified in February 2022, has spread faster than earlier variants[12,13] while its effects have proven to be less severe than previous strains. The Delta variant, isolated in a region of India in October 2020, has emerged as the dominant global variant alongside the Alpha[14].

Since April 2021, the literature has been enriched by many works focusing on the impact of variants on SARS-CoV-2 S protein modification[15–19] and the importance of mutations at the sequence level has been considered a crucial step in the analysis and prediction of variants. These studies were made possible thanks to the availability of a large volume data sets. The study of mutational landscape at the sequence level has been facilitated by the large volume of data stored in public databases (such as the GISAID database)[20]. On the contrary, the impact of the mutation at the structural level suffers from the lack of experimental data on protein structures. Existing structures stored in PDB database[21] are mainly related to various structural domains of the Spike protein and its mutations[17].

We present a systematic sequence-structure analysis of Spike protein for the three variants Alpha, Delta, and Omicron. We analysed the variants according to the following procedure: (i) we first checked the mutual differences between strains in sequence space; (ii) we computed the between strain differences in terms of residue contact network distances; (iii) we checked the existence of a general linear relation between sequence and structure metrics; (iv) we built up a phenomenological sequence/structure relation in terms of PCN descriptors of single mutated residues. We relied on our experience in Protein Contact Network (PCN) framework to perform the first sequence-structure analysis. Protein Contact Networks (PCNs)[22,23] can catch the protein structure modularity at the basis of domain functional partition of protein molecular structures and allosteric regulation[24–26]. We compared the selected variants in terms of network invariants stemming from the PCN approach, complementing the classical sequence-based comparison. The distance-to-distance correlation analysis[27] on global sequence and structure has 21 mutated residues (while the other two strains only 6 and 4 point mutations)hence an outlier pointing to a clear separation (in both sequence and structure metrics) of this strain with respect to the other two. Focusing on changes in the Eigenvector Centrality (EC) descriptor[28], the unique mutational pattern of Omicron with its highly non-random distribution of EC changes emerged as a cue for rationalising the functional consequences of the observed mutations. We report a marked excess of centrality value in the Receptor Binding Domain (RBD), and a decrease in the cleavage-allosteric region, pointing to a significant restructuring consistent with phenotype changes (vaccine escape, lower lethality) and finally observed structural modifications of the Omicron strain[29]

Related work. The study of protein sequence-structure (also referred to as Protein sequence-structure analysis - PSSA -) consists of an integrated analysis of protein sequences and structures. In recent years extensive research has been devoted to prediction by means of sequence comparison and alignments of sequences and structures[30–32]. The current outbreak of COVID-19 has opened up an unprecedented field of application for such methods. In[33], the authors analysed Membrane (M) and Envelope (E) proteins of COVID-19 and the comparison to *homologous* proteins in MERS, SARS, and bat viruses, and found that many regions of E and M proteins of SARS-CoV-2 are similar to SARS and bat ones. Conversely, the MERS virus *proteins* differ in many respects.

Cherian et al.[34], analysed four mutations of the Spike protein (L452R, T478K, E484Q and P681R), during the second wave of COVID-19 in Maharashtra (India). In particular, focusing on the impact of these mutations on the RBD domain structure, they found a possible increase of the ACE2 binding affinity in L452R, T478K, E484Q mutation, while postulating an increased transmission rate for P681R. The analysis of mutations of Spike protein has also been also treated in[35,36], focusing on mutations affecting antigenicity. Ortuso et al.[17] found certain Spike mutations in the RBD : S477N, N439K, N501Y, Y453F, E484K, K417N, S477I and G476S. Among these, they found that mutation N501Y, in particular, is one of the characteristic features of the SARS-CoV-2 Delta. Using mutation analysis for SaRS-CoV-2, Di Giacomo et al.[37] reported a study on T478K mutation of S protein, integrating both sequence and structure analysis. All the above studies have employed a local (typical of structural biology and biochemistry studies) approach aimed at rationalising a structural (and/or functional) phenotype based on mutations present in the primary sequence.

The study of the mutations has been stimulated by the availability of a large amount of data over the world and each lineage has been characterised in terms of mutation, spatial distribution, and impact on disease evolution and transmissibility. For instance, the mutations of the Delta variants on S protein have shown a reduction in the reaction to antibodies[38]. Similarly, the mutation-based analysis for the Gamma variant evidenced a neutralising reduction of some antibodies[39].

Moreover, the effects of the mutations on the vaccination have been investigated in many works such as[4,40–43]. Genomics sequence integration has been proposed in[44] where a library of human antibodies and topological analysis of the sequences evidenced the evolution for S proteins and their impact on vaccination. They found that most common variants present the strengthening of infectivity and vaccine escape on the RBD domain of S protein. They also found that infectivity strengthening results from the evolution of the virus, and the emergence of possible vaccine-escape mutations is more likely to occur in highly vaccinated populations. Moreover, in[45] the authors point to the ability of the Omicron strain to mutate in order to reduce the effect of the neutralising antibodies, while keeping a close affinity with the ACE2 receptor[45].

