**Abstract: Modeling the Cleavage Potential of Cross-Species TMPRSS2 Variants to Subunits of SARS-CoV-2**

**Tyshawn Ferrell1,2, Kelly Pierce1, Johnny Aldan1, Ethan Ho1, Dairian Balai 1**

1Texas Advance Computing Center**, University of Texas at Austin, Austin, TX, 78712-78758**

2Department of Biology, **Albany State University, Albany, GA 31705**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a positive-stranded RNA virus, has caused >100,000 infections and >50,000 deaths during the 2020 pandemic. SARS-CoV-2 successful infection rate has been due to structural configurations of its envelope and specific cell receptor targeting. The spike (S) glycoprotein mediates cell entry by recognizing the angiotensin converting enzyme II (ACE2) receptor, which initializes cleavage of the protein to induce cell membrane fusion. However, in previous studies transmembrane protease/serine subfamily member 2 (TMPRSS2) have increased cleavage efficiency towards the S protein subunits. This study aims to characterize cross-species variants of TMPRSS2 and its potential to cleave deviates of S proteins in silico. We hypothesize that TMPRSS2-like endoprotease aids in the spillover of SARS-CoV-2 to alternate host by cleaving S protein subunits. We investigated the conserveness of TMPRSS2 sequences amongst species and assess their similarity to the human isoform. We utilized protein databanks to form concise multiple sequence alignments of seven domestic animals for comparisons with the human variant. We characterized differences in catalytically important residues based off the position specific score matrix and the property changes with the sequence. Our results demonstrated that important residues within TMPRSS2 are highly conserved, substitutions led to had minor affect on catalytic activity, in theory. This study has provided insight on possible spillover into alternate host with structurally similar TMPRSS2 variants compared to humans; posing an immense threat for epidemics in domesticated animal populations.

**Introduction**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel betacoronavirus, that curtailed the world’s health system within a matter of months. As of July 20th, 2020, 14,348,858 laboratory-confirmed infections and 603,000 SARS-CoV-2 related deaths have been reported globally [12]. This is due to SARS-CoV-2 mode of transmission and high bonding affinity for the angiotensin converting enzyme II (ACE2) receptor in airway epithelial cells. More specifically the S1 subunit of the S protein binds to ACE2 via its receptor binding domain (RBD), while the S2 subunit is used for virus-cell membrane fusion [1,3,6]. In previous studies it has been shown that SARS-CoV have a preference for cells with an endoprotease, transmembrane protease/serine subfamily member 2 (TMPRSS2), which increases the ability for virus-host membrane fusion [6,7,9]. In silico studies have shown that SARS-CoV-2 ability to infect alternate host depends on ACE2 based on the conserveness of the enzyme compared to the human strain [8]. Though ACE2 is an essential component for attachment of the S protein, it is the host cell proteases that activates the S2 subunit that holds the transmembrane fusion machinery.

TMPRSS2 is a member of the type II Transmembrane Serine Protease (TTSP), specifically belonging to the Hepsin/TMPRSS subfamily. TTSPs have four defining features: an N-terminal intracellular domain, transmembrane domain, “stem” domain, and a proteolytic domain. TMPRSS2 is synthesized as a zymogen that requires proteolytic processing to activate [2]. This activation results from the cleavage of extracellular substrates initialized by its Serine residue at the active site. In previous studies, TMPRSS2s proteolytic domain has shown to have increase pathogenesis by cleaving monobasic cleavage sites of virus transmembrane subunits within the coronavirus family [3-7]. The S protein in SARS-Cov-2 has an multibasic S1/S2’ cleavage site containing several arginine residues that are cleave by host cell proteases, which increases the efficacy of cell-cell spread [3]. However, in order for the virus-cell transmembrane fusion to occur, the monobasic S2’ cleavage site must be cleaved by TMPRSS2 [3-4]. TMPRSS2 proteolytic activity results from its catalytic domain, which consist of Ser-His-Asp, but Ser initializes the peptide hydrolysis by attacking the acyl compound of a lys or Arg residue in the S protein [2,4]. During this reaction two tetrahedral intermediates and one acylenzyme but finalizes with the protonation of an amine leaving group, carboxylic acid, water, and the catalytic triad reforms [2]. Even though TMPRSS2 is a crucial part for virus-host transmembrane fusion, much of its functionality is still unclear. In previous studies, ACE2 variants have been observed in multiple species, which may have lead SARS-CoV2 high infectivity in native and foreign host [1,8]. Since there are over 9000 sequences of TMPRSS2, a comparative analysis of its amino acid sequence can predict its cleaving efficacy of SARS-CoV-2 with similar ACE2 receptor sequences. This will yield valuable insight into SARS-CoV-2 zoonotic transmission potential to spill over in specie populations.

