Journal of proteome research

pubs.acs.org/jpr

Reviews

Applications of Protein Secondary Structure Algorithms in SARS-CoV-2 Research

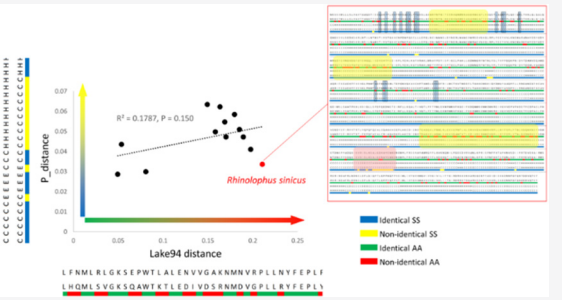
Alibek Kruglikov, Mohan Rakesh, Yulong Wei, and Xuhua Xia*

Cite This: https://dx.doi.org/10.1021/acs.jproteome.0c00734

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Since the outset of COVID-19, the pandemic has prompted immediate global efforts to sequence SARS-CoV-2, and over 450 000 complete genomes have been publicly deposited over the course of 12 months. Despite this, comparative nucleotide and amino acid sequence analyses often fall short in answering key questions in vaccine design. For example, the binding affinity between different ACE2 receptors and SARS-COV-2 spike protein cannot be fully explained by amino acid similarity at ACE2 contact sites because protein structure similarities are not fully reflected by amino acid sequence similarities. To comprehensively compare protein homology, secondary structure (SS) analysis is required. While protein structure is slow and difficult to obtain, SS predictions can be made rapidly, and a well-predicted SS structure may serve as a viable proxy to gain biological insight. Here we review algorithms and information used in predicting protein SS to highlight its potential application in pandemics research. We also showed examples of how SS predictions can be used to compare ACE2 proteins and to evaluate the zoonotic origins of viruses. As computational tools are much faster than wet-lab experiments, these applications can be important for research especially in times when quickly obtained biological insights can help in speeding up response to pandemics.



KEYWORDS: *COVID-19, spike protein, secondary structure, protein similarity, SARS-CoV-2*

1. INTRODUCTION

Since the outbreak of COVID-19 in late December of 2019, more than 450 000 full genomes of SARS-CoV-2 have been sequenced and deposited in GISAD database (https://www.gisaid.org/, last accessed February 1, 2021). Both SARS-CoV-2[1] and SARS-CoV[2−4] encode a Spike (S) protein, hereafter respectively referred to as SARS-2-S and SARS-S. The S1 receptor binding domain (RBD) binds to host Angiotensin-converting enzyme 2 (ACE2) receptor to mediate cell entry. The efficacy of this interaction determines host specificity and severity of infection.[4−6] Given a mammalian species, a high similarity between human ACE2 (hACE2) and mammalian ACE2 at S protein contact sites implies high susceptibility, and one can expect to determine species susceptibility to SARS-CoV or SARS-CoV-2 infections by comparative amino acid sequence analyses at contact sites at the ACE2 receptors.

2. SECONDARY STRUCTURE STUDIES ARE REQUIRED TO UNDERSTAND HOST SUSCEPTIBILITY TO SARS-COV-2

The above expectation, while largely correct, is not completely accurate. For example, of the 18 amino acid (aa) sites in contact between hACE2 and the RBD of SARS-S, nine aa sites differ between ferret ACE2 and hACE2, but both ferret ACE2 and hACE2 are effective as receptors for binding to RBD and mediating viral entry into host cells. In contrast, ACE2 from mouse and rat also differ from hACE2 by nine aa sites, but they cannot support viral RBD binding and viral entry.[2] This discrepancy invokes two simple explanations. First, aa sites beyond the 18 contact sites may also contribute to structural interactions and those sites might be more similar between hACE2 and ferret ACE2 than between hACE2 and mouse and rat ACE2. Second, structural similarity is not fully reflected in sequence similarity; i.e., structural similarity between hACE2 and ferret ACE2 may be greater than that between hACE2 and the mouse and rat ACE2. Only through structural studies can we hope to gain mechanistic insights into the differences in mammalian susceptibility to SARS-CoV-2.

Nevertheless, protein structure is difficult to obtain, and well-predicted protein secondary structure (SS) may serve as the next best answer. The Protein Data Bank (PDB) is the main depository of experimentally determined 3D protein structures, and around 160 thousand protein structures are

