Cell Systems

Transparent Peer Review Record

A sweep of earth's virome reveals host-guided viral protein structural mimicry and points to determinants of human disease

Gorka Lasso, Barry Honig, Sagi D. Shapira

Summary

Initial Submission: Received Jul. 24, 2020
Preprint: https://doi.org/10.1101/2020.06.18.159467
Deposited on bioRxiv, Jun. 18, 2020
Scientific editor: Quincey Justman, Ph.D.

First round of review: Number of reviewers: Three
Three confidential, zero signed
Accepted Sep. 18, 2020

Data freely available: Yes
Code freely available: Yes

This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Editorial decision letter with reviewers' comments, first round of review

Dear Sagi,

As I mentioned in my email last week, the reviews of your manuscript are back. I'm very pleased to let you know on the strength fo them, the formal peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication. Congratulations! When you

submit your final manuscript, please do include a point-by-point response to the reviewers comments, because I may need to consult with them informally and ask for minor tweaks (I'll let you know if so). Note that we'll be publishing this paper as a Report (up to 4 Figures).

Below, you'll find the comments from the reviewers. I've also made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager. ***We hope to receive your files within 10 business days, but we recognize that the COVID-19 pandemic may challenge and limit what you can do. Please email me directly if this timing is a problem or you're facing extenuating circumstances.***

I'm looking forward to going through these last steps with you. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,
Quincey

Quincey Justman, Ph.D.
Editor-in-Chief, Cell Systems

Editorial Notes

Title: Your title is excellent, although I'd avoid the semi-colon if possible. I always ask authors to make sure their titles are as effective as they can be. Note that an effective title is easily found on Pubmed and Google. A trick for thinking about titles is this: ask yourself, "How would I structure a Pubmed search to find this paper?" Put that search together and see whether it comes up is good "sister literature" for this work. If it does, feature the search terms in your title.

Abstract: Our abstracts have a 150 word limit. I've edited your abstract down to size. See what you think and please feel free to make changes.

> Viruses deploy an array of genetically encoded strategies to coopt host machinery and support viral replicative cycles. Here, we use protein structure similarity to scan for molecular mimicry, manifested by structural similarity between viral and endogenous host proteins, across thousands of catalogued viruses and hosts spanning broad ecological niches and taxonomic range, including bacteria, plants and fungi, invertebrates and vertebrates. This survey identified over 6,000,000 instances of structural mimicry; more than 70% of viral mimics cannot be discerned through protein sequence alone. We demonstrate that the manner and degree to which viruses exploit molecular mimicry varies by genome size and nucleic acid type and identify 145 human ***[correct?]*** proteins that are mimicked by coronaviruses. Our observations point to molecular mimicry as a pervasive strategy employed by viruses and indicate that the protein structure space used by a given virus is

dictated by the host proteome. A record of this paper's Transparent Peer Review process is included in the Supplemental Information.

Manuscript Text: Your text is excellent! Please leave it essentially as it is and address reviewer concerns surgically. If it's technically "too long," don't worry, that's fine. Note two minor points about "house style" when preparing the final version of your text:

- House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.), especially the Results section. These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. "Notably" is suitably neutral to use once or twice if absolutely necessary.
- We don't allow "priority claims" (e.g. new, novel, etc.). For a discussion of why, read: http://crosstalk.cell.com/blog/getting-priorities-right-with-novelty-claims, http://crosstalk.cell.com/blog/novel-insights-into-priority-claims. In this context, please use "previously unidentified" (or something similar).

Figures and Legends: Your figures are excellent, but please indicate how large each group is, either within the figure (if it's legible and not too cluttered) or within the legend. Visually, it's too easy to "round this study down" to something much smaller than it is. Clarity and distillation is important in figures, but please don't do this at the expense of making your study seem less comprehensive than it is! Also, note that each of your figures can take up an entire page.

Thank you!