In this work we used sequence-structure analysis and a systemic perspective. We some might argue that, since we consider only three mutant strains, we have little information[28], making the overall sequence structure-relation severely biased. However, the computation of network centrality descriptors at the level of single mutated residues provides allow a direct appreciation of structural consequences of each sequence change. Indeed, we noted that the same point mutation presents many different behaviours in terms of topological changes of the network representing the S protein. This is a proof-of-concept of the unique feature of the proposed approach to translate purely local information into systemic terms.

Results and discussion

Sequence analysis. We analysed the sequence of the Spike protein considering three selected variants (Alpha, Delta, Omicron) in three different states (Closed, Open 1RBD-Up, Complex with ACE2). We obtained the best multiple alignment of considered sequences using Clustal Omega routine[46]. Afterwards, we created a distance matrix from a multiple sequence alignment, calculating the evolutionary distance between each pair of sequences in a multiple sequence alignment[47].

Structure analysis through protein contact networks. We build PCNs using PCN-miner as depicted in Fig. 1. (PCNs)[22] allow us to model a protein structure into a graph that can be analysed[22,23,48]. PCNs are networks whose nodes represent the $C-\alpha$ atoms of the backbone of proteins, while their edges represent a relative spatial distance between 4 and 8 angstroms. Topological descriptors of PCNs, such as centrality measures, are used to discover protein properties such as allosteric regions[24,48,49]. The structural distances between PCNs are computed by means of the Frobenius metric, which stems from the pair-wise comparison across 18 adjacency matrices (6 mutants and 3 aggregation states). The Frobenius metrics indicates the number of corresponding pairs of residues differing in the two structures. The Frobenius norm between two matrices is defined as the square root of the sum of the absolute squares of their elements.

Correlation between sequence and structure distance matrices. We studied the correlation between the sequence distances and the related structure distances, as shown in Fig. 2. The x-axis of the Figure represents the pair-wise sequence distances that are identical for the three aggregation states and depend only upon the differences in the primary structure, while the y-axis corresponds to the pair-wise mutual distances relative to the three aggregation states for the mutants. Each point represents the correlation of a pair sequence-structure. We report results for the wild-type, alpha, delta and omicron variants and the three different structural conformations, open, closed and complex with ACE2.

The Figure shows that there is no correlation between sequence and structure. In other words, the entity of sequence distance has no explanatory content for the resulting structural distance. We should note that the Omicron variant is by far the most mutated species of all other forms bearing 21 mutated locations, while alpha