Received: September 20, 2020

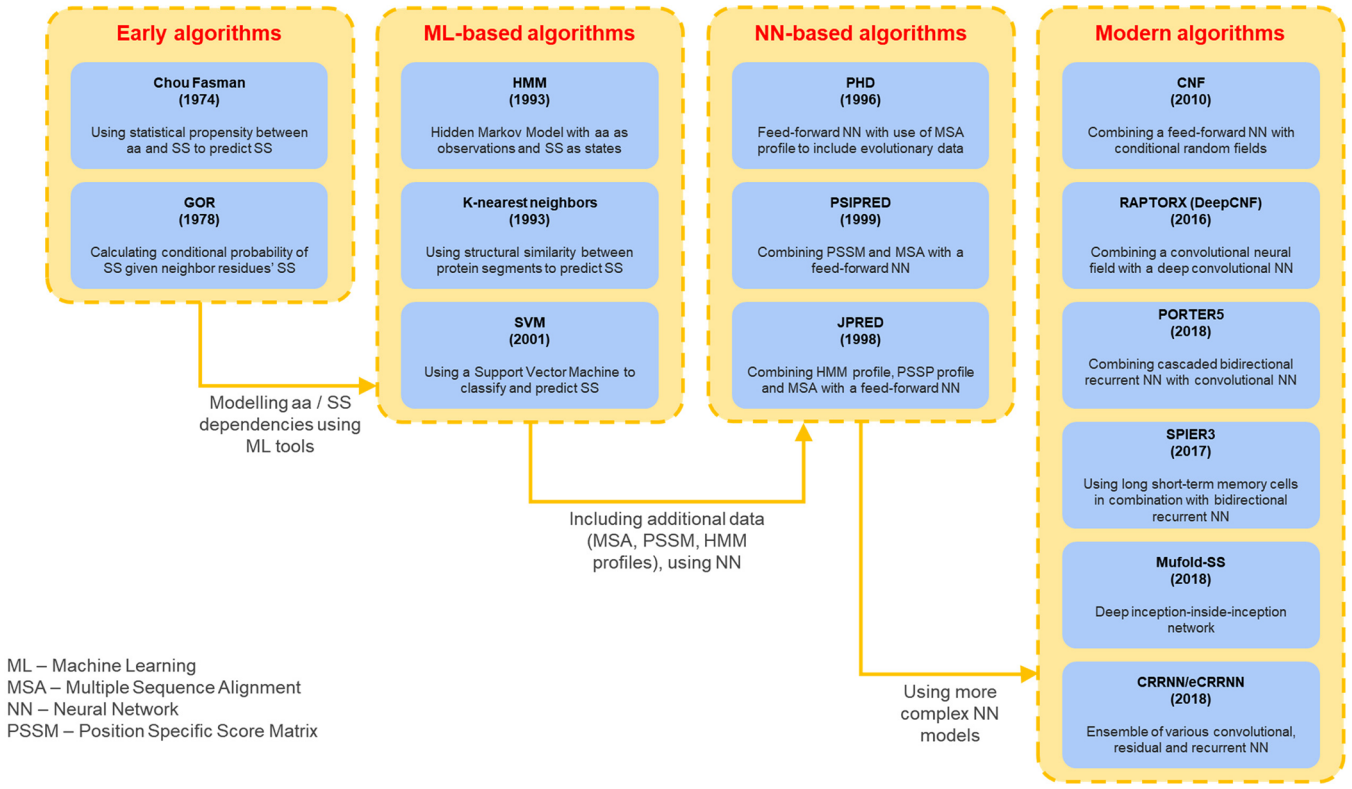


Figure 1. An overview of PSSP programs and implemented computational algorithms[18−31] developed over the past 50 years.

Table 1. A Comparison of PSSP Programs by Q3 Accuracy Assessments[a]

program	TS115 (%)	CASP10 (%)	CASP11 (%)	CASP12 (%)	TS2019 (%)	CB513 (%)
JPRED4[25]	77.1	81.6	80.4	78.8	76.6	81.7
PSIPRED v4.0[24]	80.2	81.2	80.7	80.5	82.3	79.2
CNF[26]	–	78.9	79.1	–	–	78.3
RAPTORX (DeepCNF)[27]	82.3	84.4	84.7	82.1	–	82.3
SPIDER3[29]	83.9	82.6	81.5	79.9	84.4	–
PORTER5[28]	–	–	–	–	84.5	–
MUFOLD-SS[30]	–	86.5	85.2	83.4	85.9	82.7
CRRNN[31]	–	86.1	84.2	82.6	–	87.3
eCRRNN[31]	–	87.8	85.9	83.7	–	87.8

[a]Accuracy scores (in percentage) are obtained from the programs' publication papers and from Yang et al.[32] and Smolarczuk et al.[33]

deposited.[7] In comparison, over 216 million aa sequences can be found in the NCBI GenBank database as of May 2020.[8] This inequality arises because experimental determination of structures is an expensive and lengthy process.[9,10]

In silico structure prediction techniques are faster and cheaper, and they have been useful in many research areas. For example, SS predictions have been used in enzyme structure similarity calculations,[11] ribosomal protein comparison,[12] protein activity mechanisms,[13] COVID-19 proteomics,[14] and many other areas. In section 3 we review examples of protein secondary structure predictions (PSSP) algorithms, and in section 4 we review their practical uses in pandemics research. In section 5, we describe examples of our own PSSP analyses on S protein-ACE2 binding to study species' susceptibility to SARS-CoV and SARS-2-CoV. The examples described in this review highlight how PSSP can be a useful tool in pandemics research.

3. AN EVALUATION OF CURRENT PSSP ALGORITHMS

In protein structure models, aa sequences are used to predict secondary and tertiary protein structures. SS are often classified in either three states or eight states of structures. Early PSSP models predict three secondary structure states: helix (H), strand (E), and coil (C), whereas in recent years, PSSP models have shifted to predict structures in eight states. Figure 1 summarizes PSSP programs developed over the years.

In addition to PSSP, protein structures can be modeled at the 2D level as contact maps[15] and at the 3D level as tertiary structures.[16,17] While modeling in 2D or 3D are appealing, there are several reasons why PSSP can be practical. First, unlike 2D or 3D structures, PSSP is reported as a sequence and can be used together with aa chains in multiple sequence alignments. This makes PSSP modeling useful in determining proteins that might be more similar in structures than in nucleotide or aa sequence. Second, the sequential nature allows alignment of SS elements with known or exploratory protein

https://dx.doi.org/10.1021/acs.jproteome.0c00734
J. Proteome Res. XXXX, XXX, XXX−XXX

hotspots. Lastly, PSSP is faster and less computation-heavy than 3D predictions.

