Title: Evolutionary differences in the ACE2 reveals the molecular origins of COVID-19 susceptibility

Authors: Ryan R. Cheng[1*], Esteban Dodero-Rojas[1], Michele Di Pierro[1, 2], José N. Onuchic[1, 3, 4, 5]

Affiliations:

[1] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005.

[2] Department of Physics, Northeastern University, Boston, Massachusetts 02115.

[3] Department of Chemistry, Rice University, Houston, Texas 77005.

[4] Department of Physics & Astronomy, Rice University, Houston, Texas 77005.

[5] Department of Biosciences, Rice University, Houston, Texas 77005.

*To whom the correspondence should be addressed: ryan.r.cheng@gmail.com

Abstract:

We explore the energetic frustration patterns associated with the binding between the SARS-CoV-2 spike protein and the ACE2 receptor protein in a broad selection of animals. Using energy landscape theory and the concept of energy frustration—theoretical tools originally developed to study protein folding—we are able to identify interactions among residues of the spike protein and ACE2 that result in COVID-19 resistance. This allows us to identify whether or not a particular animal is susceptible to COVID-19 from the protein sequence of ACE2 alone. Our analysis predicts a number of experimental observations regarding COVID-19 susceptibility, demonstrating that this feature can be explained, at least partially, on the basis of theoretical means.

Introduction

The coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has affected the lives of millions of people in a worldwide pandemic. The hallmark of COVID-19 is its high degree of contagiousness between individuals.

SARS-CoV-2 is believed to gain entry in to the host cell through its interaction with the Angiotensin-converting enzyme 2 (ACE2) receptor on the host cell surface[1], similar to SARS-CoV-1[2]. Recently, the structure of the SARS-CoV-2 viral spike glycoprotein bound to the human ACE2 receptor was determined using X-ray crystallography [3], providing a crucial starting point for any molecular modeling of the viral interaction with ACE2. The structure of this crucial complex has also been independently determined using cryo-EM[4] and X-ray crystallography[5].

What are the molecular interactions that give rise to interaction specificity between the viral spike and ACE2 receptor? Hints at the molecular origins of COVID-19 susceptibility can be found analyzing the susceptibility of different organisms to the coronavirus. The ACE2 receptors are found in a diverse span of the animal kingdom, including mammals, birds, and aquatic life, which have varying degrees of COVID-19 susceptibility[6-16].

For example, it is known that mice are immune to COVID-19 while on the contrary the Bronx Zoo tiger Nadia[12] had tested positive for COVID-19 and also exhibited many of the symptoms observed in infected humans. To date, a number of animals have been classified as either being susceptible or immune to COVID-19[6-16].

The evolutionary divergence in the sequences of the ACE2 receptor found in different organisms can be related to such susceptibility of infection[17-23]. Here, we explore the molecular mechanisms by which some sequence variants of the ACE2 receptor appear to confer resistance to infection by virtue of their reduced binding affinity to the viral spike protein. We examine the sequences for ACE2 receptor across a selection of 63 representative animals and identify the

residue interactions that are responsible for the reduced binding affinity, and thus COVID-19 resistance, using the concept of energetic frustration from the theory of protein folding [24, 25]. Here, energetic frustration refers to unfavorable interactions between residues in a given protein structure that cannot be mitigated without structural rearrangement or residue level mutations. In the context of the ACE2/spike complex, frustrated interactions between residues of the ACE2 and the spike glycoprotein can also exist for a given structure of the protein complex.

The rarity of kinetics traps observed in the folding of proteins indicates that, in general, proteins do not exhibit a high amount of energetic frustration, which would instead create those kinetic traps[24, 25]. While folding kinetics suggest proteins to be “minimally frustrated”, some local frustration may be present; for example, local frustration could be functionally useful for tuning conformational dynamics. In protein complexes, a site frustrated in the monomeric protein may become less frustrated when the protein is bound to its counterparts, thus guiding specific association[26, 27].

