The Role of Machine Learning in SARS-CoV-2 Susceptibility Classification

Zak White, 2022

Introduction / Motivation

Machine learning in bioinformatics

- Application of machine learning to classify various host species as susceptible or immune to SARS-CoV-2

Validating the results of an existing study

Introduction / SARS-Cov-2

COVID-19 is caused by the virus SARS-CoV-2

- Likely of zoonotic origin
- Some non-human animals are known to be susceptible; some are not

Introduction / ACE2

The SARS-CoV-2 spike protein's target in potential hosts is the ACE2 protein (angiotensin-converting enzyme 2)

- Differences in ACE2 sequences can result in different interaction with the spike protein, or possibly inhibit interaction altogether

Introduction / ML in bioinformatics

Artificial intelligence (AI) and machine learning (ML) are used in various fields for planning, decision-making, and process efficiency

ML has been able to outperform humans in certain analytical tasks within the field of biology

- Kakadiaris et al. (2018): SVM model for classifying risk of cardiovascular disease

The use of ML models (especially blackbox models) without a biology-driven explanation has been criticized. (Rudin, 2019)

Introduction / Sequence classification

Proteins are biomolecules made up of amino acids

Sequence of acids can be mapped to a feature vector for use in classification problems

- Features are categorical and nonordinal
- This problem is the basis of this study

```
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWS
AFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTILNTMSTIYSTGKVCNPDNPQE
CLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLYEEYVVLKNEMARANHYEDYGDYWRGDYEVN
GVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLYS
LTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAWD
LGKGDFRILMCTKVTMDDFLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHLKS
IGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEMKREIVGVVEP
VPHDETYCDPASLFHVSNDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLFNMLRL
GKSEPWTLALENVVGAKNMNVRPLLNYFEPLFTWLKDQNKNSFVGWSTDWSPYADQSIKVRISLKSALGD
RAYEWNDNEMYLFRSSVAYAMRQYFLKVKNQMILFGEEDVRVANLKPRISFNFFVTAPKNVSDIIPRTEV
EKAIRMSRSRINDAFRLNDNSLEFLGIQPTLGPPNQPPVSIWLIVFGVVMGVIVVGIVILIFTGIRDRKK
KNKARSGENPYASIDISKGENNPGFQNTDDVQTSF
```

$$x = (M, S, S, S, W, ..., D, V, Q, T, S, F)$$

Introduction / Objective

Objective:

- Compare the performance of ML-based classification models of susceptibility with the performance of a biology-driven AI model
  - ML-based baseline model
  - ML-based eliminative model
  - Bio-based structural model

Hypothesis:

- Expect the biology model to outperform the ML models

Introduction / Related Work

Aldraimli et al. (2020)

- Designed 36 ML models for classifying patients' level of susceptibility to visceral fat-associated diseases based on discretized visceral adipose tissue measurements.
- The best models reach over 75% classification accuracy

Luan et al. (2020)

- Performed a structural analysis to identify influential sites and used the results to predict susceptibility to SARS-CoV-2 of 42 species based on ACE2 sequences

Zhao et al. (2020)

- Performed structural and phylogenetic analyses to predict susceptibility

Methods / Data

41 ACE2 sequences were collected from the NCBI protein database, each identified as susceptible or immune (14 immune and 27 susceptible), based on:

- investigations into non-human primates as models for human infection;
- lab studies that assessed binding affinities between the SARS-CoV-2 spike protein and ACE2 in various hosts;
- studies focusing on transmission between humans and livestock, house pets, and city animals; and
- reports on city and zoo animals contracting the virus.

Each was aligned with the human sequence to standardize indexing using Needleman-Wunsch

Methods / Baseline model

A baseline classification model was created using a feature importance algorithm which scored the acids at each index as a feature.

- Univariate feature selection scored with chi-squared tests with 97.5% confidence threshold

Results used to create categorical binary decision tree

Methods / Eliminative model

Classification model designed based on the results of an existing study (White, 2021) that isolated mutations that appeared exclusively in immune sequences.

