Link prediction of viral spike proteins and cell receptors using structural perturbation method

Iris Zhou

Ladue Horton Watkins High School, 1201 South Warson Rd, St. Louis, MO 63124

October 2021

1 Abstract

Many protein receptors for animal and human viruses have been discovered in decades of studies. The main determinant of virus entry is the binding of the viral spike protein to host cell receptors, which mediates membrane fusion.

In this work, a bilayer network is constructed by integrating the similarity network of the viral spike proteins, the similarity network of host receptors, and the association network between viruses and receptors. The structural perturbation method (SPM) is used to predict possible emerging infection of a virus in potential new host organisms. The reliability of this method is based on the hypothesis that the major barrier to virus infection is the differences in the compatibility of spike proteins and cell receptors, which is determined by the amino acid sequences among species.

2 Introduction

With the rising threat of viral pandemics in the past 50 years – as illustrated by SARS-COV [14], COVID-19 [9, 18], Ebola [1], Zika [15] and HIV [4] – questions have arisen about how we can predict similar large-scale outbreaks to have time to avoid them or reduce their impact on daily life. The majority of these viral infections emerge first in another species (for instance, SARS-CoV-2 most likely originated in bats [19, 18]) before evolving to cross species barriers and expanding their life cycles in other hosts, including humans [6]. Currently, genome sequencing and computational analysis are common techniques used to identify and predict which pathogens are most likely to evolve to infect humans [3]. Specifically, the gene sequences coding the amino acid chains of viral spike proteins of different viruses and the genes of corresponding host cell receptors can be compared to identify their similarities and compatibility. Spike proteins are transmembrane glycoproteins (composed of an amino acid sequence with a carbohydrate side chain) folded into a spike shape to surround viruses [12, 14]. This "spike" shape is especially prominent in coronaviruses, hence their name. The spike proteins of viruses attach to human cell receptors and if compatible, will trigger the fusing of the viral and host cell membranes, allowing the viral genome to sneak into the cell uninhibited [12, 14]. Clearly, the spike proteins and the receptors interact to play a key role in the success of a virus in infecting a host. The spike protein is also a main element of diversity between different viral strains [12]. As new viruses or viral variants develop, their spike proteins evolve to changing environmental conditions, such as increased immune response (natural or vaccine-induced) or the use of medication inside a host's body.

Here I postulate that a virus from a reservoir has the ability to cross species barriers and adapt to a new host if the similarity between receptor proteins of the potential host and its original host is high enough. Based on this assumption, I present an approach to apply the structural perturbation method (SPM) [13] to predict the propensities of viral cross-species infections.

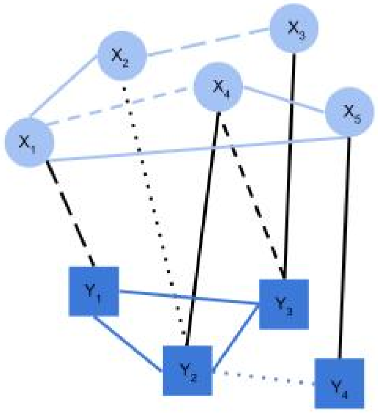
3 Computational Problem Formulation

In this work, I model the prediction of potential viral infection through prediction of the links of a graph, as described below. The similarities of the sequences of both the spike proteins and the receptors of different species are graphed onto an adjacency matrix. In addition to metrics for both virus-virus and receptor-receptor similarities, virus-receptor compatibility characterized in known viral infections is included in the adjacency matrix as well. Thus, since both animal and human cell receptors will be included in the adjacency matrix, it will allow for the identification of viral infections in animals that have yet to infect humans, but may do so in the future. This matrix can be analyzed using a linear algebra technique called structural perturbation method (SPM) [13]. Because this matrix will inevitably include unknown values – for virus-receptor interactions that haven't been studied yet – it can be diagonalized with the use of eigenvalues, after a set of data is removed, to predict the unknown values. These values can be used to predict viruses that are most likely to cross species and infect humans in the future.