In the case of the complex formed between the SARS-CoV-2 viral spike glycoprotein and the ACE2 receptor, we use changes in energy frustration as a proxy for changes in binding affinity. We use the crystal structure of the viral spike bound to the human ACE2 as a template to construct molecular models of the interaction between the viral spike and the ACE2 receptors of these different animals. We then calculate the changes in frustration with respect to the reference point constituted by the human ACE2 sequence [26, 27]. This allows us to identify key residues of the ACE2 protein that appear to inhibit the binding of the spike glycoprotein and to predict whether or not a particular animal will be susceptible to COVID-19. The novelty of our approach and the key to our results resides in the fact that, while our procedure is based on the only input of the protein sequences of ACE2 receptor, our approach does incorporate a great deal of structural information about the protein complex, which is extracted from the crystal structure[3], and physico-chemical details about the energetics of protein folding and docking, which is synthetized in the energy function and results from decades of developments[24].

Materials & Methods

ACE2 protein sequences

The majority of ACE2 protein sequences were previously annotated from the genome assembles and sequencing data from the DNA Zoo Consortium[28] (Available for download: https://www.dnazoo.org/post/the-first-million-genes-are-the-hardest-to-make-r). The full length protein sequences of the ACE2 proteins for mouse (*Mus musculus*), ferret (*Mustela putorius furo*), chicken (*Gallus gallus domesticus*), pig (*Sus*), duck (*Anas platyrhynchos*), Syrian golden hamster (*Mesocricetus auratus*), and mink (*Neovison vison*) were obtained from the Uniprot database[29] to supplement the sequences derived from the DNA Zoo. In total, 63 representative ACE2 sequences were used in our study (Table S1). For comparative analysis, a multiple sequence alignment was generated for the ACE2 sequences using Clustal Omega[30].

Homology Modeling

The crystal structure of the SARS-Cov-2 glycoprotein spike bound to the human ACE2 protein served as our starting template for constructing models of the glycoprotein spike bound to the ACE2 protein of other animals. We used the SWISS-MODEL [31] to create homology models of 63 representative animals (Table S1) for a full list.

Frustration Analysis

We performed an energy landscape analysis on the predicted ACE2-spike complex for different animals using the configurational frustration index[26, 27]:

$$F_{ij} = \frac{\left(H_{ij} - \left\langle H_{i'j'}^{decoy} \right\rangle\right)}{\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(H_{i'j'} - \left\langle H_{i'j'}^{decoy} \right\rangle\right)}} \quad (1)$$

Here, H_{ij} represents the pairwise interaction energy between residues i and j in a given structure using the Associative Memory, Water Mediated, Structure and Energy Model (AWSEM)[32], a

coarse-grained model widely used to study problems of protein folding and protein-protein association and assembly. The native energies H_{ij} are compared directly to N number of different configurational realizations between residues i and j, thereby generating a distribution of decoy energies with a mean of $\left\langle H_{i'j'}^{decoy} \right\rangle$ and a standard deviation of $\sqrt{N^{-1}\sum_{k=1}^{N}\left(H_{i'j'} - \left\langle H_{i'j'}^{decoy} \right\rangle\right)}$. Hence, F_{ij} is a type of Z-score that measures how favorable a particular pair of interactions are within a protein or protein complex with respect to a distribution of decoys. Frustrated (unfavorable interactions) are denoted by $F_{ij} < 0$ while $F_{ij} > 0$ are considered favorable; in particular, $F_{ij} < -1$ is considered highly frustrated while $F_{ij} > 1$ is considered minimally frustrated.

