Graph-based coevolutionary approach on SARS-CoV-2 spike protein

Author: Alice Novell Mazzara
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain, and Barcelona Supercomputing Center, Girona 29, 08028 Barcelona, Spain.

Advisor: Marta Ibañes
BSC advisors: Camila Pontes and Victoria Ruiz

Abstract: Amino acids that coevolve can be indicative of functionality, so coevolution-based methods can be used to detect important amino acids in proteins, known as hotspots. Here, we apply a recently published method based on network metrics and coevolutionary information to detect functional hotspots in the SARS-CoV-2 spike protein. We found 275 potential hotspots with available experimental information in the literature for 4 of them, for example, position 614 (ASP), known to increment the infectivity of SARS-CoV-2 towards human host cells. In addition, a hotspot enrichment analysis was performed, as well as a study of the relative solvent accessibility of hotspot versus non-hotspot positions for the receptor binding domain. The hotspots showed less surface area available when bound to the human receptor compared to when not bound, which does not occur for non-hotspot positions, indicating that the hotspots obtained may be important for the binding of the spike protein to the receptor of the host cell.

I. INTRODUCTION

Proteins are complex molecules that play many critical roles in living organisms. By interacting with other molecules, they do most of the work in cells. Proteins are composed of hundreds of small units called amino acids, which are attached to each other in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's three-dimensional structure by a physical process called protein folding [1]. The correct three-dimensional structure is essential for the protein to perform its specific function.

The folding of a protein is a complex process, involving four stages, from a primary to a quaternary structure [2]. The primary structure is the linear sequence of amino acids (Fig. 1a). The secondary structure is generated by the formation of intramolecular hydrogen bonds, which folds the chains into either alpha-helices or beta-sheets (Fig. 1b). Tertiary structure is formed by the folding of the secondary structure sheets or helices into one another (Fig. 1c). This structure describes the three-dimensional shape of the protein. Quaternary structure results from tertiary structures interacting further with each other (Fig. 1d).

An amino acid is a molecule constituted by a hydrogen atom, a carboxylic acid group, an amine group, and a side chain, known as R group, specific to the type of amino acid. The four constituents are attached to a carbon atom, the α carbon. The R groups have different properties, such as shape, size, charge, and polarity. The interactions between the side chains of the different amino acids are what allow each protein to fold into a specific three-dimensional shape and perform its biological functions.

Some amino acids play a more important role than others. It can be argued that the most important amino acids are the ones responsible for performing the specific function of the protein. We consider these functional amino acids as "hotspots".

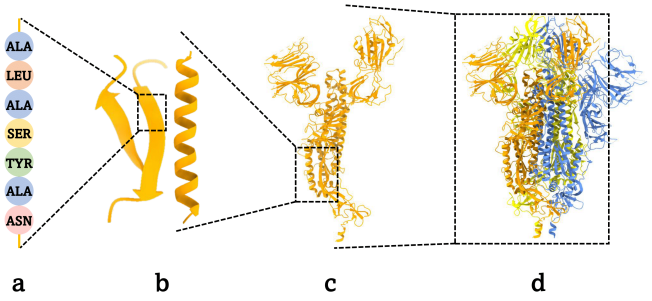

FIG. 1: Four structures of proteins. (a) Primary structure, (b) secondary structure, (c) terciary structure and (d) quaternary structure. The cartoon representation of the protein is obtained with ChimeraX [3].

The use of experimental methods for the identification of hotspots has proven laborious and time-consuming [4, 5]. In order to deal with this problem, many computational methods have been developed to predict hotspots [6–8].

One way to detect hotspots is through coevolution [9], an evolutionary process in which a heritable change in one entity establishes selective pressure for a change in another entity. These entities can range from nucleotides to amino acids, to proteins, to entire organisms.

In coevolution at the amino acid level, changes in specific amino acids trigger changes in other amino acids because of the pressure to maintain function. This means that amino acids in a sequence are depending on each other, they are correlated. We can study which amino acids in a family of sequences are correlated, which could be indicative of functionality.

Coevolution at the amino acid sequence level is studied

using multiple sequence alignments (MSA) [10]. A MSA is the result of aligning three or more related sequences adding gaps where necessary so that the maximum characters from each sequence are matched in each column, resulting in a rectangular array where each row is a different protein and each column is a position that can take the value of one of the 20 types of amino acids or a gap. Once sequences are aligned, patterns across them can be identified. Using this information, there are several methods available to determine which positions are correlated [11, 12].