4 Methods

4.1 Forming an Adjacency Matrix from a Graph

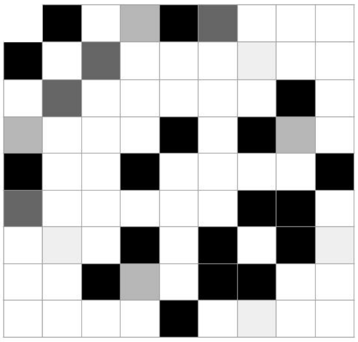
An adjacency matrix of a graph is formed by labeling each vertex of the graph with a number, which will then become the rows and columns of the matrix. In a simple unweighted graph, each (v_i, v_j) is marked with either 0 or 1, depending on whether the two vertices are linked or not. In weighted graphs, each (v_i, v_j) is marked with the strength of the relationship between i and j, usually on a scale from 0 to 1. For graphs with no self-loops, the main diagonal of the matrix contains all zeroes.



Using the above graph as an example, a 9x9 adjacency matrix can be generated by labeling each row (and column) $(x_1, x_2, ...x_5, y_1, ...y_4)$. By assigning weights based on the proximity of the dashed lines, the following matrix A can be produced:

$$\begin{bmatrix} 0 & 1 & 0 & 0.5 & 1 & 0.75 & 0 & 0 & 0 \\ 1 & 0 & 0.75 & 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0.75 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0 & 1 & 0 & 1 & 0.5 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0.75 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0.25 & 0 & 1 & 0 & 1 & 0 & 1 & 0.25 \\ 0 & 0 & 1 & 0.5 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0.25 & 0 & 0 \end{bmatrix}$$

This can also be represented visually with the following plot (darker colors symbolize a stronger relationship in the graph, or a higher weight in the matrix):



4.2 Eigenvalues, Diagonalization and Spectral Decomposition

If A is an nxn matrix in a field F, then the eigenvalue λ in F and the (nonzero) eigenvector $\vec{v}$ in F^n are defined such that

$$A\vec{v} = \lambda\vec{v} \tag{1}$$

Manipulating this equation, we get:

$$(A - I\lambda)\vec{v} = 0 \tag{2}$$

where I is the identity matrix.

Thus, if $\vec{v}$ is nonzero, this equation can only hold if:

$$\det(A - I\lambda) = 0 \tag{3}$$

All λ (both real and complex) that satisfy the above equation (referred to as the characteristic equation of A) are the eigenvalues of the matrix. Since the equation is of the nth order, there must be n total eigenvalues $[\lambda_1, ...\lambda_n]$. Each λ_i can be substituted back into (2) to obtain the corresponding eigenvectors $\vec{v}$. By putting all of the n eigenvalues and eigenvectors back into (1), the following equation is obtained:

$$A[\vec{v_1}, \vec{v_2}...\vec{v_n}] = [\lambda_1\vec{v_1}, \lambda_2\vec{v_2}...\lambda_n\vec{v_n}] = [\vec{v_1}, \vec{v_2}...\vec{v_n}] \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & \lambda_n \end{bmatrix} \tag{4}$$

If the definitions $\Lambda = diag[\lambda_1, \lambda_2, ...\lambda_n]$ and $V = [\vec{v_1}, \vec{v_2}...\vec{v_n}]$ are employed, a more compact form of the equations can be produced:

$$AV = \Lambda V \tag{5}$$

$$A = V\Lambda V^{-1} \tag{6}$$

This is called the spectral decomposition of the matrix, where a real symmetric matrix can be written in terms of its eigenvalues and eigenvectors. Since the matrix of eigenvectors V is orthogonal (by definition, they form an orthonormal basis in R^N), then $V^{-1} = V^T$. Thus, equation (6) can be rewritten as:

$$A = V\Lambda V^T \tag{7}$$

In addition, for any m:

$$(\sum_n \lambda_n v_n v_n^T) v_m = \lambda_n v_n = A v_n \tag{8}$$

Hence:

$$A = \sum_n \lambda_n v_n v_n^T \tag{9}$$

4.3 Structural Perturbation Method

After generating an adjacency matrix, a random set of vertices must be perturbed (removed) and the removed values will be replaced with zeroes in the matrix. If the set of perturbed values is denoted by ΔE and the remaining set by E^R, and their weighted adjacency matrices are ΔA and A^R respectively, then $A = A^R + \Delta A$. Note that both the perturbed and remaining matrix contain the same dimensions as the original matrix; the missing values are simply set to zero.

Since A^R is a real symmetric matrix (based on how the adjacency matrix is formed), using equation (9) it can be diagonalized as follows:

$$A^R = \sum_{k=1}^{N} \lambda_k v_k v_k^T \tag{10}$$

where λ_k and v_k are respectively the eigenvalues and eigenvectors of each perturbation k of A^R, with N total perturbations.

After perturbation, the eigenvalue is adjusted to $\lambda_k + \Delta\lambda_k$ and the eigenvector to $v_k + \Delta v_k$. Thus, from equation (1):

$$(A^R + \Delta A)(v_k + \Delta v_k) = (\lambda_k + \Delta\lambda_k)(v_k + \Delta v_k) \tag{11}$$

Next, this equation is left multiplied by v_k^T. To make the problem easier, a linear approximation is utilized by neglecting the second order terms $v_k^T \Delta A \Delta v_k$ and $\Delta\lambda_k v_k^T \Delta v_k$. Thus:

$$\Delta\lambda_k \approx \frac{v_k^T \Delta A v_k}{v_k^T v_k} \tag{12}$$

This linear approximation allows the eigenvalues to change but keeps the eigenvectors constant (since all terms containing Δv_k^T are cancelled). Thus, the perturbed matrix can be rewritten as as:

$$A' = \sum_{k=1}^{N} (\lambda_k + \Delta\lambda_k) v_k v_k^T \tag{13}$$

However, this process neglects the degenerate case in which some eigenvalues are repeated. The repeated eigenvalues must be picked out and changed into perturbed unique eigenvalues. If the eigenvalues λ_{ki} are defined such that i indicates M-associated eigenvectors derived from the same eigenvalues and k indicates distinct eigenvalues, and the chosen repeated eigenvalues are defined as $v'_{ki} = \sum_{j=1}^{M} \beta_{kj} v_{kj}$, the equation (1) yields:

$$(A^R + \Delta A) v'_{ki} = (\lambda_{ki} + \Delta\lambda'_{ki}) v'_{ki} \tag{14}$$

Again using a linear approximation (neglecting second order terms), the following is obtained:

$$\Delta\lambda'_{ki} \sum_{j=1}^{M} \beta_{kj} v_{kj} = \sum_{j=1}^{M} \beta_{kj} \Delta A v_{kj} \tag{15}$$

For all $n = 1...M$, we can multiply equation (13) by v_{kn}^T to obtain:

$$\Delta\lambda'_{ki} \beta_{kn} = \sum_{j=1}^{M} \beta_{kj} v_{kn}^T \Delta A v_{kj} \tag{16}$$

This condenses to the matrix form:

$$W B_k = \Delta\lambda'_k B_k \tag{17}$$

where W is an MxM matrix defined by $W_{nj} = v_{kn}^T \Delta A v_{kj}$ and B_k is the column vector of β_{kj} (essentially the adjusted eigenvector). Noticing the parallels between equation (17) and equation (1), one realizes that the perturbed matrix becomes:

$$A' = \sum_{k=1}^{N} (\lambda_k + \Delta\lambda'_k) v'_k v_k^T \tag{18}$$

This perturbed matrix A' can be computed for as large N as necessary, dividing by N to obtain the average (in order to maintain the same scaling and structural consistency). For each perturbation and subesequent diagonalization, a set number or percentage of perturbed values must be used. This perturbed matrix can be used to make a prediction about the value of certain unknown links. The method described in this section is known as the structural perturbation method (SPM) [13].