Typically, three metrics are used to evaluate accuracy of PSSP programs: Q3, Q8, and Segment Overlap (SOV) scores. Q3 and Q8 represent the percentages of SS sequence positions correctly predicted by the models using three or eight structure states, respectively. SOV is a more complex measure that represents the percentage of segment overlap between predicted and correct sequences. Different protein databases can be used for the evaluation, and the best practice is to use multiple data sets. Tables 1 and 2 show a collection of different

Table 2. A Comparison of PSSP Programs by Q8 Accuracy Assessments[a]

program	CASP10 (%)	CASP11 (%)	CASP12 (%)	TS2019 (%)	CB513 (%)
CNF[26]	64.8	65.1	–	–	64.9
RAPTORX (DeepCNF)[27]	71.8	72.3	69.8	–	68.3
PORTERS[28]	–	–	–	73.6	–
MUFOLD-SS[30]	76.5	74.5	72.1	74.9	70.6
CRRNN[31]	73.8	71.6	68.7	–	71.4
eCRRNN[31]	76.3	73.9	70.7	–	74.0

[a]Accuracy scores (in percentage) are obtained from program publication papers and from Yang et al.[32] and Smolarczuk et al.[33]

PSSP models' accuracies calculated using various protein data sets.[27−33] Note that models are continually retrained with new protein structures, so there are discrepancies in reported accuracy values. Also, depending on data sets and metrics used, results of PSSP programs comparisons vary.

In addition to prediction accuracy, it is important to consider the programs' usability and their limitations. While some programs are readily available through web servers, predictions through server are often limited by sequence length or number. For example, Mufold-SS only allows sequences of up to 700 aa long and Jpred4 only allows sequences of up to 800 aa long. In addition, most web servers only allow prediction of one protein sequence at a time, which is often impractical when working with a large number of sequences. Standalone versions of the programs do not have the restrictions of the web servers.

4. PSSP METHODS HAVE BEEN USED WIDELY IN PANDEMICS RESEARCH

4.1. Structural Conformation at SARS-CoV nsp5 Protein

Lu et al.[34] explored the structure of the SARS-CoV *nsp5* gene. With reference to SARS-CoV strain GD, comparative sequence analyses with 110 strains at *nsp5* showed that five *nsp5* had mutations. Secondary structure predictions were performed at the five mutated strains using PSIPRED and the analysis showed that all five mutated strains had identical predicted secondary structure, which implies that *nsp5* encoded proteins retain a conserved structure and may be a better therapeutic target than more rapidly evolving genes.

4.2. Rapid Evolution of Pandemic Norovirus Genogroups

Bull et al.[35] examined RNA polymerase and capsid protein similarities in five norovirus genogroups, of which the GII.4 genogroup was associated with acute gastroenteritis global outbreaks. To evaluate whether this highly pathogenic genogroup had a greater epidemiological fitness than the other four genogroups, rate of mutation at RNA polymerase and capsid secondary structures were modeled using the CPHmodels Server.[36] The PSSP model revealed that the 15 varying amino acid residues on capsid were located on the exposed loops in GII.4. Moreover, more pathogenic genogroups had more similarities with GII.4 in structure than less pathogenic ones.

4.3. Identification of a Potential Inhibitor of H1N1 Neuraminidase

Seniya et al.[37] studied the potential effect of the *Boesenbergia pandurata* metabolite 4-hydroxy panduratin A to inhibit spread of Influenza A H1N1 (swine flu) infection. Influenza has two major surface proteins, neuraminidase (NA) and hemagglutinin (HA), to facilitate viral breach into host cell. To evaluate the potential of 4-hydroxy panduratin A to dock into active binding pockets of H1N1 NA, a homology-based protein structure prediction program, Modeler 9.10,[38] was used. In addition, I-TASSER[39] prediction was also used in combination with ab initio methods of modeling. These steps required secondary structure templates which were predicted using the PSIPRED server and rated using Z scores in LOMETS.[40] The combination of PSSP and I-TASSER enabled the downstream analysis of protein interactions between the viral NA and the plant metabolite.

4.4. Determining Conserved Segments of H7N9 Hemagglutinin

Sarkar et al.[16] examined the Avian Influenza A (H7N9) hemagglutinin (HA) protein to determine conserved HA regions that could serve as potential peptide vaccines. As aforementioned, HA is one of the two major surface proteins that facilitate viral entry into host cells. In addition, HA can also elicit an antibody response during infection. The PSSP server, SABLE,[41] was used to predict accessible surface area (ASA) in 120 HA sequences from H7N9 strains, and Jpred[42] and HHpred[43] were used to verify results. ASA, like secondary structure, is a 1D prediction; the aa sequence is converted to a sequence of numerical values, between 0 and 100, that describes aa sites accessibility in the solvent. Eight highly accessible regions were predicted by ASA and through epitope prediction, four regions were found with promising immunogenic potential.

4.5. Computationally Designed Peptides to Block Binding between SARS-2-S and Host ACE2

Good binding between SARS-2-S and host ACE2 receptor is crucial for viral entry into host cells. This interaction has been extensively explored by experimental research as a COVID-19 vaccine target and by computational research aiming to design competitive binding peptides[44] to bring forth new avenues to COVID-19 treatment. Using computational tools EvoEF2[45] and EvoDesign,[46] Huang et al.[44] designed peptide sequences that potentially bind competitively to SARS-2-S to limit viral entry. On the basis of a hACE2 structure template, they explored thousands of peptide designs through 3D modeling and selected best candidates by SARS-2-S binding affinity scored by PSSP performed in EvoDesign. The computational nature of this study allowed results to be obtained rapidly; currently, the computationally designed peptides are being evaluated experimentally.[44]

https://dx.doi.org/10.1021/acs.jproteome.0c00734
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 3. Average PSSP Program Accuracies as Measured Using ACE2 and Spike Protein Data from PDB[a]

protein set	metric	PORTER5[28] (%)	MUFOLD-SS[30] (%)	PSIPRED[24] (%)	JPRED4[25] (%)
totals (other 2 sets combined)	Q3	75.2	77.1	77.7	76.5
	Q8	62.8	64.0	61.0	60.9
	SOV	57.6	57.8	60.3	58.3
hACE2 (1r42:A, 6m0j:A, 6m18:B, 6m1d:B, 6m17:B)	Q3	81.2	82.0	82.0	80.5
	Q8	69.9	70.8	65.2	65.1
	SOV	71.2	67.5	72.3	69.7
SARS-2-S S1 (6vxx:A, 6vyb:A, 6m0j:E, 6m17:E)	Q3	67.8	71.0	72.4	71.4
	Q8	54.0	55.5	55.7	55.6
	SOV	40.6	45.8	45.4	44.0

[a]PDB IDs are shown below the set names.