In the present study, we focus on a recently published method based on coevolutionary information and network metrics [13] and apply it to detect hotspots in the particular case of the spike protein of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus. We validate the results with available experimental information and perform further analyses such as a hotspot enrichment analysis and a relative solvent accessibility assessment for the hotspots obtained lacking bibliographic information.

Studying SARS-CoV-2 is relevant in the current pandemic scenario since the newly detected hotspots can be, for instance, potential drug interaction targets.

Viruses of the coronavirus family have proteins protruding from their surface [14] (Fig. 2.a). These protrusions are known as spike proteins. The spike proteins play an important role in how these viruses infect their hosts as they bind only to certain receptors on the host cell. They are essential for both host specificity and viral infectivity [15]. Because of their crucial role in viral entry into cells, their study is of great interest to the development of vaccines and therapeutics, which is why we have chosen to investigate them.

More specifically, the spike protein is a homotrimer, i.e. it consists of three equal chains of amino acids (Fig. 2.b). Each chain has 1273 amino acids.Within the chain there are different regions, known as domains, that are distinguished because they perform different functions (Fig. 2.c).

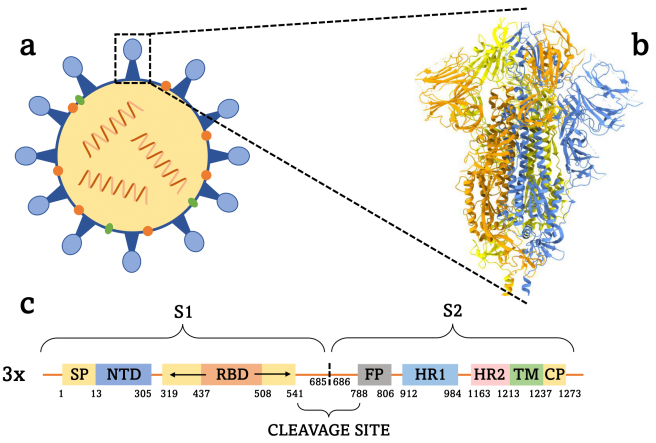

FIG. 2: (a) Schematic representation of SARS-CoV-2 virus, (b) cartoon representation the spike protein, with each chain in a different color and (c) domains of each chain of the spike protein. The cartoon representation of the protein is obtained with ChimeraX [3].

II. METHOD

In the following sections the workflow we followed is explained.

The code used to perform section B was provided by Dra. Camila Pontes and the code used for sections C and D was provided by the writers of the paper presenting the method [13].

A. Data

The data used was provided to me by my BSC advisors, as they had recently published a paper using the same data [16]. The input data are the files describing the three-dimensional structure of the spike protein of 122 different Betacoronavirus species. These files are usually found in the Protein Data Bank [17], where the structures of proteins are experimentally resolved. In this case, the three-dimensional structures were not obtained experimentally, they were predicted computationally using AlphaFold [18], an artificial intelligence program that makes protein structure predictions using deep learning.

In addition to the three-dimensional structure information of each protein, a MSA of the sequences is generated. The MSA was generated using the Mafft algorithm [19].

The method from the original paper explained in sections B, C and D is performed for each of the 122 proteins and the combination of the separate results is done according to section E.

B. Interaction matrix

The coevolutionary analysis protocol followed is based on Direct Coupling Analysis (DCA) [7, 8]. The objective is to detect statistical coupling between amino acid occupancies of any two columns of the MSA. The main inputs of DCA are reweighed frequency counts for single MSA columns and column pairs:

$$f_i(A) = \frac{1}{M_{eff}+\lambda}\left(\frac{\lambda}{q} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\right)$$
$$f_{ij}(A,B) = \frac{1}{M_{eff}+\lambda}\left(\frac{\lambda}{q^2} + \sum_{a=1}^{M}\frac{1}{m^a}\delta_{A,A_i^a}\delta_{B,A_j^a}\right) \quad (1)$$

where A_i^a (and B) refer to each position in the MSA, with i being the MSA column and a the MSA row. In this equation, q=21 for the number of different amino acids (also counting the gap). The weighting factor $1/m^a$ aims

at correcting for the sampling bias, where m^a is the number of similar sequences of A^a. $M_{eff} = \sum_{a=1}^{M} 1/m^a$ is the effective number of independent sequences. This equation also has a pseudo-count λ, a standard tool for estimating probabilities from counts in biological sequence analysis.