SARS-CoV-2, a zoonotic pathogen, has evolved to successfully infect a broad host range using common host cell machinery. With its high infectivity and unpredictive nature, SARS-CoV-2 is increasingly difficult to treat and contain. In previous studies TMPRSS2 has shown to cleave SARS-CoV at a monobasic Arg site, causing cell-cell spread and increased severity of coronaviruses in infected lung cells of mice [20-21]. However, TMPRSS2 variants in other species with similar ACE2 receptors (compared to human strains) could lead to unfavorable bonding conditions for SARS-CoV-2. Additionally, TMPRSS2’s proteolytic activation takes place after SARS-CoV S protein undergoes receptor-induced conformational modifications [21]. Therefore, we hypothesis that cross-species conservation of TMPRSS2 catalytic residues may play a role in alternate host spillover. By comparing residue alterations of TMPRSS2 multiple sequence alignments between various species and human, we’ll be able to score the susceptibility/severity of infection for each species, in silica. Also, we’ll be able to categorize each substitution’s property delta, then reflect that change in Pymol. By determining the molecular configuration leading to cell entry, we’ll gain insight on spillover potential of SARS-CoV-2 to alternate host. Which could arbor adaptive advantages for SARS-CoV-2 immune evasion and overall infectivity within broad an host range.

**Methods**

**Data Collection and Curation**

All 9757 TMPRSS2 protein sequences from 120 species were retrieved from EggN0G, as of June 30, 2020 [13]. Using Clustal Omega we generated the full set of each protein sequence alignment independently, which verified the authenticity of each alignment in the EggN0G dataset [14]. We utilized functions in Biopython, a python biological computation package, to parse the lines within the alignment file to convert header and sequence pairs into iterables. Then filtered out target mammalian species (Homo sapiens, Mus musculus, Canis Lupus familiaris, Felis catus, Bos Taurus, Equus caballus, and Gallus gallus) from the multiple sequence alignment of all 120 species, and modified it using the multiple sequence alignment function. From the target species list we extracted the H. sapien homolog sequence based upon UniProt’s TMPRSS2 homolog [17]. After trimming the H. sapien sequence using N-terminus and C-terminus indices, we generated the consensus sequence for the other target mammalian species, which were later trimmed using the same N/C-terminus indices as the H. sapien sequence. After modification of all species, we used the multipleSeqAlignment function between the human and other species’ sequences. We also use the position specific score matrix (PSSM) attribute to find the percent probability of a particular residue at a specific location.

**Protein Sequence Analysis**

We identified 24 TMPRSS2 amino acid residues that were previously reported to be important for proteolytic cleavage and conformation [2,4,15,16,17]. These residues include the assumed monobasic recognition site (K223 and K224), disulfide bonds (C113, C120, C126, C133, C139, C148, C172, C185, C231, C241, C281, C297 C365, C410, C426, C437, and C465), catalytic triad (D345, H296, and S441), and the binding site (D435). We compared the residues of each mammalian species to the TMPRSS2 homolog sequence in search of differences in particular residues of interest. Each substitution at a residue of interest was categorized as a mutation, which was then analyzed for its property difference based on data from previous studies (tryna get the reference from a prof). We retrieved the predicted structure of TMPRSS2 from SWISS-MODEL [18]. We modified the structure of TMPRSS2 to reflect protein substitutions based off the comparisons using Pymol [19].