5 Results

5.1 Data set

As mentioned in the introduction, the techniques explained in section 4 will be applied to determine the compatibility of a set of viruses and protein receptors. For this specific application, the amino acid sequences of 22 viral spike proteins, along with all of the known sequences of the 18 receptors (40 from animal hosts and 20 from human hosts) were collected [5]. These are listed below in Table 1.

All of the spike proteins, and all of the the receptor proteins, were then compared in pairs based on sequence alignment using MUSCLE [7]. The alignment results were used to calculate the sequence similarity, using absolute and relative distance. The absolute distance is an estimate of the evolutionary divergence between sequences and is defined as the ratio of the number of aligned amino acid residues to the alignment length. The relative distance is the pairwise distance divided by the maximum distance value calculated from the distance analysis results. In this application, the relative distances were used as the weights of the links between viral spike proteins and the links between receptor proteins. The data behind the links between viral spike proteins and receptors are taken from 60 known pairs of virus-receptor interactions [5]. Thus, after generating a graph, a matrix is created using the method demonstrated in Section 4.1.

virus	receptor
HIV	CD4
Hantaviruses	Integrin $\alpha v \beta 3$
Foot-and-mouth disease virus	Integrin $\alpha v \beta 3$
SARS-CoV-1	ACE2
SARS-CoV-1 (bats)	ACE2
Rabies virus	nAchR
Echovirus (E-6, E-7, E-11, E-12, E-20, E-21 and E-70)	CD55
Coxsackievirus A and B (CV-A21, CV-B1, CV-B3 and CV B5)	CD55
HCoV-229E (severe acute respiratory syndrome-associated coronavirus)	APN
Vesicular stomatitis virus	PS receptor
Encephalomyocarditis virus	VCAM1
Hepatitis A virus	HAVCR1
Measles virus vaccine strains	CD46
Measles virus wild-type strains	SLAM
MERS coronavirus	DPP4
Nipah virus	Ephrin B2
Nipah virus	Ephrin B3
Lassa virus	DAG1
Lymphocytic choriomeningitis virus	DAG1
Junin arena virus	TRFC
Machupo virus	TRFC
SARS-CoV-2	ACE2
Sendai virus	ASGR2

Table 1: List of the 18 virus receptors and 22 viruses

5.2 Evaluation of Structural Perturbation Method

The accuracy of the structural perturbation method outlined in section 4.3 must be verified in the context of the specific application, before it can be used on the data set for prediction. In order to do so, we perform a 5-fold cross-validation [10]. In this method, all of the 60 known virus-receptor interactions are randomly divided into 5 equally sized subsets. Four of these groups are used training data; the other is the test data. To determine its accuracy, SPM will be performed and its results will be compared with the already known information regarding the virus-receptor interactions using the following metrics:

$$precision = \frac{TP}{TP + FP} \tag{19}$$

$$recall = \frac{TP}{TP + FN} \tag{20}$$

where TP, FP and FN are respectively the number of true positive, false positive and false negative samples, with respect to a virus-receptor interactions.