Then the maximum-entropy principle [20] is applied to obtain the least-constrained model $P(A_1, ..., A_L)$ possible. This model takes the mathematical form of as a Boltzman distribution with pairwise couplings $e_{ij}(A, B)$ and local fields $h_i(A)$:

$$P(A_1, ..., A_L) = \frac{1}{Z} exp\left\{ \sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\} \tag{2}$$

In this equation appears the normalization factor Z, known as the partition function in statistical physics. This function is defined as:

$$Z = \sum_{(A_1,...,A_L)} exp\left\{ \sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\} \tag{3}$$

Its direct calculation is infeasible for any realistic protein length and approximations have to be used. The approach used is based on a small-coupling expansion. The exponential $\sum_{i<j} e_{ij}(A_i, A_j)$ in (Eq. 3) is expanded into a Taylor series. We keep only the linear order of the expansion, obtaining the mean-field equations:

$$\frac{f_i(A)}{f_i(q)} = exp\left\{ h_i(A) + \sum_A \sum_{i \neq j} e_{ij}(A, B) f_j(B) \right\} \tag{4}$$

The equation that then allows to solve the original inference problem in mean-field approximation is:

$$e_{ij}(A, B) = -(C^{-1})_{ij}(A, B) \tag{5}$$

where $C_{ij}(A, B)$ is calculated:

$$C_{ij}(A, B) = f_{ij}(A, B) - f_i(A) f_j(B) \tag{6}$$

e_{ij} contains the interaction information for all amino acid pairs possible in all sequences. To study each protein, a new matrix E_{ij} is calculated, where only the interaction information for the amino acid pairs present in that protein is used.

C. Local interaction matrix

E_{ij} contains interaction information of all amino acid pairs of a given protein. To add structural information we calculate a contact matrix for all pairs of amino acids of the protein to act as a filter.

$$Q_{ij} = \begin{cases} 1, & \text{if } |r_i - r_j| \leq 10\text{Å} \\ 0, & \text{if } |r_i - r_j| > 10\text{Å} \end{cases} \tag{7}$$

where $|r_i - r_j|$ is the distance between the C_β's of the i-th amino acid and the j-th amino acid (the beta carbon C_β is the first atom of the R chain in the amino acid).

To combine the two matrices we perform a Hadamard product (an element-by-element multiplication of the two matrices) between the contact matrix and the interaction matrix, obtaining the local interaction matrix.

$$L_{ij} = Q_{ij} \otimes E_{ij} \tag{8}$$

If the distance between two amino acids is less than 10 Å their corresponding interaction term is considered, on the other hand, if the distance between the two amino acids is greater than 10 Å their interaction will be set to 0.

D. Network analysis

The next step is to construct a network from the local interaction matrix. First, each amino acid of the protein is considered a node of a weighted network. A weighted network is one in which a value is assigned to each edge connecting two nodes. In this network, we connect two nodes if the interaction value obtained in the local interaction matrix between the respective amino acids is greater than an iteratively defined threshold value: we start building a graph considering the maximum energy E_{ij} of the matrix, obtaining a non-connected graph (i.e. a graph of isolated nodes except the two residues with the strongest interaction). At this point, we iteratively lower the value until we obtain a connected graph.

Finally, to detect hotspots, we study the betweenness centrality parameter of the previously obtained network. The betweenness centrality of a node $B(k)$ is:

$$B(k) = \sum_{i,j} \frac{\xi(i, j|k)}{\xi(i, j)} \tag{9}$$

where $\xi(i, j|k)$ is the number of the shortest paths in the graph that connect i and j passing through k, and $\xi(i, j)$ is the number of the shortest paths in the graph that connect i and j.

A node is central in the network when the flow of information passes through it to connect different parts of the protein, i.e. the betweenness centrality of a central node will be higher. We consider an amino acid as a hotspot if the betweenness centrality of its associated node is greater than half of the maximum betweenness centrality of all nodes.

E. Filtering

The last step consists of filtering the data. For each protein, different hotspot positions are obtained. To combine this information, only hotspots that appear in 70% or more proteins are taken into account.

III. RESULTS AND ANALYSIS

275 potential hotspots have been found for the spike protein of the SARS-CoV-2 virus (Fig. 3). The hotspots at positions 343 (ASN), 501 (ASN), 614 (ASP), and 986 (PRO), appearing in orange in the figure, also appear on the list of experimentally found functional positions in UNIPROT [21]. Position 614 (ASP) is of special interest, since there is experimental evidence of its great importance for viral infectivity in human cells [22, 23].