5. USING PSSP MODELS TO GAIN BIOLOGICAL INSIGHT INTO SARS-COV-2 AND SARS-COV INFECTIVITY

Focusing on SARS-CoV-2, we tested the ability of several PSSP programs to predict SS of hACE2 and SARS-2-S S1 domain. We used experimentally derived SS from ACE2 structures available on PDB (1r42:A, 6m0j:A, 6m18:B, 6m1d:B, and 6m17:B; S1: 6vxx:A, 6vyb:A, 6m0j:E, and 6m17:E) to compare with SS predictions. Table 3 shows that the accuracy metrics of SS predicted for ACE2 and for SARS-2-S S1 were much lower than test scores from Tables 1 and 2, possibly because membrane protein structures are hard to predict. Another possible reason is that the training data used for the PSSP programs were not specific enough to predict ACE2 and S1 proteins more accurately. The Q8 results for PSIPRED and JPRED4, which only predict three structure states, were expected to be lower than that of PORTER5 and MUFOLD-SS, which predicted eight structure states. However, Q8 results were similar for all four programs (Table 3), possibly because extra types of secondary structures are rare in the studied proteins.

As previously mentioned, mammalian susceptibility to SARS-CoV cannot always be accurately predicted by differences in ACE2 aa sequences. This problem can be viewed as a mismatch between empirical and theoretical results. Using ACE2 PSSP instead of aa sequences, we attempt to explain this mismatch. To showcase that PSSP can circumvent this mismatch, Table 4 shows the P_distance, a measurement of differences in predicted SS between hACE2 and other species' ACE2. Here, we choose to use Mufold-SS to predict ACE2 SS (Table 3). P_distance is based on Q3 and Q8 scores, and the formula used for calculation is shown in eq 1, where M is the number of residues that are the same in both windows and L is sequence length (analogous to Q3/Q8 evaluations). Mufold-SS can be robust with three states but not with eight states, as it assumes equal weight for all SS differences. Hence, all calculated P_distances (Table 4) were based on three-state SS predictions.

$$\text{P\_distance} = 1 - \left(\frac{M}{L}\right) \quad (1)$$

The P_distance shows that SS variations better explain patterns of SARS-CoV infectivity than hotspot aa differences. First, unlike differences in ACE2 aa, differences in ACE2 SS corroborate the finding that rats[47] are less susceptible to SARS-CoV than palm civets[48] and mice,[49] with P_distances of 0.0509 (rats) vs 0.0472 (palm civets and mice). Second, ACE2 SS explains why Chinese horseshoe bats (P_distance = 0.0335) are more susceptible to SARS-CoV than Pearson's horseshoe bats (P_distance = 0.0410).[50] Nonetheless, our findings cannot be generalized further, as not all patterns of infectivity are explained through P_distance. For example, P_distance cannot explain why palm civets (0.0472) are more susceptible to SARS-CoV than Pearson's horseshoe bat (0.0410).[48,50]

Table 4. P_distances between hACE2 SS and Mammalian ACE2 SS[a]

SS sequence	P_distance
NM_001135696_Macaca_mulatta (Macaque)	0.0286
XM_008988993_Callithrix_jacchus (Marmoset)	0.0298
GQ999936_Rhinolophus_sinicus (Chinese horseshoe bat)	0.0335
EF569964_Rhinolophus_pearsonii (Pearson's horseshoe bat)	0.0410
AY996037_Cercopithecus_aethiops (African green monkey)	0.0435
NM_001130513_Mus_musculus (Mouse)	0.0472
AY881174_Paguma_larvata (Civet)	0.0472
XM_005074209_Mesocricetus_auratus (Hamster)	0.0497
NM_001012006_Rattus_norvegicus (Rat)	0.0509
AB211998_Procyon_lotor (Raccoon)	0.0547
NM_001310190_Mustela_putorius_furo (Ferret)	0.0584
EU024940_Nyctereutes_procyonoides (Raccoon dog)	0.0622
NM_001039456_Felis_catus (Cat)	0.0634

[a]ACE2 SS are predicted by Mufold-SS.[30]

To further examine the ACE2 of species shown in Table 4, we calculated aa sequence similarities using the Lake94[51] phylogenetic distance with hACE2 as reference. Indeed, with respect to hACE2, aa sequence similarities as measured by Lake94 poorly reflect similarities at SS as measured by P_distance in many species (Figure 2: R^2 = 0.179, P = 0.150), an example is *Rhinolophus sinicus*.