In our analysis, we found that it was useful to compare the configurational frustration between an interprotein residue pair with the same pair from the human ACE2-spike complex:

$$\Delta F_{ij}^{(Species)} = F_{ij}^{(Species)} - F_{ij}^{(Human)} \qquad (2)$$

We find that $\Delta F_{ij}^{(Species)} < -1.5$ robustly identifies highly frustrated interactions that result in COVID-19 resistence. On the other hand, if all of the inter-protein residue interactions between the ACE2 receptor and the spike do not exhibit high levels of frustration (i.e., $\Delta F_{ij}^{(Species)} > -1$) we identify that species as being highly susceptible to COVID-19. For completeness, if the most frustrated interprotein interactions fall between $-1.5 < \Delta F_{ij}^{(Species)} \leq -1$ that species is predicted to be moderately susceptible.

Evolutionary distance between ACE2 proteins

The Jukes-Cantor distance is used to quantify the evolutionary distance between aligned ACE2 proteins in our study: $d = -\frac{19}{20}\log\left(1 - \frac{20}{19}p\right)$, where p is the p-distance—i.e., the number of residue sites between two compared sequences that are different divided by the sequence length of the multiple sequence alignment.

Results & Discussion

Comparative frustration analysis of ACE2-spike complex for different species

By examining the comparative differences between the inter-protein interactions with respect to the human ACE2-spike complex, we are able to identify residue-interactions outliers that represent a significant disruption to the ACE2/spike interaction relative to the human ACE2-spike interaction.

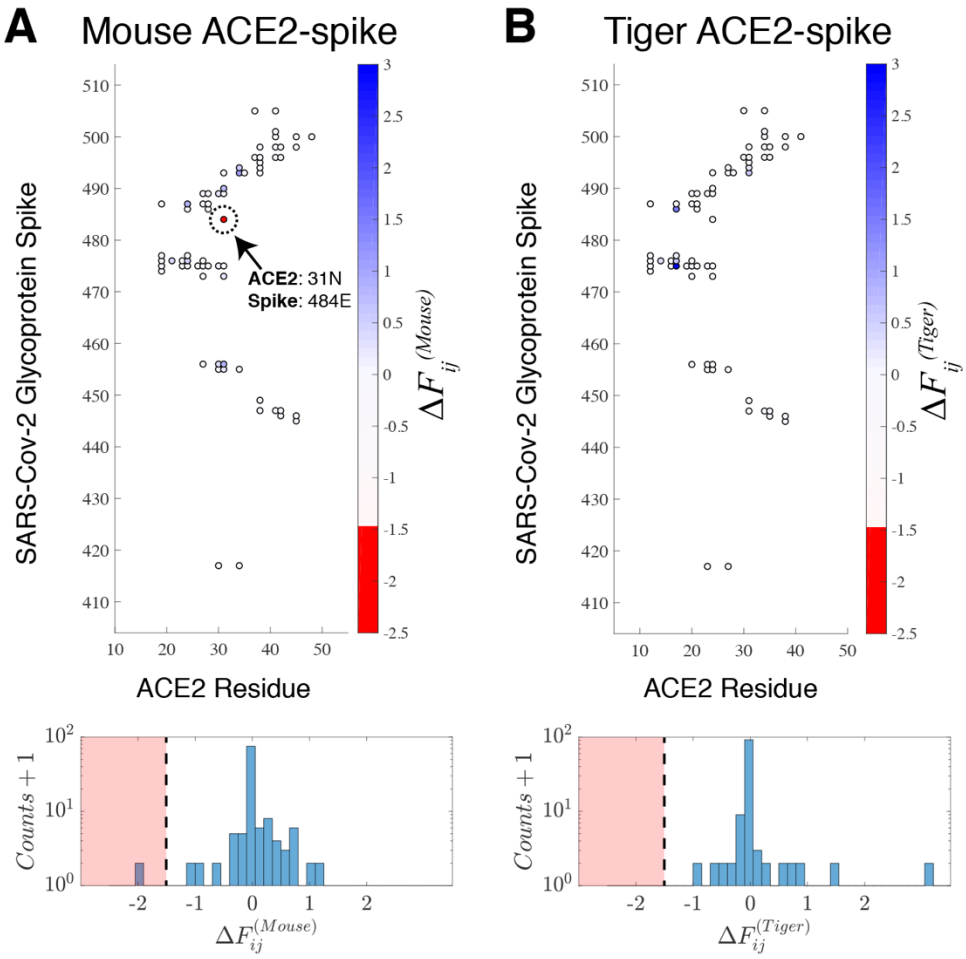
Shown in Figure 1 are plots of our frustration analysis for mouse (*Mus musculus*) and tiger (*Panthera tigris*), which are used as representative examples of animals have been experimentally observed to be resistant[14] and susceptible[12] to COVID-19, respectively. We observe a single frustrated residue pair between 31N of the mouse ACE2 and 484E of the spike protein on the map of configurational frustration (Figure 2A/2C), which appears as an outlier in the histogram of $\Delta F_{ij}^{(Mouse)} \sim -2$. Similar highly frustrated outliers can be found for the other animals that are known to resist COVID-19 (Figure S1), such as chicken (*Gallus gallus domesticus*)[6] and duck (*Anas*)[6]. We find that a threshold value of $\Delta F_{ij}^{(Species)} < -1.5$ robustly identifies residue pairs that appear to confer COVID-19 resistence. Likewise, a histogram of $\Delta F_{ij}^{(Tiger)}$ exhibits comparable levels of frustration to that of the human ACE2 and spike (Figure 1B)—similar findings are obtained for other animals with known susceptibilities to COVID-19 (Figure S2), such as white-tailed deer (*Odocoileus virginianus*)[15], European rabbit (*Oryctolagus cuniculus*)[7], and pig (*Sus scrofa*)[13].