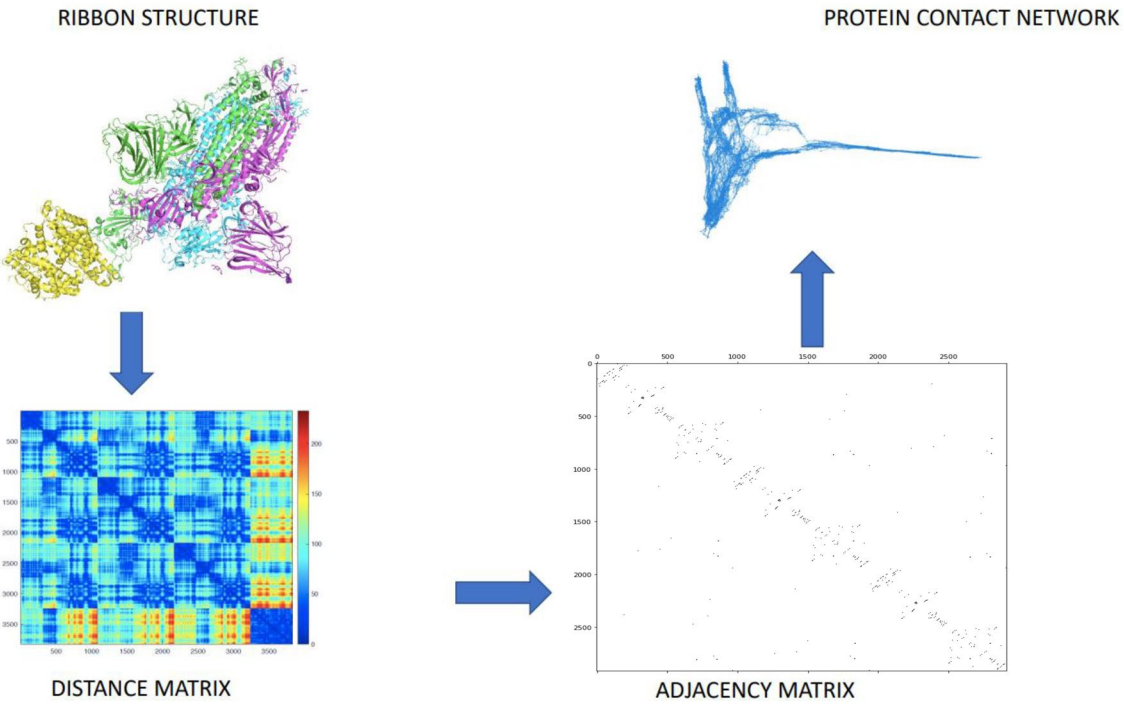

Figure 1. Scheme of the PCN construction on the close conformation of SARS-CoV 2 spike protein (PDB code 6vyb): starting from the structure (upper left). it is possible to compute the distance matrix (lower left), then the adjacency matrix (lower right) and finally the PCN (upper right).

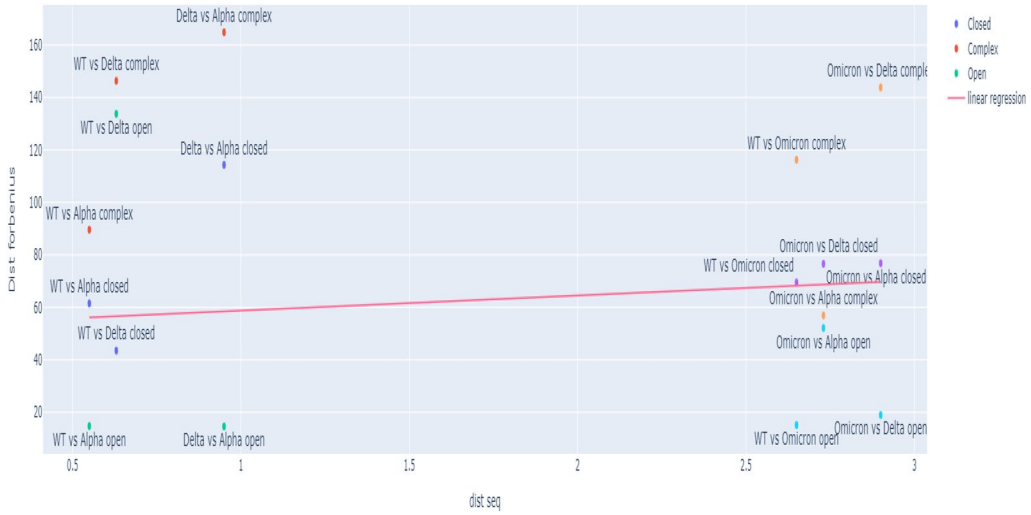

Figure 2. The x-axis of the Figure represents the distance of the sequences while the y-axis the distance of the structures. Each point represents the correlation of a pair sequence/structure. We report the wild type, alpha, delta and omicron variant and three structural conformations open, closed and complex.

and delta strains have 6 and 4 mutated locations, respectively. The range restriction effect completely hides any general sequence/structure relation that collapses to the pure registration of the singular position of the Omicron strain in both sequence and structure spaces. This well-known fact[50] evidence of the continuity breaking of sequence distance with the oncoming of a mutant (Omicron) bearing a higher number of mutations than the others is a puzzling event pointing to strong selective pressure regarding random mutational events. The Omicron strain is notably different from the others, as unfortunately shown by the poor performance of the vaccines[51,52]. Second, we note that the wider distances of structures (y-axis) are related to the complex of S and ACE2 proteins; this may be related to the higher infectivity rate of this strain. This result tells us that structural (and consequently functional) differences are highly context-dependent at the fine scale of analysis, preventing any simple extrapolation from sequences.

Network invariants. Having assessed the singular position of Omicron vis-à-vis Alpha and Delta strains by both sequence and structure (entire network wiring) spaces, network invariant analysis enables us to explain the structural changes. The change in network invariants regarding the initial Wuhan strain was estimated in terms of log ratio. Thus a value of zero (0 in number) corresponds to no change. Positive and negative values point to an increase or decrease, respectively. We calculated main centrality measures for each node: Closeness (CC), Eigenvector EC, and Betweenness (BC) Centrality[53,54]. The Closeness Centrality (CC) measures how close the nodes are to each other in term of shortest paths. The Eigenvector Centrality (EC) indicates the importance of a node in a network. The Betweenness Centrality (B)C estimates show how many shortest paths go through a given node, thereby revealing its important role in signal transmission throughout the network. Recently, Barozi and coworkers applied these centrality metrics to analyse the Molecular Dynamics of the Omicron S-protein by identifying a specific evolutionary pattern towards an increased allosteric regulation of the S-protein RBD-hACE2 binding[55].