Iterative sequence comparison algorithm

- Compared acids at each positions
- Assigned weights to mutations based on number of mutations in the sequence
- Summed weights to identify most influential mutations

```
for each negative sequence S⁻ do
    for each i < length of S⁻ do
        isInfluentialIndex ← True;
        for each positive sequence S⁺ do
            if i^th acid in S⁻ == i^th acid in S⁺ then
                isInfluentialIndex ← False;
            end
        end
        if isInfluentialIndex then
            // The mutation may be influential
            Record i, i^th acid in S⁻
        end
    end
end
```

Methods / Structural model

A third model was designed by further investigating the results of the original study with a manual structural analysis

- The structure of each mutation was compared with that of the original sequence
- The mutations that had no effect on the structure were considered extraneous and removed from the new model

Results / ML approaches

Baseline: various mutations at six sites were selected

Pos	Acids that impede susceptibility
83	Phenylalanine (F)
211	Aspartate (D)
212	Alanine (A), Aspartate (D), Serine (S)
246	Arginine (R)
251	Valine (V)
687	Not Alanine (A) or Serine (S)

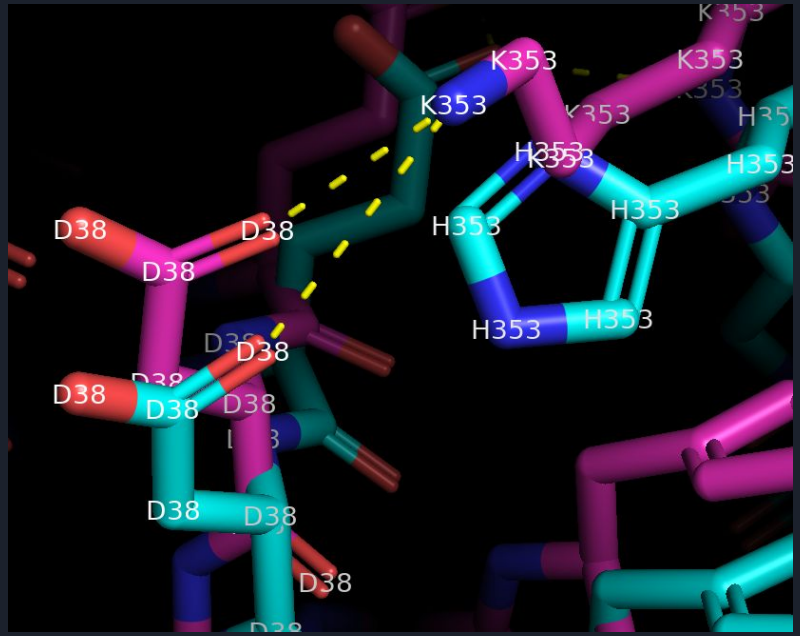
Eliminative: eight mutations were identified

Pos	Mutation		
31	Lysine (K)	→	Aspartate (D)
41	Tyrosine (Y)	→	Alanine (A)
66	Glycine (G) Arginine (R)	⇒	Alanine (A)
83	Tyrosine (Y)	→	Phenylalanine (F)
113	Serine (S) Arginine (R)	⇒	Asparagine (N)
353	Lysine (K)	→	Histidine (H)
426	Proline (P)	→	Serine (S)
679	Isoleucine (I)	→	Valine (V)

Results / Structural approach

Based on the structural analysis, the structural model only focused on five mutations.

Pos	Mutation		
31	Lysine (K)	→	Aspartate (D)
83	Tyrosine (Y)	→	Phenylalanine (F)
113	Serine (S) Arginine (R)	⇒	Asparagine (N)
353	Lysine (K)	→	Histidine (H)
426	Proline (P)	→	Serine (S)



Results / Sites of focus

The adjacent table shows which residues were used in each model, and compares them with the UniProt database's ACE2 mutagenesis overview.