Reviewer comments:

Reviewer #1: Lasso et al. identified potential viral mimics (not potential interactors) in human. Their findings about (i) viruses utilizing structural mimicry varies by genome size and nucleic acid type and (ii) viruses with smaller genomes - like ssRNA viruses - mimicking human proteins to a greater extent than their dsDNA counterparts are very intriguing. It is also an interesting finding that while virus-host PPIs has enabled convergence of evolutionary unrelated viruses around key pathways, the structurally mimicked proteins show little convergence.
Overall, this is a timely contribution.

A few comments:

- Since they analyzed all coronaviruses, what are the differences they observed between SARS-CoV-2 and other CoV members? Do the differences suggest any hypothesis on the molecular basis of high contagiousness of SARS-CoV-2 compared to SARS-CoV or other beta-CoV members?

- They should mention the shortcomings of their approach in the Discussion. For instance, their approach starts with sequence alignment, but some viral/bacterial proteins have unique sequences such that they

do not have any sequence homology to any other protein (protein from human or any other organism).

- It would be helpful if they clarify why they calculate "human protein mimics per residue". It would be good to compare Figure 2c with a new figure where they show the number of mimics per virus, not per residue.

- Please provide the sequence identities of the examples in Figure S2.

- The examples they provide in Figure 3 are compelling. Do they observe any differences between CoV and SARS-CoV-2 in terms of the predicted mimics provided in this figure?

Reviewer #2: In the manuscript, "A sweep of earth's virome reveals host-guided viral protein structural mimicry with implications for human disease" from Lasso et al., the structural similarity of known 3-dimensional protein structures to viral proteins was investigated using global structural alignment and similarity scoring. Essentially, the authors determine host protein structural neighbors of viral proteins. A highlight of this manuscript is the comprehensive set of viral proteins (~337,000 from 7,486 viruses) across diverse families and host taxonomic ranges that were included in this analysis. These comparisons produced ~6 million instances of structural mimicry. The scope of the analysis is only limited by the differences in protein structure availability across taxonomic divisions.
The authors made several observations that broadly enrich our understanding of host-pathogen co-evolution as well as specific discoveries of previously unrecognized host-pathogen structural relationships. For example, a striking finding was that > 70% of mimicry owes to structural features, and by analogy, would not be present in primary sequence features (mimics typically had < 20% sequence identity). The authors confirmed previously known functional mimics with their approach, while expanding structural mimics across several viruses. For instance, 61 herpesvirus structural mimics were assigned that would otherwise be missed if evaluated by sequence identity alone.
Additionally, the authors showed that genome size and nucleic acid type are key regulators of the mimicry phenotype. For example, dsDNA viruses had a larger number of mimics converging on specific biological pathways, such as innate immune sensing, while ssRNA viruses are more efficient mimickers based on their limited genome size. The authors also had a specific focus on coronavirus proteins, highlighting non-structural proteins, such as NSP13 and NSP3, that mimic regulators of immune anti-viral programs, e.g. helicase proteins and proteins involved in STAT1-mediated signaling, respectively. Also, these data identified mimics that could contribute to coagulation-associated disorders commonly associated with several viruses, which is clinically relevant, as previous evidence has suggested that a history of macular degeneration, a complement-associated syndrome, is a significant risk factor for morbidity and mortality in SARS-CoV-2 infected patients (independent of age).

The manuscript presents a thought-provoking perspective on the different selective pressures of viruses and their hosts, and how these pressures relate to genome size and virus (Baltimore) classification. Moreover, the analysis of mimics across host species has the potential to identify pathogenic proteins with mimicry across closely related species, and thus may more readily facilitate zoonotic infections.
The one drawback of the study, which the authors acknowledged, was that it remains to be shown whether the newly classified structural mimics are bona fide functional mimics. Overall, the contribution of

this manuscript is significant given the current SARS-CoV-2 pandemic. The strong evidence that zoonotic transfer of coronaviruses (SARS, MERS, SARS-2) is the primary source of human pathogenesis highlights the importance of understanding the structural conservation and mimicry potential of virus-host protein pairs.
The authors could address a few points detailed below (either in the Results or the Discussion), which would give the manuscript additional insights and context.