Only the EC showed a striking variation from a global null effect with Omicron, highlighting a marked and statistically significant difference regarding the other strains ($F = 14.03$, $p < 0.0001$ as for an absolute change in EC). It is worth stressing that this result has no necessary relation with the number of mutated sites, given that it corresponds to the average change per mutated residue (see Table 1, and Fig. 3).

We also report in Table 2 that the EC changes, setting a threshold of relevant change at EC = |2| and considering values less than −2 (i.e., < − 2) as 'negative' and greater than 2 (i.e. values > 2) as 'positive', we note that only Omicron mutations lead to the changes in their role in PCN wiring. The distribution of these mutations is far from random; positive changes are only present in RBD domain, while negative changes concentrate on the splicing domain[29]. The Eigenvector Centrality (EC) of a node can be equated to the loading of a variable on the first principal component of a multivariate data set[56]. Thus, the increase of EC of a set of nodes of the same protein domain corresponds to a drastic contraction of the domain marked by an increase in the amount of structural variance explained by the first component. On the contrary, a *proxy* of structural relaxation corresponds to a decreased value of EC coordinates.

This implies that the Omicron RBD has a more compact structure than the original strain, which is consistent with observations in other works, such as[29]. The analysis is labor intensive as well as dependent on the availability of the structures from the PDB database. In addition, we observed relaxation of the splicing domain. Considering that the splicing domain encompasses the allosteric site of spike protein, we achieve a highly consistent structural explanation of this peculiar phenotype of Omicron. We also observe that Omicron presents these characteristics: (i) from a molecular point of view it exhibits stabilisation of RBD and increased sensitivity

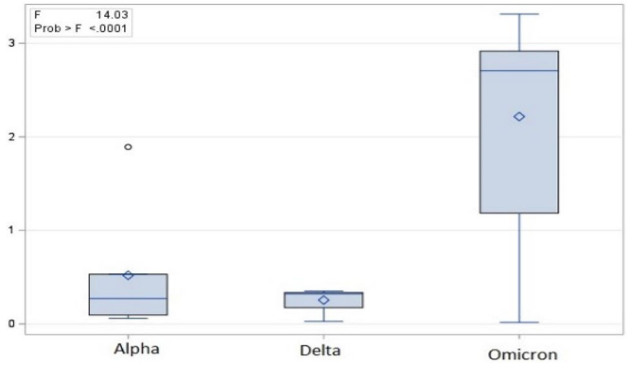

Figure 3. The absolute values of eigenvector centrality (EC) variations relative to the three analysed strains are reported as box-plots. In terms of the changes it is evident that the position of Omicron is quite different from that of the other strains. The use of module values is crucial for a statistically significant result ($F = 14.03$, $p < 0.0001$).

Strain	Mean (EC)	Mean (|EC|)	SD (EC)	SD (|EC|)
Alpha	− 0.469	0.521	0.739	0.696
Delta	0.066	0.256	0.324	0.153
Omicron	0.988	2.147	2.201	1.015

Table 1. The table reports mean and standard deviation for both real and absolute ($|EC|$) values of eigenvector centrality for the three strains. It is worth noting the neat departure from no effect for the Omicron strain together with the elevated standard deviation (SD (EC)) of current values pointing to the presence of both highly positive and negative EC changes for mutated residues. On the contrary, for the other two strains mutations do not provoke any important change in EC.

to microenvironment allosteric signalling (relaxation of splicing domain); (ii) from a phenotypic point of view, data from the literature reports higher infectivity and vaccine escape. Thus, we hypotesise that molecular changes determine the changes to the phenotype.

A final remark is related to the considerably different EC change provoked by the mutation of the same residues by the same amino acid substitution as for 501 and 614 positions in the three strains (Table 2). While these mutations do not provoke any notable EC change in Alpha and Delta strains, they culminate in a drastic EC change in Omicron. This provides evidence of the 'context dependent' character of network invariants of mutation patterns with respect to the purely local consideration of both the sequence-based and global structural views.

Methods

Protein contact networks. A protein structure can be represented as a complex three-dimensional object formally defined by the coordinates in the 3D space of its atoms. Despite the wide availability of data on protein molecular structures, the protein structure-function relationship is far from fully understood. For this reason, it is necessary to define simple descriptors that can describe protein structures with few numerical variables. Structure and function are based on the complex network of inter-residue interactions, where residues are identified by amino acid sequences[22]. The interaction of residues, therefore, is the way to define protein contact networks (PCN) that represent the protein structure by means of α-carbon location. The spatial position of C_α is still reminiscent of the protein backbone, and this allows us to also highlight the most important features of the three-dimensional structure. Starting from spatial distribution of the C_α, a distance matrix d is evaluated where each $d_{i,j}$ represents the Euclidean distance in the 3D space between the i-th and j-th residues, defined as