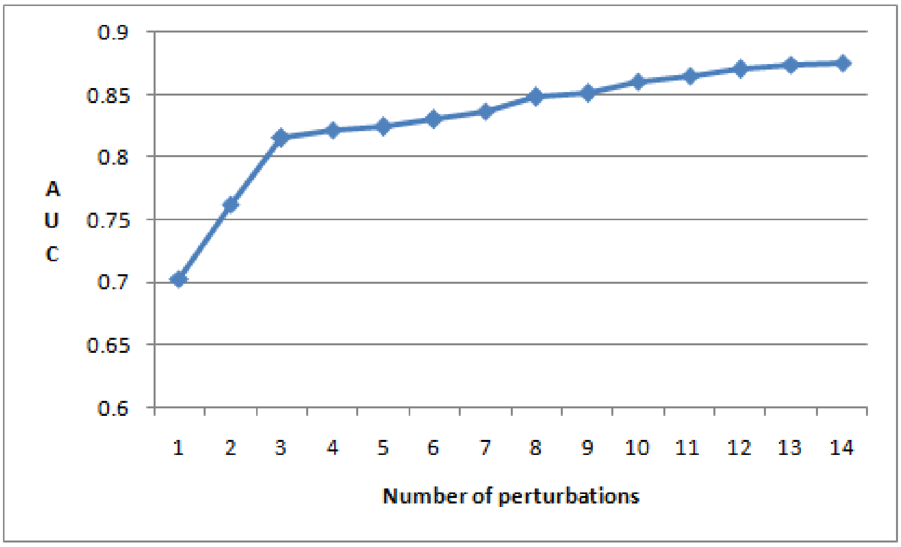
The receiver operating characteristic curve (ROC) curve is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true positive rate (TPR) and the false positive rate (FPR) are calculated by the following formula [16]:

$$TPR = \frac{TP}{TP + FN} \tag{21}$$

$$FPR = \frac{FP}{TN + FP} \tag{22}$$

The area under the ROC curve (AUC) – calculated via integration – represents the overall success of the structural perturbation method in predicting the correct outcomes [2].

Using the previously mentioned data set of 60 known virus-receptor interactions split into 5 groups, the plot showing AUCs of different numbers of perturbations is shown below.



As the plot illustrates, the accuracy plateaus between 85% and 90% once 8 perturbations have been performed. Since 50% represents a worthless test and 100% represents a perfect test, this statistic demonstrates an accurate prediction rate for the application.

5.3 Results

Finally, the structural perturbation method outlined in section 4.3 was applied to the full data set described in section 1. This was calculated using the Matlab software of SPM [13]. Table 2 shows cases where there is a high (greater than 90 percent) chance of cross-species host susceptibility to viruses in the prediction results. Interestingly, SPM predicted that SARS-CoV-2 would infect humans. However, if all data regarding SARS-CoV spike proteins are excluded, then the link between SARS-CoV-2 and its receptor ACE2 could not be predicted.

virus	host 1 (known)	host 2 (predicted)	weight of predicted link
MERS-CoV	C. dromedarius	H. sapiens	0.99
VSV	B. taurus	H. sapiens	0.99
SARS-CoV	F. catus	H. sapiens	0.99
SARS-CoV (Rm1/2004)	Chiroptera	H. sapiens	0.98
HIV (SIV;Lentivirus)	P. troglodytes	C. aethiops	0.98
HIV (SIV;Lentivirus)	C. aethiops	H. sapiens	0.94
Hantavirus	M. musculus	R. norvegicus	0.93
Lassa virus, LCMV	R. norvegicus	H. sapiens	0.92
Hantavirus	R. norvegicus	H. sapiens	0.91
Lassa virus, LCMV	unknown	H. sapiens	0.91
Measles virus wild-type strains	H. sapiens	M. mulatta	0.91
Nipah virus	S. scrofa	H. sapiens	0.90
Nipah virus	M. brandtii	H. sapiens	0.90
Rabies virus	C. lupus familiaris	H. sapiens	0.90
SARS-CoV-2	unknown	H. sapiens	0.90

Table 2: cross-species infection propensity

5.4 Discussion

At this moment, the COVID-19 pandemic is still ongoing all over the world. It is believed that there is a reservoir of viruses in wild animals, since increasing numbers of viruses have been discovered in wild animals. With more sequences of viruses and their receptors available, computational prediction models such as the structural perturbation method used in this work can be applied as an early warning system to prevent infections and propagation in humans.

Although many other proteins also contribute to the process of virus production and host invasion, the spike protein located on the surface of the virus plays the most important role in the binding of the cell receptor and membrane fusion [8, 11, 17]. Therefore, it is the most important factor to determine host range [14, 8, 11, 17]. In this work, with the sequences of SARS-COV spike protein and its receptor included, SPM model successfully predicted that SARS-CoV-2 has the potential to cross species barriers to infect humans.