1. The authors discussed mimicry among virus classes, mostly as a function of whether convergence was observed across biological processes (fig 2d). Do the authors structural analyses allow conclusions to be drawn on whether a small number of specific host proteins are "highly mimicked"? Either across all viruses or only within specific virus families for a specific host?
2. In Fig 3, was it the authors intention to only high non-structural proteins? Is the absence of structural protein biologically significant or a technical issue?
3. In the Discussion, could the authors comment on whether there are specific examples of new structural mimics found that would guide selection of antigens in vaccine development? Or more broadly, is there any indication that the observed common or unique features of virus-host structural pairs could guide therapeutic strategies?

Minor points
1. Issues with sentence completion or structure in lines 89-90, 249-50, 290 - 292.

Reviewer #3: In this paper, the authors perform a structure-based comparison between viral proteins and host proteins at a very large scale. Their analysis covers >300,000 viral proteins encoded by >7,000 viruses belonging to a broad taxonomic range. The hosts infected by the studied viruses belong to one of the following broad taxa: bacteria, plant and fungi, vertebrates, invertebrates, and human. Based on their approach, the authors find that structural similarity between viral proteins and host proteins is enriched among the infected host taxon compared to other taxa that are not infected by the virus. They conclude that the structure space mimicked by viruses is guided by the structure space of the infected host proteome.

Several studies have shown before that viral proteomes tend to mimic specific host proteomes in their structural properties as a means to hijack cellular processes and pathways. This mimicry is more evident at the structural domain level and protein interaction interface level. The current study is the first attempt to detect virus-host structural mimicry at a large scale covering a broad taxonomic range. The analysis and results are therefore of broad interest. In addition, the reported case studies relating to disease are also relevant and timely.

The authors determine structural mimicry using a two-step approach. In brief, the authors first select a structural template for the viral protein based on sequence similarity. This template may be selected from any species. Then, the authors perform a structural similarity search to identify other structures in PDB that have high structural similarity with the selected viral protein template. They find that these highly similar structural neighbors are enriched among the host which is known to be infected by the virus.

A minority (<30%) of structural mimicry cases can be detected through sequence similarity alone. Are these highly similar sequence neighbors also enriched among the host which is known to be infected by the virus? These highly similar sequence neighbors are interesting because they are clearly related to the viral protein via divergent evolution.

The authors correctly point out that in general, it is difficult to determine whether structural mimicry (with no detectable sequence similarity) is the result of divergent evolution or convergent evolution. However, given the global nature of structural similarity between viral and host proteins detected in this study, it seems less likely that through convergent evolution, viral and host proteins will independently evolve tertiary structures that are highly similar to each other in a global way. Rather, it seems more likely that through convergent evolution, viral and host proteins will independently evolve different tertiary structures to carry out similar functions, e.g., to bind the same interface in the host molecular circuitry ("interface mimicry").

One caveat of the study is that while the survey identified >6 million instances of structural mimicry, it is difficult to know what fraction of these instances are true positives. For every true positive case of structural mimicry between viral and host proteins, there could be numerous false positive cases of other proteins from the same host or proteins from other hosts that have nothing to do with the virus but are incorrectly detected because they share similar tertiary structures with the true positive host protein. The authors are encouraged to comment on this caveat.

I also recommend that the authors review the text to remove errors in sentence structure. For example, on Page 4 Line 89 there is an extra unneeded phrase "Ska was used". Page 11 Line 348, < 1x10^{-6} should be > 1x10^{-6}.

Authors' response to the reviewers' first round comments

Attached.

Reviewer comments:

Reviewer #1: Lasso et al. identified potential viral mimics (not potential interactors) in human. Their findings about (i) viruses utilizing structural mimicry varies by genome size and nucleic acid type and (ii) viruses with smaller genomes - like ssRNA viruses - mimicking human proteins to a greater extent than their dsDNA counterparts are very intriguing. It is also an interesting finding that while virus-host PPIs has enabled convergence of evolutionary unrelated viruses around key pathways, the structurally mimicked proteins show little convergence. Overall, this is a timely contribution.

A few comments:

- Since they analyzed all coronaviruses, what are the differences they observed between SARS-CoV-2 and other CoV members? Do the differences suggest any hypothesis on the molecular basis of high contagiousness of SARS-CoV-2 compared to SARS-CoV or other beta-CoV members?