$$d_{i,j} = \sqrt{((x_i - x_j)^2) + ((y_i - y_j)^2) + (z_i - z_j)^2)} \tag{1}$$

where (x_i, y_i, z_i) and (x_j, y_j, z_j) respectively are the Cartesian coordinates of residue i and j. Matrix d is used to define a Protein Contact Network[2257]. It is possible to build up adjacency matrix A, whose generic element is defined as:

$$A_{ij} = \begin{cases} 1 & \text{if} \quad 4 \leq d_{ij} \leq 8 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Thus, we define a link between two residues i and j if their mutual distance lies between 4 and 8 Å. The lower end excludes all covalent bonds, which are not sensitive to environmental change (hence to protein functionality), while the upper end eliminates of weaker non-covalent bonds (hence not significant for protein functionality).

Strain	Mutated residue	modEC	effect
Alpha	501	− 1.89267	Null
Alpha	570	− 0.16767	Null
Alpha	614	− 0.375	Null
Alpha	716	0.060667	Null
Alpha	1118	0.095333	Null
Delta	142	− 0.352	Null
Delta	452	0.321667	Null
Delta	614	0.321667	Null
Delta	950	0.028	Null
Omicron	67	− 0.786	Null
Omicron	142	− 1.08433	Null
Omicron	339	2.449667	Positive
Omicron	371	3.125667	Positive
Omicron	373	2.703333	Positive
Omicron	375	2.721333	Positive
Omicron	417	3.207	Positive
Omicron	440	2.709667	Positive
Omicron	493	2.906333	Positive
Omicron	496	2.876	Positive
Omicron	498	2.930667	Positive
Omicron	501	2.826	Positive
Omicron	547	1.134667	Null
Omicron	614	− 2.14667	Negative
Omicron	655	− 2.32667	Negative
Omicron	764	− 0.82067	Null
Micron	796	− 3.03567	Negative
Omicron	856	1.23533	Null
Omicron	954	0.017667	Null
Omicron	969	3.31266	Null

Table 2. The EC change (logratio) with respect to the original strain for each mutated amino acid residue is reported. The absence of any relevant change in EC of the Alpha and Delta strain is worth than nothing, while 14/21 (67%) Omicron mutations imply a significant change with respect to the original strain. Moreover the distribution of positive and negative changes is far from random, positive changes being concentrated in RBD domain, while negative changes are only found in the splicing domain.

The adjacency matrix of a graph is unique regarding the ordering of nodes. With proteins where the order of nodes (residues) corresponds to the residue sequence (primary structure), the evidence shows that its corresponding network is unique: this establishes a one-to-one correspondence between the protein and its network.

In the case of SARS CoV-2 spike protein, this formalism has been used to detect the allosteric site of the S protein[58] through an integrated structural/dynamic approach[59].

Finally, we consider the distance between two PCN as the Frobenius distance of their adjacency matrices.

Datasets. We consider SARS-CoV-2 genomic sequences extracted from GISAID[20] database on March 2022. Sequences used in this work can be found at https://github.com/hguzzi/Multiscalemodelling. We downloaded thirteen protein structures from the Protein Data Bank (PDB https://www.rcsb.org/): 6vxx, 7wk2, 7sbk, 7fet, 6vyb, 7edf, 7w92, 7tgw, 7df4, 7fem, 7wk4, 7w98, 7vxm. Coordinates of the Carbon-α atoms were used to obtain PCNs.

Sequence comparison. Sequence alignment was performed by the Smith-Waterman and the Clustal Omega algorithm[46]. Regarding pair-wise alignment, we used EMBOSS[47] (European Molecular Biology Open Software Suite), that is a high-quality package of open source software tools for molecular biology. It uses the Smith-Waterman algorithm (changed for speed enhancements) to calculate the local alignment of two sequences. For multiple alignment, we used Clustal Omega, a multiple sequence program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. The pairwise alignment tool contains a file that reports input parameters (i.e., two sequences in FASTA format), alignment obtained and expressed as a percentage. PCN-Miner software http://github.com/hguzzi/PCN-MINER was used to build PCNs[26].

Centrality measures. We considered following centrality measures with respect to PCNs: degree, closeness, betweenness and eigenvector.

Degree centrality is the number of adjacent nodes to w_i which can be defined as follows.

$$C_{deg}(w_i) = deg(w_i).$$

Closeness centrality is to be considered as a central node close to the others in terms of distance. Formally, the closeness centrality of node w_i is the reciprocal of the average shortest distance to w_i over all $n-1$ reachable nodes, i.e.

$$C_{closeness}(w_i) = \frac{n-1}{\sum_{j=1}^{j=n-1, j\neq i} d(w_i, w_j)}.$$

where $d(w_i, w_j)$ is the shortest distance between w_i and w_j.