We next performed multiple sequence alignment (MSA) using MAFFT[52] on ACE2 aa sequence and on predicted ACE2 SS sequence for *Rhinolophus sinicus* highlighted in red in Figure 2. Hotspot sites were highlighted in the alignment, representing hACE2 sites S19, Q24, D30, K31, H34, E35, E37, D38, Y41, Q42, L79, M82, Y83, K353, and R393 that form contact with SARS-2-S at sites K417, G446, Y449, L455, F456, A475, F486, N487, Y489, Q498, T500, N501, G502, and Y505, as previously identified through X-ray crystallography experiments.[53,54]

Rhinolophus sinicus ACE2 seems to be more conserved at hotspot locations (boxed in light blue) than other regions at the SS level (Figure 3). Furthermore, lack of SS differences at some aa substitution sites can be explained by the nature of aa substitutions: some aa substitutions are considered conservative as they have similar physicochemical properties.[55] Indeed,

https://dx.doi.org/10.1021/acs.jproteome.0c00734
J. Proteome Res. XXXX, XXX, XXX−XXX

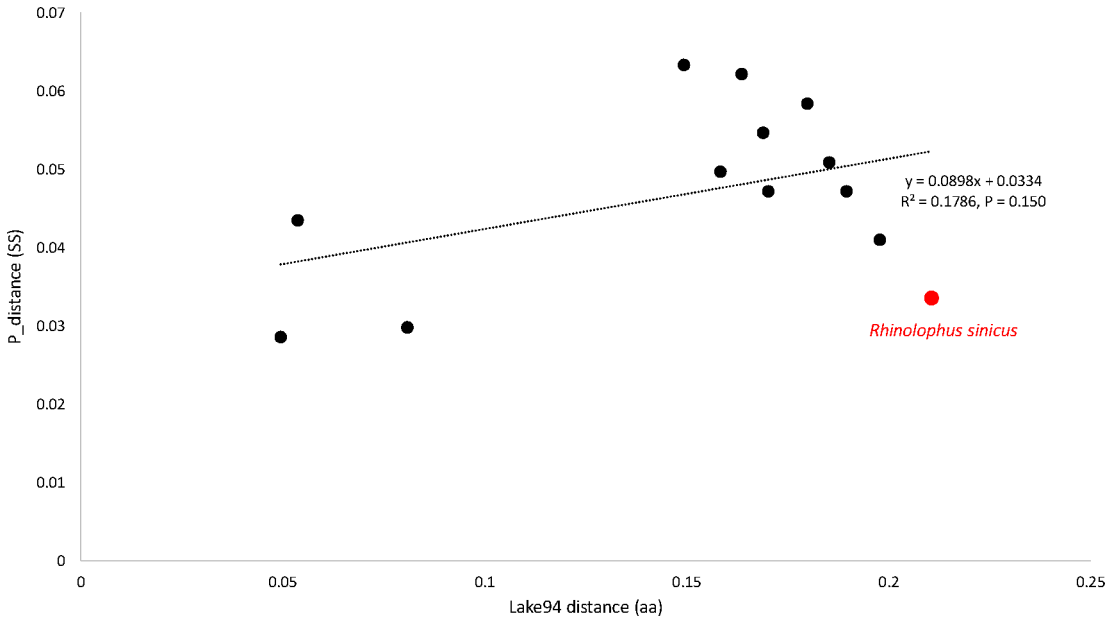


Figure 2. Lake94 distances measured at ACE2 aa sequences poorly correlate P_distance measured at ACE2 SS. Sequence distances in mammalian ACE2 are calculated with respect to hACE2, and the 13 species considered are those listed in Table 4.