The referee raises a very interesting and important question. It is tempting to suggest that the precise constellation of proteins mimicked by different CoV's may inform about differences in their pathophysiology. However, we have not identified structural mimics that differentiate SARS-CoV-2 (or any other CoV) from all other CoVs. In fact, the majority (96%) of CoV structural mimics are shared by three or more CoVs. So, it is likely that it is conservation or divergence at key amino-acid residues that determine the functional consequences mimicry. Future analytical and experimental interrogation will be necessary to identify molecular features of virus-encoded mimics which impact disease. We have added a note to this effect in the expanded discussion.

- They should mention the shortcomings of their approach in the Discussion. For instance, their approach starts with sequence alignment, but some viral/bacterial proteins have unique sequences such that they do not have any sequence homology to any other protein (protein from human or any other organism).

The reviewer is absolutely correct in pointing out that a limitation of our approach is identification of sequence relationships (albeit distant ones) between query proteins (virus) and the PDB (which we use to capture structural information). Indeed, as the breadth and scope of PDB increases (with more experimentally solved protein structures), so will the ability to leverage computational approaches to resolve structural information of novel proteins as well as capture pathogen encoded structure mimics. We have added a note to this effect in the expanded discussion.

- It would be helpful if they clarify why they calculate "human protein mimics per residue". It would be good to compare Figure 2c with a new figure where they show the number of mimics per virus, not per residue.

Following the referee's suggestion, we have clarified the reasoning behind calculating the number of human protein mimics per residue: please see amended text in lines 146-152.

- Please provide the sequence identities of the examples in Figure S2.

We thank the referee for pointing out this major oversight. We have corrected Figure S2 to include pairwise sequence identities.

- The examples they provide in Figure 3 are compelling. Do they observe any differences between CoV and SARS-CoV-2 in terms of the predicted mimics provided in this figure?

Again, the referee raises an important question. Please see our response to the first comment.

Reviewer #2: In the manuscript, "A sweep of earth's virome reveals host-guided viral protein structural mimicry with implications for human disease" from Lasso et al., the structural similarity of known 3-dimensional protein structures to viral proteins was investigated using global structural alignment and similarity scoring. Essentially, the authors determine host protein structural neighbors of viral proteins. A highlight of this manuscript is the comprehensive set of viral proteins (~337,000 from 7,486 viruses) across diverse families and host taxonomic ranges that were included in this analysis. These comparisons produced ~6 million instances of structural mimicry. The scope of the analysis is only limited by the differences in protein structure availability across taxonomic divisions.

The authors made several observations that broadly enrich our understanding of host-pathogen co-evolution as well as specific discoveries of previously unrecognized host-pathogen structural relationships. For example, a striking finding was that > 70% of mimicry owes to structural features, and by analogy, would not be present in primary sequence features (mimics typically had < 20% sequence identity). The authors confirmed previously known functional mimics with their approach, while expanding structural mimics across several viruses. For instance, 61 herpesvirus structural mimics were assigned that would otherwise be missed if evaluated by sequence identity alone.

We thank the referee for the kind reading of our manuscript. Indeed, structural similarity offers an opportunity to identify evolutionary relationships. While our method is limited by the structural coverage of viral and host proteins, as the referee points out, to look beyond sequence space by taking advantage of the fact that structure is better conserved than sequence. Of course, since our method relies on the existence of experimentally resolved crystal structures, as the number and diversity of these increases, so will the scope of our approach.

Additionally, the authors showed that genome size and nucleic acid type are key regulators of the mimicry phenotype. For example, dsDNA viruses had a larger number of mimics converging on specific biological pathways, such as innate immune sensing, while ssRNA viruses are more efficient mimickers based on their limited genome size. The authors also had a specific focus on coronavirus proteins, highlighting non-structural proteins, such as NSP13 and NSP3, that mimic regulators of immune anti-viral programs, e.g. helicase proteins and proteins involved in STAT1-mediated signaling, respectively. Also, these data identified mimics that could contribute to coagulation-associated disorders commonly associated with several viruses, which is clinically relevant, as previous evidence has suggested that a history of macular degeneration, a complement-associated syndrome, is a significant risk factor for morbidity and mortality in SARS-CoV-2 infected patients (independent of age).