Given an adjacency matrix $\mathcal{A}$, the relative centrality score a node v can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{w \in Neigh(v)} x_w = \frac{1}{\lambda} \sum_{w \in \mathcal{G}} \mathcal{A}_{v,w} x_w$$

Eigenvector centrality estimates the influence of a node in a network. It scores the nodes of a network on the basis of the idea that high-scoring nodes contribute more to the score of the node than connections to low-scoring nodes. It has been shown that eigenvector centrality identifies the role of residues in allosteric signal transmission, both on a local and global scale[60].

Given an unweighted undirected graph G and its adjacency matrix $\mathcal{A}$ we can estimate the EC (x_v) for each node v as

$$x_v = \frac{1}{\lambda} \sum_{w \in Neigh(v)} x_w = \frac{1}{\lambda} \sum_{w \in \mathcal{G}} \mathcal{A}_{v,w} x_w$$

where $Neigh(v)$ is the set of neighbours of v, and λ is a constant. The previous equation may be written as in vector notation as the eigenvector equation $\mathbf{Ax} = \lambda \mathbf{x}$, where λ is an eingenvalue for which a non-zero eigenvector solution exists.

Betweenness centrality is defined as follows:

$$C_{betweennes}(w_i) = \sum_{i \neq j \neq k} \frac{\sigma_{j,k}(i)}{\sigma_{j,k}}$$

where $\sigma_{j,k}$ is the total number of the shortest paths from node w_j to node w_k and $\sigma_{j,k}(i)$ is the number of those paths that pass through i.

Conclusion

It is widely recognised that mutations of protein sequences impact first on their structure and then their function. The recent pandemic has provided an unprecedented scenario for the analysis of protein mutation, focusing on the mutations of SARS-CoV-2 viral proteins. We relied on this information to give a proof-of-concept of the 'quantum-leap' in terms of extraction of hypothesis on structure-function relations provided by a mesoscopic approach, such as PCN. Our results clearly show that mutations in the Omicron sequence cause the increase and the decrease of EC in two distinct regions. Moreover, as evidenced by the ANOVA test, Omicron mutations, regardless of the number and region, cause a more marked shift in EC, confirming their different pattern of mutation. As regard sequence/structure/function protein studies, this result leads to a shift from episodic local considerations of single mutations to a context-dependent evaluation of structural consequences of point mutations along the lines of Quantitative Structure Activity (QSAR) studies of small organic molecules.

Data availability

The website https://github.com/hguzzi/Multiscalemodelling contains data and code used in this work. More material may be shared upon reasonable request. Please contact Pietro Hiram Guzzi hguzzi@unicz.it for any request.

Received: 1 November 2022; Accepted: 15 February 2023
Published online: 17 February 2023

References

1. Kumar Das, J., Tradigo, G., Veltri, P., Guzzi, H. & Roy, P. S. Data science in unveiling covid-19 pathogenesis and diagnosis: Evolutionary origin to drug repurposing. *Briefings Bioinform.* **22**, 855–872 (2021).
2. Guzzi, P. H., Mercatelli, D., Ceraolo, C. & Giorgi, F. M. Master regulator analysis of the sars-cov-2/human interactome. *J. Clin. Med.* **9**, 982 (2020).
3. Gordon, D. E. *et al.* A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
4. Guzzi, P. H., Petrizzelli, F. & Mazza, T. Disease spreading modeling and analysis: A survey. *Briefings Bioinform.* **23**(4), bbac230. https://doi.org/10.1093/bib/bbac230 (2022).
5. Satarker, S. & Nampoothiri, M. Structural proteins in severe acute respiratory syndrome coronavirus-2. *Arch. Med. Res.* **51**, 482–491 (2020).

6. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war-host adaptation and its constraints on virus evolution. *Nat. Rev. Microbiol.* **17**, 321–328 (2019).
7. Roossinck, M. J. Symbiosis versus competition in plant virus evolution. *Nat. Rev. Microbiol.* **3**, 917–924 (2005).
8. Modonesi, C. & Giuliani, A. Epidemiology, ecology, and evolution of human-virus interaction: An overview of the relevance to human health and disease. *Org. J. Biol. Sci.* (2021).
9. López-Cortés, G. I. *et al.* The spike protein of sars-cov-2 is adapting because of selective pressures. *Vaccines* **10**, 864 (2022).
10. Vizza, P., Curcio, A., Tradigo, G., Indolfi, C. & Veltri, P. A framework for the atrial fibrillation prediction in electrophysiological studies. *Comput. Methods Prog. Biomed.* **120**, 65–76 (2015).
11. Mercatelli, D., Triboli, L., Fornasari, E., Ray, F. & Giorgi, F. M. Coronapp: A web application to annotate and monitor sars-cov-2 mutations. *J. Med. Virol.* **93**, 3238–3245 (2021).
12. He, X., Hong, W., Pan, X., Lu, G. & Wei, X. Sars-cov-2 omicron variant: Characteristics and prevention. *MedComm* **2**, 838–845 (2021).
13. Hui, K. P. *et al.* Sars-cov-2 omicron variant replication in human bronchus and lung ex vivo. *Nature* **603**, 715–720 (2022).
14. Bian, L. *et al.* Impact of the delta variant on vaccine efficacy and response strategies. *Expert Rev. Vaccines* **20**, 1201–1209 (2021).
15. Kumar, S., Thambiraja, T. S., Karuppanan, K. & Subramaniam, G. Omicron and delta variant of sars-cov-2: A comparative computational study of spike protein. *J. Med. Virol.* **94**, 1641–1649 (2022).
16. Galicia, J. C., Guzzi, P. H., Giorgi, F. M. & Khan, A. A. Predicting the response of the dental pulp to sars-cov2 infection: A transcriptome-wide effect cross-analysis. *Genes Immun.* **21**, 360–363 (2020).
17. Ortuso, F., Mercatelli, D., Guzzi, P. H. & Giorgi, F. M. Structural genetics of circulating variants affecting the sars-cov-2 spike/human ace2 complex. *J. Biomol. Struct. Dyn.* **40**, 6545–6555 (2021).
18. Cannataro, M., Guzzi, P. H., Mazza, T., Tradigo, G. & Veltri, P. Using ontologies for preprocessing and mining spectra data on the grid. *Future Gener. Comput. Syst.* **23**, 55–60 (2007).
19. Dubanevics, I. & McLeish, T. C. Computational analysis of dynamic and control in the sars-cov-2 main protease. *J. R. Soc. Interface* **18**, 20200591 (2021).
20. Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data-from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
21. Sussman, J. L. *et al.* Protein data bank (pdb): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**, 1078–1084 (1998).
22. Di Paola, L., De Ruvo, M., Paci, P., Santoni, D. & Giuliani, A. Protein contact networks: An emerging paradigm in chemistry. *Chem. Rev.* **113**, 1598–1613 (2013).
23. Gu, S., Jiang, M., Guzzi, P. H. & Milenković, T. Modeling multi-scale data via a network of networks. *Bioinformatics* **38**, 2544–2553 (2022).
24. Khan, T. & Ghosh, I. Modularity in protein structures: Study on all-alpha proteins. *J. Biomol. Struct. Dyn.* **33**, 2667–2681 (2015).
25. Guzzi, P. H. & Zitnik, M. Editorial deep learning and graph embeddings for network biology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 653–654 (2022).
26. Guzzi, P. H., Di Paola, L., Giuliani, A. & Veltri, P. Pcn-miner: An open-source extensible tool for the analysis of protein contact networks. *Bioinformatics* **38**, 4235–4237 (2022).
27. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794. https://doi.org/10.1214/009053607000000505 (2007).
28. Giuliani, A., Zbilut, J. P., Conti, F., Manetti, C. & Miccheli, A. Invariant features of metabolic networks: A data analysis application on scaling properties of biochemical pathways. *Phys. A Stat. Mech. Appl.* **337**, 157–170. https://doi.org/10.1016/j.physa.2004.01.053 (2004).
29. Cui, Z. *et al.* Structural and functional characterizations of infectivity and immune evasion of sars-cov-2 omicron. *Cell* **185**, 860-871.e13. https://doi.org/10.1016/j.cell.2022.01.019 (2022).
30. Pearce, R. & Zhang, Y. Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* **297**(1), 100870. https://doi.org/10.1016/j.jbc.2021.100870 (2021).
31. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
32. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
33. Aldaais, E. A., Yegnaswamy, S., Albahrani, F., Alsowaiket, F. & Alramadan, S. Sequence and structural analysis of covid-19 e and m proteins with mers virus e and m proteins-a comparative study. *Biochem. Biophys. Rep.* **26**, 101023. https://doi.org/10.1016/j.bbrep.2021.101023 (2021).
34. Cherian, S. *et al.* Sars-cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of covid-19 in Maharashtra, India. *Microorganisms* **9**, 1542 (2021).
35. Harvey, W. T. *et al.* Sars-cov-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
36. Das, J. K., Roy, S. & Guzzi, P. H. Analyzing host-viral interactome of sars-cov-2 for identifying vulnerable host proteins during covid-19 pathogenesis. *Infect. Genet. Evol.* **93**, 104921 (2021).
37. Di Giacomo, S., Mercatelli, D., Rakhimov, A. & Giorgi, F. M. Preliminary report on severe acute respiratory syndrome coronavirus 2 (sars-cov-2) spike mutation t478k. *J. Med. Virol.* **93**, 5638–5643 (2021).
38. Tao, K. *et al.* The biological and clinical significance of emerging sars-cov-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).
39. Hitchings, M. D. *et al.* Effectiveness of chadox1 vaccine in older adults during sars-cov-2 gamma variant circulation in São Paulo. *Nat. Commun.* **12**, 1–8 (2021).
40. McLean, G. *et al.* The impact of evolving sars-cov-2 mutations and variants on covid-19 vaccines. *Mbio* **13**, e02979-21 (2022).
41. Sassi, M. B. *et al.* Phylogenetic and amino acid signature analysis of the sars-cov-2s lineages circulating in Tunisia. *Infect. Genet. Evol.* **102**, 105300 (2022).
42. Mlcochova, P. *et al.* Sars-cov-2 b. 1.617. 2 delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).
43. Petrizzelli, F., Guzzi, P. H. & Mazza, T. Beyond covid-19 pandemic: Topology-aware optimization of vaccination strategy for minimizing virus spreading. *Comput. Struct. Biotechnol. J.* **20**, 2664–2671 (2022).
44. Wang, R., Chen, J., Hozumi, Y., Yin, C. & Wei, G.-W. Emerging vaccine-breakthrough sars-cov-2 variants. *ACS Infect. Dis.* **8**, 546–556 (2022).
45. Cao, Y. *et al.* Imprinted sars-cov-2 humoral immunity induces convergent omicron rbd evolution. *Nature* 1–3 (2022).
46. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* **7**, 539 (2011).
47. Rice, P., Longden, I. & Bleasby, A. Emboss: The European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
48. Di Paola, L., Hadi-Alijanvand, H., Song, X., Hu, G. & Giuliani, A. The discovery of a putative allosteric site in the sars-cov-2 spike protein using an integrated structural/dynamic approach. *J. Proteome Res.* **19**, 4576–4586 (2020).
49. Tasdighian, S. *et al.* Modules identification in protein structures: The topological and geometrical solutions. *J. Chem. Inf. Model.* **54**, 159–168 (2014).
50. Gobeil, S.M.-C. *et al.* Structural diversity of the sars-cov-2 omicron spike. *Mol. Cell* **82**, 2050-2068.e6. https://doi.org/10.1016/j.molcel.2022.03.028 (2022).
51. Mannar, D. *et al.* Sars-cov-2 omicron variant: Antibody evasion and cryo-em structure of spike protein-ace2 complex. *Science* **375**, 760–764. https://doi.org/10.1126/science.abn7760 (2022).