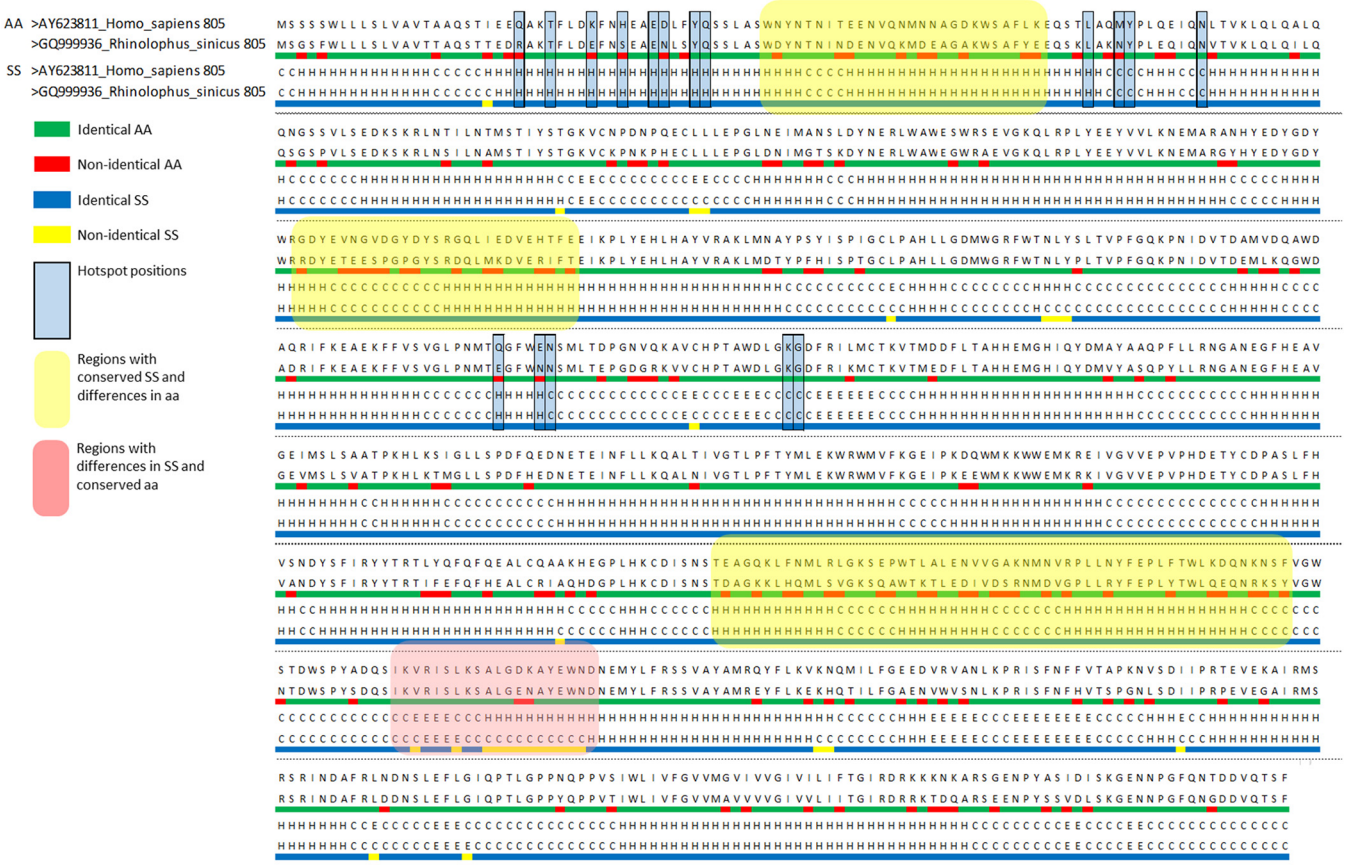


Figure 3. SS and aa alignments between *Rhinolophus sinicus* ACE2 and hACE2. Match and mismatch sites are respectively indicated by green and red for aa alignment and by blue and yellow for SS alignment. Notable regions where conservation levels differ between aa and SS alignments are boxed in light red and yellow. Hotspot positions boxed in light blue represent SARS-2-S contacting sites at hACE2.[53,54]

conservative D ↔ E, D ↔ N, E ↔ N, E ↔ Q, and K ↔ R are present at the regions boxed in yellow (Figure 3); these amino acids have similar properties and reduced substitution effects on predicted SS folding. On the other hand, some regions have many SS differences but relatively conserved aa (Figure 3: boxed in light red), one explanation for this discrepancy is that aa substitutions may influence SS at distant loci rather than closer ones due to complexities of hydrogen bond formation.

https://dx.doi.org/10.1021/acs.jproteome.0c00734
J. Proteome Res. XXXX, XXX, XXX−XXX

Moreover, Lysine has been reported as preferred amino acids at C-terminus of proteins for α-helix formation,[56] and reduced helix stabilization in the light red region could be caused by the K → N substitution.

6. CONCLUSION

Here we reviewed potential applications of PSSP programs to gain biological insights. These fast methods can be helpful to obtain important answers as an immediate response in pandemics research. Because some mutations, especially substitutions, might not induce structural changes, analysis on SS expands upon analysis of aa. In this review, we evaluated some of the current PSSP programs and discussed PSSP applications in pandemics research. Additionally, we offered examples of PSSP analyses with a focus on SARS-CoV and SARS-CoV-2. Because coronavirus infection is achieved through binding between the viral Spike protein and the host ACE2 receptor, mammals with similar ACE2 structures could be potentially susceptible to these viruses. To identify ACE2 similarities between mammals and humans, comparisons were made at aa and SS levels. We showed that variations between predicted SS is not always consistent with variations in corresponding aa sequences. Specifically, differences at aa rarely led to different SS at ACE2 hotspot locations in *Rhinolophus sinicus*. The example above, along with other practical examples reviewed, highlight potential applications of PSSP algorithms in pandemics research.

AUTHOR INFORMATION

Corresponding Author

Xuhua Xia − *Department of Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada; Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada;* orcid.org/0000-0002-3092-7566; Email: xxia@uottawa.ca

Authors

Alibek Kruglikov − *Department of Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada;* orcid.org/0000-0001-6074-8764

Mohan Rakesh − *Department of Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada*

Yulong Wei − *Department of Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00734

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to X.X. [RGPIN/2018-03878], and NSERC Doctoral Scholarship to Y.W. [CGSD/2019-535291].

REFERENCES

(1) Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; Chen, H.-D.; Chen, J.; Luo, Y.; Guo, H.; Jiang, R.-D.; Liu, M.-Q.; Chen, Y.; Shen, X.-R.; Wang, X.; Zheng, X.-S.; Zhao, K.; Chen, Q.-J.; Deng, F.; Liu, L.-L.; Yan, B.; Zhan, F.-X.; Wang, Y.-Y.; Xiao, G.-F.; Shi, Z.-L. A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* **2020**, *579* (7798), 270−273.

(2) Lu, G.; Wang, Q.; Gao, G. F. Bat-to-Human: Spike Features Determining 'Host Jump' of Coronaviruses SARS-CoV, MERS-CoV, and Beyond. *Trends Microbiol.* **2015**, *23* (8), 468−478.

(3) Hulswit, R. J. G.; de Haan, C. A. M.; Bosch, B.-J. Chapter Two: Coronavirus Spike Protein and Tropism Changes. In *Advances in Virus Research*; Ziebuhr, J., Ed.; Coronaviruses; Academic Press, 2016; Vol. *96*, pp 29−57. DOI: 10.1016/bs.aivir.2016.08.004.

(4) Hoffmann, M.; Hofmann-Winkler, H.; Pöhlmann, S. Priming Time: How Cellular Proteases Arm Coronavirus Spike Proteins. *Activation of Viruses by Host Proteases* **2018**, 71−98.

(5) Coutard, B.; Valle, C.; de Lamballerie, X.; Canard, B.; Seidah, N. G.; Decroly, E. The Spike Glycoprotein of the New Coronavirus 2019-NCoV Contains a Furin-like Cleavage Site Absent in CoV of the Same Clade. *Antiviral Res.* **2020**, *176*, 104742.