As the referee suggests, the set of structural mimics reported in this work offers a valuable resource to functionally interrogate viral proteins. While structural similarity between viral and host proteins serves as a scaffold enabling viral proteins to coopt host pathways, amino acid identities at key functional sites will ultimately shape the extent to which host pathways can be intervened. So, while the biological implications for the structural mimics we identify are yet to be determined, the results enable formulation of experimentally testable hypotheses that focus on interrogating individual protein residues.

The manuscript presents a thought-provoking perspective on the different selective pressures of viruses and their hosts, and how these pressures relate to genome size and virus (Baltimore) classification. Moreover,

the analysis of mimics across host species has the potential to identify pathogenic proteins with mimicry across closely related species, and thus may more readily facilitate zoonotic infections.

The one drawback of the study, which the authors acknowledged, was that it remains to be shown whether the newly classified structural mimics are bona fide functional mimics. Overall, the contribution of this manuscript is significant given the current SARS-CoV-2 pandemic. The strong evidence that zoonotic transfer of coronaviruses (SARS, MERS, SARS-2) is the primary source of human pathogenesis highlights the importance of understanding the structural conservation and mimicry potential of virus-host protein pairs.

The authors could address a few points detailed below (either in the Results or the Discussion), which would give the manuscript additional insights and context.

1. The authors discussed mimicry among virus classes, mostly as a function of whether convergence was observed across biological processes (fig 2d). Do the authors structural analyses allow conclusions to be drawn on whether a small number of specific host proteins are "highly mimicked"? Either across all viruses or only within specific virus families for a specific host?

We thank the referee for the question. We have amended the text as follows in an effort to address the question: "*While we do not observe recursive structural mimicry of human proteins across the human virome, we find that viruses belonging to the same viral family (including astroviruses, coronaviruses, filoviruses, orthomyxoviruses, paramyxoviruses, rhabdoviruses and picornaviruses; data available for download at https://github.com/gorkaLasso/VirusHost_mimicry) tend to share a large fraction of mimics (where >75% of viruses within the family share >75% structural mimics).*"

2. In Fig 3, was it the authors intention to only high non-structural proteins? Is the absence of structural protein biologically significant or a technical issue?

As noted by the referee, Figure 3 does not list all viral proteins within *Coronaviridae*. This reflects our inability to identify structural templates for the missing viral proteins – impeding us from inferring structural mimics. The missing viral proteins are part of the "dark" viral proteome for which no protein structure has been solved experimentally or can be modelled by homology (at least though a conservative modeling approach).

3. In the Discussion, could the authors comment on whether there are specific examples of new structural mimics found that would guide selection of antigens in vaccine development? Or more broadly, is there any indication that the observed common or unique features of virus-host structural pairs could guide therapeutic strategies?

This is a great point that certainly needs further exploration. It's reasonable to speculate that structural mimicry may influence immunogenicity of viral proteins and by proxy vaccine design (since both affinity and avidity determine protein interactions and antibody maturation). Speculation aside though, in a recent publication we used knowledge of CoV encoded mimics to guide the identification of disease determinants. We have referenced our findings and highlight their implications in the context of viral mimicry.

Minor points

1. Issues with sentence completion or structure in lines 89-90, 249-50, 290 - 292.

   We thank the referee for their careful reading for the manuscript. We have amended the text accordingly.

Reviewer #3: In this paper, the authors perform a structure-based comparison between viral proteins and host proteins at a very large scale. Their analysis covers >300,000 viral proteins encoded by >7,000 viruses belonging to a broad taxonomic range. The hosts infected by the studied viruses belong to one of the following broad taxa: bacteria, plant and fungi, vertebrates, invertebrates, and human. Based on their approach, the authors find that structural similarity between viral proteins and host proteins is enriched among the infected host taxon compared to other taxa that are not infected by the virus. They conclude that the structure space mimicked by viruses is guided by the structure space of the infected host proteome.