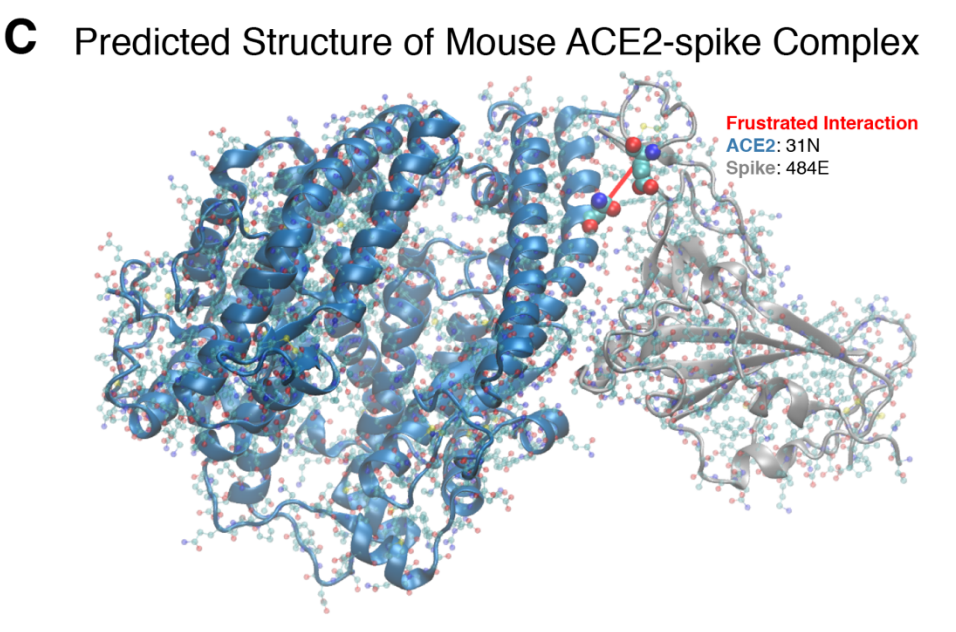
We further apply this analysis for identifying frustrated outliers in our other modeled complexes, thereby predicting whether a particular animal is susceptible to COVID-19. A detailed summary of our results is shown in Figure 2, which includes experimental observations that corroborate or are inconsistent with our predictions. Other animals that have been experimentally observed to be susceptible to COVID-19, such as mink (*Neovison vison*) [11], and Syrian golden hamster (*Mesocricetus auratus*)[9], are identified as being moderately susceptible by our computational approach. Coronavirus consensus PCR-primer sequences have been detected with high frequency in populations of straw-colored fruit bats (*Eidolon helvum*)[16], which have been predicted to exhibit moderate susceptibility by the computational approach.