Cross-validations showed that the accuracy of SPM only reaches close to 90%. Clearly, there is still room to improve. In addition to sequence similarity, other features of spike proteins and their receptors can be considered. Physicochemical properties, secondary structural motifs and compositional information of animo acids or dipeptides of viral spike proteins and cell receptors are all candidates to be considered in the future.

6 Conclusion

In this paper, a bilayer network, which integrated the similarity network of the viral spike proteins, the similarity network of host receptors, and the association network between viruses and receptors, was analyzed using the structural perturbation method. Through SPM, I managed to compare the links between several viruses, animal receptors and human receptors to predict the compatibility between certain viral spike proteins and human receptors, which can determine which viruses are most likely to cause the next major outbreak. Computing techniques like SPM should be harnessed to act as an early detection system for future propensity of viral infections in humans.

7 Acknowledgments

This work was initiated for a Calculus 3 course project. I appreciate my teacher, Dr. John Pais, for his guidance. I also appreciate Professor Ting Wang from the Department of Genetics, and Professor Nan Lin from the Department of Mathematics and Statistics in Washington University in St. Louis for their suggestions and advice for improvement of this work.

References

[1] L. Baseler et al. “The Pathogenesis of Ebola Virus Disease.” In: *Annual Review of Pathology* 12 (Jan. 2017).

[2] A. P. Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern Recognition* 30.7 (July 1997).

[3] A. Calistri et al. “New generation sequencing in pathogen discovery and microbial surveillance”. In: *Expert Review of Anti-infective Therapy* 11.9 (Feb. 2013).

[4] B.A. Castro et al. “HIV heterogeneity and viral pathogenesis”. In: *AIDS* 2.suppl 1 (1988).

[5] M Cho and H. S. Son. “Prediction of cross-species infection propensities of viruses with receptor similarity”. In: *Infection, Genetics and Evolution* ().

[6] E. Domingo. “Mechanisms of viral emergence”. In: *Veterinary Research* 41.6 (Nov. 2010).

[7] R. C Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5 (Mar. 2004).

[8] T. Heald-Sargent and T. Gallagher. “Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence.” In: *Viruses* 4.4 (2012).

[9] B. Hu et al. “Characteristics of SARS-CoV-2 and COVID-19”. In: *Nature Reviews Microbiology* 19 (Oct. 2020).

[10] P. A Lachenbruch and M. R. Mickey. “Estimation of error rates in discriminant analysis”. In: *Technometrics* 10.1 (Feb. 1968).

[11] M. Letko, A. Marzi, and V. Munster. “Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses.” In: *Nature MIcrobiology* 5 (Feb. 2020).

[12] F. Li. “Structure, Function, and Evolution of Coronavirus Spike Proteins”. In: *Annual Review of Virology* 3 (Sept. 2016).

[13] L. Lu et al. “Toward link predictability of complex networks”. In: *Proceedings of the National Academy of Sciences of the United States* 112.8 (Feb. 2015).

[14] V.D. Menachery et al. “A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence”. In: *Nature Medicine* 21.12 (Dec. 2015).

[15] D. Musso and D.J. Gubler. “Zika Virus”. In: *Clinical Microbiology Reviews* 29.3 (July 2016).

[16] K. Spackman. “Signal detection theory: Valuable tools for evaluating inductive learning”. In: *Pro-ceedings of the Sixth International Workshop on Machine Learning* (1989).

[17] D. Wrapp et al. “Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.” In: *Science* 367.6483 (Mar. 2020).

[18] A.G. Wrobel et al. “SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects”. In: *Nature Structural Molecular Biology* 27 (July 2020).

[19] X. Zhou P. an Yang, Wang x., and et al. “A pneumonia outbreak associated with a new coronavirus of probable bat origin”. In: *Nature* 579.7798 (Mar. 2020).