(6) Andersen, K. G.; Rambaut, A.; Lipkin, W. I.; Holmes, E. C.; Garry, R. F. The Proximal Origin of SARS-CoV-2. *Nat. Med.* **2020**, *26* (4), 450−452.

(7) Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In *Protein Crystallography*; Springer, 2017; pp 627−641.

(8) Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2016**, *44*, D67−D72.

(9) Terwilliger, T. C.; Stuart, D.; Yokoyama, S. Lessons from Structural Genomics. *Annu. Rev. Biophys.* **2009**, *38* (1), 371−383.

(10) Mardis, E. R. Anticipating the $1,000 Genome. *Genome Biol.* **2006**, *7* (7), 112.

(11) Rehman, S.; Grigoryeva, L. S.; Richardson, K. H.; Corsini, P.; White, R. C.; Shaw, R.; Portlock, T. J.; Dorgan, B.; Zanjani, Z. S.; Fornili, A.; Cianciotto, N. P.; Garnett, J. A. Structure and Functional Analysis of the Legionella Pneumophila Chitinase ChiA Reveals a Novel Mechanism of Metal-Dependent Mucin Degradation. *PLoS Pathog.* **2020**, *16* (5), e1008342.

(12) Anger, A. M.; Armache, J.-P.; Berninghausen, O.; Habeck, M.; Subklewe, M.; Wilson, D. N.; Beckmann, R. Structures of the Human and Drosophila 80S Ribosome. *Nature* **2013**, *497* (7447), 80−85.

(13) Wu, Y. I.; Frey, D.; Lungu, O. I.; Jaehrig, A.; Schlichting, I.; Kuhlman, B.; Hahn, K. M. A Genetically Encoded Photoactivatable Rac Controls the Motility of Living Cells. *Nature* **2009**, *461* (7260), 104−108.

(14) Jumper, J.; Tunyasuvunakool, K.; Kohli, P.; Hassabis, D. Computational predictions of protein structures associated with COVID-19, https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.

(15) Yuan, X.; Bystroff, C. Protein Contact Map Prediction. In *Computational Methods for Protein Structure Prediction and Modeling*; Xu, Y., Xu, D., Liang, J., Eds.; Biological and Medical Physics Biomedical Engineering; Springer: New York, NY, 2007; Vol. *1*: Basic Characterization, pp 255−277. DOI: 10.1007/978-0-387-68372-0_8.

(16) Sarkar, T.; Das, S.; De, A.; Nandy, P.; Chattopadhyay, S.; Chawla-Sarkar, M.; Nandy, A. H7N9 Influenza Outbreak in China 2013: In Silico Analyses of Conserved Segments of the Hemagglutinin as a Basis for the Selection of Peptide Vaccine Targets. *Comput. Biol. Chem.* **2015**, *59*, 8−15.

(17) Kwon, S. C.; Nguyen, T. A.; Choi, Y.-G.; Jo, M. H.; Hohng, S.; Kim, V. N.; Woo, J.-S. Structure of Human DROSHA. *Cell* **2016**, *164* (1), 81−90.

(18) Chou, P. Y.; Fasman, G. D. Prediction of Protein Conformation. *Biochemistry* **1974**, *13* (2), 222−245.

(19) Kloczkowski, A.; Ting, K.-L.; Jernigan, R. L.; Garnier, J. Protein Secondary Structure Prediction Based on the GOR Algorithm Incorporating Multiple Sequence Alignment Information. *Polymer* **2002**, *43* (2), 441−449.

(20) Asai, K.; Hayamizu, S.; Handa, K. Prediction of Protein Secondary Structure by the Hidden Markov Model. *Bioinformatics* **1993**, *9* (2), 141−146.