Taken together, our findings summarized in Figure 2 show that an energy landscape-based approach can identify the molecular origins of COVID-19 susceptibility and resistance. Experimental observations corroborate 10 out of 12 of the computational predictions. One apparent inconsistency between the predictions and experimental observation regards the susceptibility of ferrets (*Mustela putorius furo*), which have been observed to replicate SARS-Cov-2 specifically in their upper respiratory tract[6]. Our analysis of the ferret ACE2-spike complex reveals a single highly frustrated inter-protein interaction between 34Y of ACE2 and 403R of the spike protein. However, it has been noted[6, 19] that ferrets have a unique respiratory biology, which may offer an explanation for this apparent discrepancy. Another apparent inconsistency is observed with our predictions for dogs (*Canis lupus familiaris*). Our frustration analysis predicts two highly frustrated inter-protein contacts within the ACE2/spike complex: 33Y of ACE2 with 417K of the spike and 325E of the ACE2 with 502G of the spike. Yet the susceptibility of dogs still remains somewhat controversial—while viral susceptibility and the production of antibody responses have been detected in dogs[8], viral replication has been reported to be poor[6].

Our energy landscape-based predictions for COVID-19 susceptibility are closely related to similar approaches that examined sequence differences in ACE2 sequences of different animals in the context of a structural model of the ACE2/spike complex[18-20]. Frustration analysis yields a benefit to computational estimates of binding affinity because it compares the interaction energies between ACE2 and the spike glycoprotein with respect to alternative configurations (i.e., decoys) to assess how favorable a particular interaction is in the binding interface. In particular, the majority of our predictions are consistent with those of Damas et al[19], which makes predictions that are consistent with the same 10 out of 12 experiment observations that are highlighted in Figure 2. However, validation of the different models that exists is limited by the relatively small number of confirmed cases of COVID-19 in animals.

By in large, we find that our simple model appears to be consistent with many experimental observations of COVID-19 infections across different animals despite only considering the ACE2-spike protein interaction.





200

201

Figure 1. Comparative analysis of the frustration indices with respect to those observed in human ACE2-spike complex reveals outliers that indicate COVID-19 resistance. The configurational frustration index relative to the frustration in the human ACE2-spike complex is shown for (A) mouse (*Mus musculus)* and (B) tiger (*Panthera tigris*) on a contact map illustrating select contacts between the SARS-Cov-2 spike and the ACE2 protein. Corresponding histograms of the frustration index between all contacts between the ACE2 and spike protein are also shown. (A) and (B) are representative examples of animals that are resistant to COVID-19 and susceptible to COVID-19, respectively. Animals that resist COVID-19 appear to have frustrated outliers that represent highly unfavorable residue interactions compared to the human ACE2-spike complex (i.e., $\Delta F_{ij}^{(Species)} < -1.5$). For the mouse, a single frustrated interaction between residue 31N of the ACE2 protein and 484E of the spike glycoprotein appears to confer COVID-19 resistance. (C) The frustrated interaction is plotted on the modeled 3D structure of the spike glycoprotein bound to the mouse ACE2 receptor.