52. Vizza, P. *et al.* Methodologies of speech analysis for neurodegenerative diseases evaluation. *International journal of medical informatics* **122**, 45–54 (2019).
53. Guzzi, P. H. & Roy, S. *Biological Network Analysis: Trends, Approaches, Graph Theory, and Algorithms* (Elsevier, 2020).
54. Ren, Y. *et al.* Pattern discovery in multilayer networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 741–752 (2021).
55. Barozi, V., Edkins, A. L. & Bishop, Ö. T. Evolutionary progression of collective mutations in omicron sub-lineages towards efficient rbd-hace2: Allosteric communications between and within viral and human proteins. *Comput. Struct. Biotechnol. J* **20**, 4562–4578 (2022).
56. Price, N. D., Reed, J. L., Papin, J. A., Famili, I. & Palsson, B. O. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys. J.* **84**, 794–804. https://doi.org/10.1016/s0006-3495(03)74899-1 (2003).
57. Di Paola, L. & Giuliani, A. Protein contact network topology: A natural language for allostery. *Curr. Opin. Struct. Biol.* **31**, 43–8. https://doi.org/10.1016/j.sbi.2015.03.001 (2015).
58. Di Paola, L., Hadi-Alijanvand, H., Song, X., Hu, G. & Giuliani, A. The discovery of a putative allosteric site in the sars-cov-2 spike protein using an integrated structural/dynamic approach. *J. Proteome Res.* **19**, 4576–4586. https://doi.org/10.1021/acs.jproteome.0c00273 (2020).
59. Di Paola, L., Mei, G., Di Venere, A. & Giuliani, A. Disclosing allostery through protein contact networks. In *Allostery: Methods and Protocols*. 7–20 (Springer, 2021).
60. Negre, C. F. *et al.* Eigenvector centrality for characterization of protein allosteric pathways. *Proc. Natl. Acad. Sci.* **115**, E12201–E12208 (2018).

Acknowledgements

PHG, PV, BP and UL were partly funded by the MISE PON-VQA project.

Author contributions

Conceptualization P.H.G., L.dP., A.G., P.V. Software and Data B.P., U.L. Manuscript Drafting P.H.G., L.dP., P.V., A.G., B.P. Manuscript Editing P.H.G., L.dP., A.G., P.V. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests .

Additional information

Correspondence and requests for materials should be addressed to P.H.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.