(21) Yi, T.-M.; Lander, E. S. Protein Secondary Structure Prediction Using Nearest-Neighbor Methods. *J. Mol. Biol.* **1993**, *232* (4), 1117−1129.
(22) Hua, S.; Sun, Z. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach1 1Edited by B. Holland. *J. Mol. Biol.* **2001**, *308* (2), 397−407.
(23) Rost, B.; Sander, C.; Schneider, R. PHD-an Automatic Mail Server for Protein Secondary Structure Prediction. *Bioinformatics* **1994**, *10* (1), 53−60.
(24) McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED Protein Structure Prediction Server. *Bioinformatics* **2000**, *16* (4), 404−405.
(25) Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G. J. JPred4: A Protein Secondary Structure Prediction Server. *Nucleic Acids Res.* **2015**, *43* (W1), W389−W394.
(26) Wang, Z.; Zhao, F.; Peng, J.; Xu, J. Protein 8-Class Secondary Structure Prediction Using Conditional Neural Fields. *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2010**, 109−114.
(27) Wang, S.; Peng, J.; Ma, J.; Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **2016**, DOI: 10.1038/srep18962.
(28) Torrisi, M.; Kaleel, M.; Pollastri, G. Porter 5: Fast, State-of-the-Art Ab Initio Prediction of Protein Secondary Structure in 3 and 8 Classes. *bioRxiv*, Oct 5, 2018, 289033. .
(29) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing Non-Local Interactions by Long Short-Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers and Solvent Accessibility. *Bioinformatics* **2017**, *33* (18), 2842−2849.
(30) Fang, C.; Shang, Y.; Xu, D. MUFOLD-SS: New Deep Inception-inside-Inception Networks for Protein Secondary Structure Prediction. *Proteins: Struct., Funct., Genet.* **2018**, *86* (5), 592−598.
(31) Zhang, B.; Li, J.; Lü, Q. Prediction of 8-State Protein Secondary Structures by a Novel Deep Learning Architecture. *BMC Bioinf.* **2018**, *19* (1), 293.
(32) Yang, Y.; Gao, J.; Wang, J.; Heffernan, R.; Hanson, J.; Paliwal, K.; Zhou, Y. Sixty-Five Years of the Long March in Protein Secondary Structure Prediction: The Final Stretch? *Briefings Bioinf.* **2018**, *19* (3), 482−494.
(33) Smolarczyk, T.; Roterman-Konieczna, I.; Stapor, K. Protein Secondary Structure Prediction: A Review of Progress and Directions. *Curr. Bioinf.* **2020**, *15* (2), 90−107.
(34) Lu, J.-H.; Zhang, D.-M.; Wang, G.-L.; Guo, Z.-M.; Li, J.; Tan, B.-Y.; Ou-Yang, L.-P.; Ling, W.-H.; Yu, X.-B.; Zhong, N.-S. Sequence Analysis and Structural Prediction of the Severe Acute Respiratory Syndrome Coronavirus Nsp5. *Acta Biochim. Biophys. Sin.* **2005**, *37* (7), 473−479.
(35) Bull, R. A.; Eden, J.-S.; Rawlinson, W. D.; White, P. A. Rapid Evolution of Pandemic Noroviruses of the GII.4 Lineage. *PLoS Pathog.* **2010**, *6* (3), e1000831.
(36) Lund, O.; Nielsen, M.; Lundegaard, C.; Worning, P. CPH Models 2.0: X3M a Computer Program to Extract 3D Models. *CASP Conference*; 2002.
(37) Seniya, C.; Khan, G. J.; Misra, R.; Vyas, V.; Kaushik, S. In-Silico Modelling and Identification of a Possible Inhibitor of H1N1 Virus. *Asian Pac. J. Trop. Dis.* **2014**, *4*, S467−S476.
(38) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234* (3), 779−815.
(39) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7.
(40) Wu, S.; Zhang, Y. LOMETS: A Local Meta-Threading-Server for Protein Structure Prediction. *Nucleic Acids Res.* **2007**, *35* (10), 3375−3382.
(41) Adamczak, R.; Porollo, A.; Meller, J. Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins. *Proteins: Struct., Funct., Genet.* **2005**, *59* (3), 467−475.
(42) Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J. JPred: A Consensus Secondary Structure Prediction Server. *Bioinformatics* **1998**, *14* (10), 892−893.
(43) Söding, J.; Biegert, A.; Lupas, A. N. The HHpred Interactive Server for Protein Homology Detection and Structure Prediction. *Nucleic Acids Res.* **2005**, *33*, W244−248.
(44) Huang, X.; Pearce, R.; Zhang, Y. De Novo Design of Protein Peptides to Block Association of the SARS-CoV-2 Spike Protein with Human ACE2. *Aging* **2020**, *12* (12), 11263−11276.
(45) Huang, X.; Pearce, R.; Zhang, Y. EvoEF2: Accurate and Fast Energy Function for Computational Protein Design. *Bioinformatics* **2020**, *36* (4), 1135−1142.
(46) Pearce, R.; Huang, X.; Setiawan, D.; Zhang, Y. EvoDesign: Designing Protein−Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *J. Mol. Biol.* **2019**, *431* (13), 2467−2476.
(47) Holmes, K. V. Adaptation of SARS Coronavirus to Humans. *Science* **2005**, *309* (5742), 1822−1823.
(48) Guan, Y.; Zheng, B. J.; He, Y. Q.; Liu, X. L.; Zhuang, Z. X.; Cheung, C. L.; Luo, S. W.; Li, P. H.; Zhang, L. J.; Guan, Y. J.; Butt, K. M.; Wong, K. L.; Chan, K. W.; Lim, W.; Shortridge, K. F.; Yuen, K. Y.; Peiris, J. S. M.; Poon, L. L. M. Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. *Science* **2003**, *302* (5643), 276−278.
(49) Li, W.; Greenough, T. C.; Moore, M. J.; Vasilieva, N.; Somasundaran, M.; Sullivan, J. L.; Farzan, M.; Choe, H. Efficient Replication of Severe Acute Respiratory Syndrome Coronavirus in Mouse Cells Is Limited by Murine Angiotensin-Converting Enzyme 2. *J. Virol.* **2004**, *78* (20), 11429−11433.
(50) Hou, Y.; Peng, C.; Yu, M.; Li, Y.; Han, Z.; Li, F.; Wang, L.-F.; Shi, Z. Angiotensin-Converting Enzyme 2 (ACE2) Proteins of Different Bat Species Confer Variable Susceptibility to SARS-CoV Entry. *Arch. Virol.* **2010**, *155* (10), 1563−1569.
(51) Lake, J. A. Reconstructing Evolutionary Trees from DNA and Protein Sequences: Paralinear Distances. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91* (4), 1455−1459.
(52) Katoh, K.; Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30* (4), 772−780.
(53) Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; Wang, X. Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor. *Nature* **2020**, *581* (7807), 215−220.
(54) Shang, J.; Ye, G.; Shi, K.; Wan, Y.; Luo, C.; Aihara, H.; Geng, Q.; Auerbach, A.; Li, F. Structural Basis of Receptor Recognition by SARS-CoV-2. *Nature* **2020**, *581* (7807), 221−224.
(55) Yampolsky, L. Y.; Stoltzfus, A. The Exchangeability of Amino Acids in Proteins. *Genetics* **2005**, *170* (4), 1459−1472.
(56) Forood, B.; Feliciano, E. J.; Nambiar, K. P. Stabilization of Alpha-Helical Structures in Short Peptides via End Capping. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (3), 838−842.

https://dx.doi.org/10.1021/acs.jproteome.0c00734
J. Proteome Res. XXXX, XXX, XXX−XXX