FIG. 3: Model of one monomer of SARS-CoV-2 spike protein with hotspots in red and hotspots appearing in bibliography in orange. The cartoon representation of the protein is obtained with ChimeraX [3].

A. Hotspot enrichment

The first analysis of these results consists of a hotspot enrichment analysis, which allows us to know whether the hotspots found are more concentrated in particular domains (Fig. 2.c). To carry out the study, a hypergeometric test is performed on the different domains (Table 1 in Appendix).

A domain is considered enriched if its p-value < 0.05. We observe that the flanking positions to the cleavage site region (positions 541-788) are enriched (p-value $= 2.9 \cdot 10^{-23}$). This region corresponds to the part of the protein that fragments when the spike protein binds to the receptor allowing for conformation change for membrane fusion. For the other domains studied, there is no significant enrichment of hotspots.

B. Relative solvent accesibility

In this section, we focus on the study of the RBD. The receptor binding domain (RBD) is a known functional domain, its function is to bind to the angiotensin converting enzyme 2 (ACE2) human receptor, allowing the virus to enter human cells. Even though there is no hotspot enrichment in this domain, a relative solvent accessibility (RSA) analysis can be performed to assess the importance of the hotspots found, given that the relative solvent accessibility (RSA) or relative accessible surface area of an amino acid is a measure of the amino acid solvent exposure. The solvent exposure of an amino acid measures how accessible is the amino acid to the solvent surrounding the protein. If amino acids are bound to a receptor, its relative accessible surface area is expected to be lower.

We calculated the RSA per amino acid of the RBD using the software Freesasa [24]. The RBD was analyzed both in its free conformation and bound to the human receptor ACE2 (Fig 4). When the RBD is bound to ACE2, the positions in the RDB that we determined as hotspots have lower accessible surface area than non-hotspot positions (p-value = 0.03). For the free states, we observe different results for the open and close conformation. The open conformation is when the RBD is up and can bind to the receptor because the amino acids responsible of binding are exposed. In this case, both hotspots and non-hotspots have a high accessible surface area. The closed conformation refers to when the RBD can not bind to the receptor because the amino acids responsible of binding are not exposed. We see that in the closed conformation the positions that we determined as hotspots also have lower accessible surface area than non-hotspot positions (p-value = 0.01).

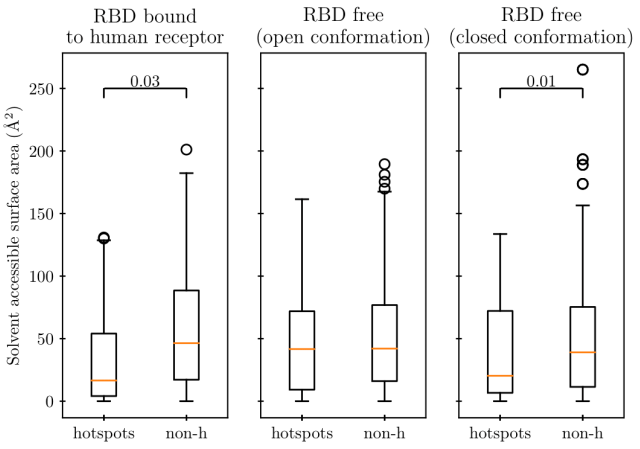

FIG. 4: Solvent accessible surfaces for hotspot and non-hotspot positions when the RBD is bound to the human receptor, the RBD is free with an open conformation and the RBD is free with a closed conformation. The p-values lower than 0.05 indicate significance of the difference between solvent accessible surface area for hotspots versus non-hotspots and are calculated with a Wilcoxon test.

IV. DISCUSSION

Here below I indicate the main results and discuss each of them:

- The method implemented allows finding a set of 275 potential hotspots, some of them confirmed by literature. A confirmed hotspot worth noting is 614 (ASP), given its known importance for SARS-CoV-2 to gain entry to human cells. This is indicative that a structure and sequence-based coevolution detection method could be sensible to key amino acids, but further evaluation is needed to assess its full potential.
- The RSA analysis indicates that the hotspots found in the RBD domain might be important for the binding of the spike protein to the human receptor ACE2. Finding a lower accessible surface area for hotspots than non-hotspots in the case the RBD bound to the receptor suggests that those amino acids are participating in the binding. This significant change suggests that exploring in this direction could lead to interesting results.
- The study done is limited and could be improved in the future by comparing the hotspots obtained with more sources of literature. Additionally, in the absence of literature, methods of energy perturbation by insilico mutation effects could help to detect the individual contribution of each hotspot on protein stability, and thus, its functionality.
- Another option for further research would be to further study the cleavage site region, given the high number of hotspots found in that domain.