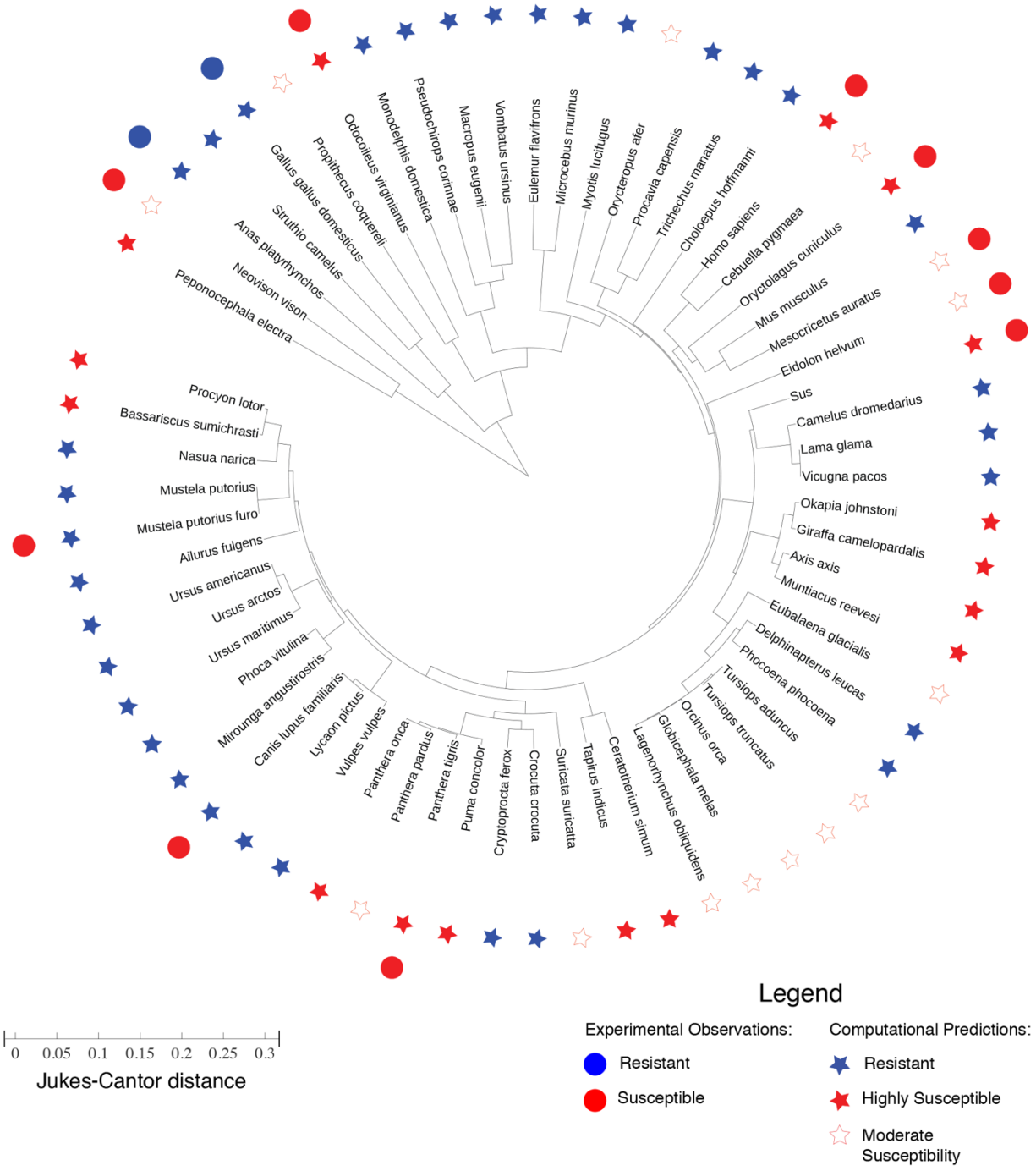


Figure 2.