- It is clear the structural model was more true to the findings outlined on UniProt, and the baseline model was least similar

	Model			
Pos	**B**	**E**	**S**	**UniProt**
31		×	×	×
41		×		×
66		×		
83	×	×	×	×
113		×	×	
211	×			
212	×			
246	×			
251	×			
353		×	×	×
426		×	×	×
679		×		
687	×			

Results / Classification metrics

Classification confusion matrices were made for each model based on testing subsets. The results were used to compute accuracy, sensitivity (true positive rate), and specificity (true negative rate)

	+	-
+	5	1
-	1	1

(a) Baseline Model

	+	-
+	14	3
-	3	6

(b) Eliminative Model

	+	-
+	24	3
-	5	9

(c) Structural Model

Results / Classification metrics

Classification confusion matrices were made for each model based on testing subsets. The results were used to compute accuracy, sensitivity (true positive rate), and specificity (true negative rate)

Model	Sensitivity	Specificity	Accuracy
Baseline	0.83	0.50	0.75
Eliminative	0.82	0.67	0.77
Structural	0.89	0.64	0.80

The structural model had the highest classification accuracy and sensitivity, and the eliminative model had the highest specificity.

Discussion

The mutation at site 113 (S → N) was used in both the eliminative and structural model

- The mutation affect the polar bond between residue 113 and the Asparagine (N) at site 117

This mutation was not referenced in the UniProt overview and could be a mutation that affects susceptibility, or an extraneous finding.

	Model			
Pos	**B**	**E**	**S**	**UniProt**
31		×	×	×
41		×		×
66		×		
83	×	×	×	×
113		×	×	
211	×			
212	×			
246	×			
251	×			
353		×	×	×
426		×	×	×
679		×		
687	×			

Discussion (2)

Predicting future sequences:

Species	B	E	S
Chinchilla (*chinchilla lanigera*)	×		
Grizzly bear (*ursus arctos horribilis*)	×	×	×
Black flying fox (*pteropus alecto*)		×	×
Sumatran orangutan (*pongo abelii*)	×	×	×
Tasmanian devil (*sarcophilus harrisii*)			
Orca (*orcinus orca*)	×	×	×
Turkey (*meleagris gallopavo*)			
Leatherback sea turtle (*dermochelys coriacea*)			

Discussion (3)

Although the results were only marginal, the structural model did prove to have a greater classification accuracy.

- Having the model be explainable from a biological point is more intuitive
- The structural model placed a greater focus on critical residues than other models

A more well-trained and -tested model that combined ML and bioinformatics could be helpful in predicting susceptibility to the virus when it is not already known.

Discussion / Limitations

Limited amount of data was a main restriction

- Less accurate models
- Less confidence in performance metrics

Data is categorical and non-ordinal

- Severely restricts the variety of classification models that can be used for the problem

Discussion / Future work

Variations on the problem

- Multinomial classification: degree of susceptibility
- Multi-label classification: susceptible, symptomatic, transmittable
- Multidimensional feature space: categorizing acids

More involved structural analysis could improve the performance of the model (especially specificity)

- Effect of mutations in consecutive positions

Conclusion

I satisfied my hypothesis! This study was

- An interesting validation study of my work from last year
- An introduction about the caveats of biology applications of ML
- An opportunity for me to familiarize myself with more concepts in bioinformatics

References

- Aldraimli et al. (2020), *Machine learning prediction of susceptibility to visceral fat associated diseases*
- Kakadiaris et al. (2018), *Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA*
- Cynthia Rudin (2019), *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*
- Luan et al. (2020), *Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection*
- Zhao et al. (2020), *Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2*
- UniProt, *ACE2 - Angiotensin-converting enzyme 2 precursor - Homo sapiens (Human) - ACE2 gene & protein*
- White (2021), *Identifying Mutations in ACE2 That Influence Susceptibility to SARS-CoV-2*

Questions

Thanks for listening!