Work in progress

**Results**

**References**

1. Damas, J., Hughes, G. M., Keough, K. C., Painter, C. A., Persky, N. S., Corbo, M., Hiller, M., Koepfli, K. P., Pfenning, A. R., Zhao, H., Genereux, D. P., Swofford, R., Pollard, K. S., Ryder, O. A., Nweeia, M. T., Lindblad-Toh, K., Teeling, E. C., Karlsson, E. K., & Lewin, H. A. (2020). Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. bioRxiv : the preprint server for biology, 2020.04.16.045302. <https://doi.org/10.1101/2020.04.16.045302>
2. Hedstrom, L. (2002). Serine protease mechanism and specificity. *Chemical reviews*, *102*(12), 4501-4524.
3. Hoffmann, M., Kleine-Weber, H., & Pöhlmann, S. (2020). A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell*.
4. Hoffmann, M., Hofmann-Winkler, H., & Pöhlmann, S. (2018). Priming time: How cellular proteases arm coronavirus spike proteins. In *Activation of Viruses by Host Proteases* (pp. 71-98). Springer, Cham.
5. Hofmann, H., & Pöhlmann, S. (2004). Cellular entry of the SARS coronavirus. Trends in microbiology, 12(10), 466-472.
6. Matsuyama, S., Nao, N., Shirato, K., Kawase, M., Saito, S., Takayama, I., ... & Sakata, M. (2020). Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. Proceedings of the National Academy of Sciences, 117(13), 7001-7003.
7. Meng, T., Cao, H., Zhang, H., Kang, Z., Xu, D., Gong, H., ... & Wei, H. (2020). The insert sequence in SARS-CoV-2 enhances spike protein cleavage by TMPRSS. BioRxiv.
8. Paniri, A., Hosseini, M. M., & Akhavan-Niaki, H. (2020). First comprehensive computational analysis of functional consequences of TMPRSS2 SNPs in susceptibility to SARS-CoV-2 among different populations. Journal of Biomolecular Structure and Dynamics, (just-accepted), 1-18.
9. Shirato, K., Kawase, M., & Matsuyama, S. (2018*). Wild-type human coronaviruses prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry*. Virology, 517, 9-15.
10. Shulla, A., Heald-Sargent, T., Subramanya, G., Zhao, J., Perlman, S., & Gallagher, T. (2011). *A transmembrane serine protease is linked to the severe acute respiratory syndrome coronavirus receptor and activates virus entry*. Journal of virology, 85(2), 873-882.
11. Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). *Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein*. Cell.
12. WHO, 2020 Who, *Novel Coronavirus(2019-nCoV) Situation Report 182*

<https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200720-covid-19-sitrep-182.pdf?sfvrsn=60aabc5c_2>

1. *Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, Peer Bork*. [Nucleic Acids Res. 2019 Jan 8; 47(Database issue): D309–D314. doi: 10.1093/nar/gky1085](https://doi.org/10.1093/nar/gky1085)
2. Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sciences. Protein Sci 27:135-145
3. Paoloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C., & Antonarakis, S. E. (1997). Cloning of the TMPRSS2 gene, which encodes a novel serine protease with transmembrane, LDLRA, and SRCR domains and maps to 21q22. 3. Genomics, 44(3), 309-320.
4. Ohno, A., Maita, N., Tabata, T., Nagano, H., Arita, K., Ariyoshi, M., ... & Kishimoto, K. (2020). Cystal structure of inhibitor-bound human MSPL/TMPRSS13 that can activate high pathogenic avian influenza. bioRxiv.
5. The UniProt Consortium  
   **UniProt: a worldwide hub of protein knowledge**  
   [Nucleic Acids Res. 47: D506-515 (2019)](https://doi.org/10.1093/nar/gky1049)
6. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research **42**, W252–258, doi:10.1093/nar/gku340 (2014).
7. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. https://pymol.org/2/ (2020).
8. Iwata-Yoshikawa, N., Okamura, T., Shimizu, Y., Hasegawa, H., Takeda, M., & Nagata, N. (2019). TMPRSS2 contributes to virus spread and immunopathology in the airways of murine models after coronavirus infection. Journal of virology, 93(6).
9. Matsuyama, S., Nagata, N., Shirato, K., Kawase, M., Takeda, M., & Taguchi, F. (2010). Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. Journal of virology, 84(24), 12658-12664.