Phylogenic tree representing the evolutionary distance between ACE2 proteins of different species. The lengths in the radial direction denote the Jukes-Cantor distance (See Materials & Methods) as a measure of evolutionary distance between any two ACE2 proteins. The experimental observation of SARS-CoV-2 resistance/susceptibility are plotted alongside the computational predictions for resistance/susceptibility based on our frustration analysis of the

ACE2-spike complex. See the Legend for more details. There is a consistency between the computational predictions and the experimental observations for mouse (*Mus musculus*), chicken (*Gallus gallus domesticus*)[6], duck (*Anas platyrhynchos*)[6], mink (*Neovison vison*)[11], bat (*Eidolon helvum*)[16], Syrian golden hamster (*Mesocricetus auratus*)[9], tiger (*Panthera tigris*), white-tailed deer (*Odocoileus virginianus*)[15], European rabbit (*Oryctolagus cuniculus*)[7], and pig (*Sus scrofa*)[13]. However, apparent inconsistencies are found for ferret (*Mustela putorius furo*)[6] and dog (*Canis lupis familiaris*) [6, 8]—however, SARS-Cov-2 has only been observed to replicate in the upper respiratory tract of ferrets[6], and viral replication has been observed to be low in dogs[6].

Conclusion

The COVID-19 pandemic and the spread of other coronaviruses in recent years requires an indirect approach to understanding the molecular determinants behind susceptibility and resistance. Here, we constructed structural models of the ACE2-spike glycoprotein complex for a wide range of animals with ACE2 receptors. Using an energy landscape theory-based analysis we are able to uncover specific inter-protein interactions between the ACE2 and spike that appear to confer COVID-19 resistance. Our predictions appear to be consistent with many of the experimental observations regarding animal susceptibility, providing a structural explanation to those observations.

Our analysis reveals that the evolutionary distance between ACE2 proteins is not sufficient to predict COVID-19 susceptibility (Figure 2). Rather, an energy landscape-based analysis appears necessary to assess the interactions between the ACE2 protein and the SARS-Cov-2 spike glycoprotein.

Acknowledgements

The authors would like to thank Matthew MacManes, Erez Aiden, and Olga Dudchenko for helpful discussions and providing data and resources through the DNA Zoo. This work was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation NSF Grant PHY-2019745. J.N.O. was also supported by the NSF-CHE-1614101 and by the Welch Foundation (Grant C-1792). J.N.O. is a Cancer Prevention and Research Institute of Texas. Scholar in Cancer Research.

1. Zhou, P., X.-L. Yang, et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-273.
2. Li, W., M.J. Moore, et al., *Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus.* Nature, 2003. **426**(6965): p. 450-454.
3. Wang, Q., Y. Zhang, et al., *Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2.* Cell, 2020. **181**(4): p. 894-904.e9.
4. Yan, R., Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, *Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2.* Science, 2020. **367**(6485): p. 1444.
5. Shang, J., G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, and F. Li, *Structural basis of receptor recognition by SARS-CoV-2.* Nature, 2020. **581**(7807): p. 221-224.
6. Shi, J., Z. Wen, et al., *Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2.* Science, 2020. **368**(6494): p. 1016.
7. Mykytyn, A.Z., M.M. Lamers, et al., *Susceptibility of rabbits to SARS-CoV-2.* bioRxiv, 2020: p. 2020.08.27.263988.
8. Sit, T.H.C., C.J. Brackman, et al., *Infection of dogs with SARS-CoV-2.* Nature, 2020. **586**(7831): p. 776-778.
9. Sia, S.F., L.-M. Yan, et al., *Pathogenesis and transmission of SARS-CoV-2 in golden hamsters.* Nature, 2020. **583**(7818): p. 834-838.
10. Gautam, A., K. Kaphle, B. Shrestha, and S. Phuyal, *Susceptibility to SARS, MERS, and COVID-19 from animal health perspective.* Open Vet J, 2020. **10**(2): p. 164-177.
11. Oude Munnink, B.B., R.S. Sikkema, et al., *Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans.* Science, 2021. **371**(6525): p. 172.
12. Goldstein, J., *Bronx Zoo Tiger Is Sick With the Coronavirus*, in *New York Times*. 2020.
13. BS, P., S. G, P. MM, E.-H. C, M. E, M. P, and L. CE, *Susceptibility of Domestic Swine to Experimental Infection with Severe Acute Respiratory Syndrome Coronavirus 2.* Emerg Infect Dis., 2021. **27**(1): p. 104-112.
14. Muñoz-Fontela, C., W.E. Dowling, et al., *Animal models for COVID-19.* Nature, 2020. **586**(7830): p. 509-515.
15. Palmer, M.V., M. Martins, et al., *Susceptibility of white-tailed deer (Odocoileus virginianus) to SARS-CoV-2.* bioRxiv, 2021: p. 2021.01.13.426628.