Acknowledgments

I would like to thank my supervisors, Dra. Camila Pontes and Victoria Ruiz, for all the advice, guidance, and patience throughout this project.

Also, my gratitude to Joan, friends, and family for the constant support during the whole degree.

[1] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223–230, 7 1973.
[2] O. Bieri and T. Kiefhaber, "Elementary steps in protein folding," *Biological Chemistry*, vol. 380, 1 1999.
[3] G. et al., "Ucsf chimerax: Meeting modern challenges in visualization and analysis," *Protein Science*, vol. 27, no. 1, pp. 14–25, 2018.
[4] D. D. L. et al., "Complete mutagenesis of the hiv-1 protease," *Nature*, vol. 340, pp. 397–400, 8 1989.
[5] C. S. Gibbs and M. J. Zoller, "Identification of functional residues in proteins by charged-to-alanine scanning mutagenesis," *Methods*, vol. 3, pp. 165–173, 12 1991.
[6] C. Baldassi and et al., "Fast and accurate multivariate gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners," *PLoS ONE*, vol. 9, p. e92721, 3 2014.
[7] F. M. et al., "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, 12 2011.
[8] M. W. et al., "Identification of direct residue contacts in protein–protein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 67–72, 1 2009.
[9] J. N. Thompson, "Concepts of coevolution," *Trends in Ecology Evolution*, vol. 4, pp. 179–183, 6 1989.
[10] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, pp. 368–373, 6 2006.
[11] F. P. et al., "Correlated mutations contain information about protein-protein interaction," *Journal of Molecular Biology*, vol. 271, pp. 511–523, 8 1997.
[12] U. G. et al., "Correlated mutations and residue contacts in proteins," *Proteins: Structure, Function, and Genetics*, vol. 18, pp. 309–317, 4 1994.
[13] F. B. et al., "Coevolutionary data-based interaction networks approach highlighting key residues across protein families: The case of the g-protein coupled receptors," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1153–1159, 2020.
[14] D. A. Brian and R. S. Baric, "Coronavirus genome structure and replication," 2005.
[15] F. Li, "Structure, function, and evolution of coronavirus spike proteins.," *Annual review of virology*, vol. 3, pp. 237–261, 2016.
[16] C. P. et al., "Unraveling the molecular basis of host cell receptor usage in sars-cov-2 and other human pathogenic -covs," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 759–766, 2021.
[17] H. M. Berman, "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 1 2000.
[18] J. J. et al., "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, pp. 583–589, 8 2021.
[19] K. et al., "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, pp. 3059–3066, 07 2002.
[20] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, 5 1957.
[21] A. B. et al., "Uniprot," *Nucleic Acids Research*, vol. 49, pp. D480–D489, 1 2021.
[22] B.-F. et al., "Sars-cov-2 viral spike g614 mutation exhibits higher case fatality rate," *International journal of clinical practice*, vol. 74, no. 8, p. e13525, 2020.
[23] J. A. P. et al., "Spike mutation d614g alters sars-cov-2 fitness," *Nature*, vol. 592, pp. 116–121, 4 2021.
[24] S. Mitternacht, "Freesasa: An open source c library for solvent accessible surface area calculations," *F1000Research*, vol. 5, p. 189, 2 2016.

V. APPENDIX

The equation used to perform the hypergeometric test is:

$$p(n, M, N, k) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}} \tag{10}$$

where M is the number of amino acids in the protein, N the number of amino acids in the domain, n the number of hotspots in the whole protein and k the number of hotspots in the particular domain.

TABLE I: Results of the hypergeometric test for different domains of the spike protein. The S1 domain includes positions 1-685, the S2 domain positions 686-1273, the NTD positions 13-305, the RBD positions 319-541, and the cleavage site region positions 541-788 of the spike protein. Some positions are lacking information and have been removed from the analysis.

	n	**M**	**N**	**k**	**p-value**
S1	275	1149	684	172	0.13
S2	275	1149	463	103	0.87
NTD	275	1149	292	44	0.99
RBD	275	1149	222	49	0.79
Cleavage site	275	1149	246	121	2.9e-23