Notes.

Partial cleavage can be accomplished by using enzymes, such as trypsin and chymotrypsin, that break bonds between specific amino acids.

* Trypsin: Cleaves the chain at the carboxyl groups of the basic amino acids lysine and arginine. The cleavage takes place in such a way that the amino acid with the charged side chain ends up at the C-terminal end of one of the peptides produced by the reaction. A peptide can be automatically identified as the C-terminal end of the original chain if its C-terminal amino acid is not a site of cleavage.
* Chymotryspin: Cleaves the chain at the carboxyl groups of the aromatic amino acids phenylalanine, tyrosine, and tryptophan.
* Protein domain- a conserved part of a given protein sequence and tertiary structure that can evolve, function, and exist independently of the rest of the protein chain.
* Quaternary structure is the final level of protein structure which is formed by two or more polypeptide chains. Each chain is termed a subunit of the structure.
  + Dimers, trimers, tetramers, consist of two, three, and four polypeptide chains, respectively.
  + Oligomer- an aggregate of several smaller units (monomers); bonding may be covalent or noncovalent.
    - Chains with noncovalent bond result in subtle changes that can alter the structure at one site of the protein molecule. This is known as allosteric hindess, which can cause conformational change in one subunit that induces a change in another subunit.
* Consensus Sequences- DNA sequences to which RNA polymerase binds; they are identical in many organisms
* Protein kinases- a class of enzymes that modify a protein by attaching a phosphate group to it
* I hypothesis that variation in the serine domain of TMPRSS2 will have an adverse effect on cross-species transmission of COVID-19. Specifically focusing on domesticated animal’s TMPRSS2 amino acid sequence. Protein folding and bonding affinity play a critical role in determining the cleavage potential for the proteolytic domain of TMPRSS2.

Interpreting Data

Hypothesis- This study aims to characterize cross-species variants of TMPRSS2 and its potential to cleave deviates of S proteins in silico. We hypothesize that TMPRSS2-like endoprotease aids in the spillover of SARS-CoV-2 to alternate host by cleaving S protein subunits.

* Compare the similarity of each animal sequence to the human strain
  + If the similarity of each sequence is high that means that TMPRSS2 has a highly conserved sequence meaning that cells supporting TMPRSS2 are susceptible to SARS-CoV-19
* Look at the residues of interest and count the frequency of each
  + Will give an assumption of the cleavage efficiency for TMPRSS2
* See the mutation rate among the species of interest
  + Will track the evolution of TMPRSS2 which will give an estimate for its ability to alter cleavage potential
* If there is a change in amino acid, what kind of change is it (volume, hydrophobicity, percent buried

Code

* I want to compare two alignment sequences to return both similarities and differences with their positions labeled
* I want to pull a particular *species*  from the main list of species and place it as a variable
* Run a multisequence alignment on the species to find the consensus sequence in the Desired species list (DS)
* Take the position of the DS that matches with the positions of the trimmed Human sequence
* Convert the DS consensus sequence to a SeqRecord to compare to the human sequence
* Run a multiple sequence alignment of both DS and human
* Find the frequency of changes at specific sequence locations compare to the human sequence
  + Find out if these changes result in changes in amino acid properties (specifically hydrophobic to hydrophilic or basic to acid)
  + Also find other differences in the amino sequence such as changes in disulfide bonding
* Find the similarities in the two sequences and record the position of those similarities