16. Kumakamba, C., F.R. Niama, et al., *Coronavirus surveillance in Congo basin wildlife detects RNA of multiple species circulating in bats and rodents.* bioRxiv, 2020: p. 2020.07.20.211664.
17. Martínez-Hernández, F., A.B. Isaak-Delgado, J.A. Alfonso-Toledo, C.I. Muñoz-García, G. Villalobos, N. Aréchiga-Ceballos, and E. Rendón-Franco, *Assessing the SARS-CoV-2 threat to wildlife: Potential risk to a broad range of mammals.* Perspectives in Ecology and Conservation, 2020.
18. Lam, S.D., N. Bordin, et al., *SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals.* Scientific Reports, 2020. **10**(1): p. 16471.
19. Damas, J., G.M. Hughes, et al., *Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates.* Proceedings of the National Academy of Sciences, 2020. **117**(36): p. 22311.
20. Luan, J., Y. Lu, X. Jin, and L. Zhang, *Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection.* Biochemical and Biophysical Research Communications, 2020. **526**(1): p. 165-169.
21. Becker, D.J., G.F. Albery, et al., *Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization: a modeling study.* bioRxiv, 2020: p. 2020.05.22.111344.
22. Melin, A.D., M.C. Janiak, F. Marrone, P.S. Arora, and J.P. Higham, *Comparative ACE2 variation and primate COVID-19 risk.* Communications Biology, 2020. **3**(1): p. 641.
23. Frank, H.K., D. Enard, and S.D. Boyd, *Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host receptor in bats compared to other mammals.* bioRxiv, 2020: p. 2020.04.20.051656.
24. Onuchic, J.N. and P.G. Wolynes, *Theory of protein folding.* Current Opinion in Structural Biology, 2004. **14**(1): p. 70-75.
25. Onuchic, J.N., Z. Luthey-Schulten, and P.G. Wolynes, *THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective.* Annual Review of Physical Chemistry, 1997. **48**(1): p. 545-600.
26. Ferreiro, D.U., J.A. Hegler, E.A. Komives, and P.G. Wolynes, *Localizing frustration in native proteins and protein assemblies.* Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19819.
27. Parra, R.G., N.P. Schafer, L.G. Radusky, M.-Y. Tsai, A.B. Guzovsky, P.G. Wolynes, and D.U. Ferreiro, *Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics.* Nucleic Acids Research, 2016. **44**(W1): p. W356-W360.
28. Dudchenko, O., S.S. Batra, et al., *De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.* Science, 2017. **356**(6333): p. 92.
29. The UniProt, C., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2021. **49**(D1): p. D480-D489.
30. Madeira, F., Y.M. Park, et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic acids research, 2019. **47**(W1): p. W636-W641.
31. Waterhouse, A., M. Bertoni, et al., *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic Acids Research, 2018. **46**(W1): p. W296-W303.
32. Davtyan, A., N.P. Schafer, W. Zheng, C. Clementi, P.G. Wolynes, and G.A. Papoian, *AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials*

and Bioinformatically Based Local Structure Biasing. The Journal of Physical Chemistry B, 2012. **116**(29): p. 8494-8503.