Several studies have shown before that viral proteomes tend to mimic specific host proteomes in their structural properties as a means to hijack cellular processes and pathways. This mimicry is more evident at the structural domain level and protein interaction interface level. The current study is the first attempt to detect virus-host structural mimicry at a large scale covering a broad taxonomic range. The analysis and results are therefore of broad interest. In addition, the reported case studies relating to disease are also relevant and timely.

The authors determine structural mimicry using a two-step approach. In brief, the authors first select a structural template for the viral protein based on sequence similarity. This template may be selected from any species. Then, the authors perform a structural similarity search to identify other structures in PDB that have high structural similarity with the selected viral protein template. They find that these highly similar structural neighbors are enriched among the host which is known to be infected by the virus.

A minority (<30%) of structural mimicry cases can be detected through sequence similarity alone. Are these highly similar sequence neighbors also enriched among the host which is known to be infected by the virus? These highly similar sequence neighbors are interesting because they are clearly related to the viral protein via divergent evolution.

The referee is correct in raising the observation that sequence homologs may reflect a common origin. There is an important distinction though. The analysis we present relies on gaining structural information of viral proteins from distantly related sequences of experimentally resolved structures. As such, our coverage is inherently limited by the scope and breadth of the PBD. So, while evaluation of protein structure conservation and mimicry is possible, proper analysis of sequence conservation among the minority of mimics (the <30% that the referee highlights) requires pairwise sequence alignment with fully sequenced host genomes (not only those proteins for which structure information can be inferred). Beyond the gaps in virus-host annotations, such analysis requires a dedicated effort that is beyond the scope of this work.

The authors correctly point out that in general, it is difficult to determine whether structural mimicry (with no detectable sequence similarity) is the result of divergent evolution or convergent evolution. However, given the global nature of structural similarity between viral and host proteins detected in this study, it seems less likely that through convergent evolution, viral and host proteins will independently evolve tertiary structures that are highly similar to each other in a global way. Rather, it seems more likely that through convergent evolution, viral and host proteins will independently evolve different tertiary structures to carry out similar functions, e.g., to bind the same interface in the host molecular circuitry ("interface mimicry").

We share the referee's view on this. It is reasonable to think that global structural mimicry relationships denote a common ancestral origin. Yet, we cannot rule out convergent evolution as a source of structural mimicry of smaller domains (e.g. <100 residues long) that would still meet the structural similarity criteria we set (SAS < 2.5Å). Nevertheless, we acknowledge the point and have amended the discussion to include a comment to better reflect this.

One caveat of the study is that while the survey identified >6 million instances of structural mimicry, it is difficult to know what fraction of these instances are true positives. For every true positive case of structural mimicry between viral and host proteins, there could be numerous false positive cases of other proteins from the same host or proteins from other hosts that have nothing to do with the virus but are incorrectly detected because they share similar tertiary structures with the true positive host protein. The authors are encouraged to comment on this caveat.

The referee raises a valid and critically important point. Indeed, we find instances of 'mimicry' between viral and host proteins with helicase and polymerase functions. In these cases, while the structural similarity is obvious, the 'mimicry' relationship is almost certainly spurious. As we commented on in the discussion, *"…some mimics that we identify may represent spurious structural relationships that stem from underlying functional class of a given pair of proteins (virus and host endoded proteases)…*" Distinguishing the mimicked host protein within a set of host proteins sharing a common ancestor (gene duplication event) is a challenging task that would require higher resolution details such as the identity of amino acids at key functional residues. This, however, falls outside the scope of the work presented here and would be an interesting question to tackle in the future. Having said that, the statistical comparisons shown in Figure 1 reveal that the 'signal' embedded in the structural mimics identified is higher than the noise.

I also recommend that the authors review the text to remove errors in sentence structure. For example, on Page 4 Line 89 there is an extra unneeded phrase "Ska was used". Page 11 Line 348, < 1x10^{-6} should be > 1x10^{-6}.

We thank the referee for the keen eye and careful reading of the manuscript. We have appropriately amended the sentence structure on page 4. As for page 11, the notation is correct. The lower the e-value the greater the significance of the sequence similarity. In our case we request the e-value to be < $1x10^{-6}$ to consider sequence homology.