Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 73 (2019) 71–80

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Prediction of cross-species infection propensities of viruses with receptor similarity

Myeongji Cho[a], Hyeon Seok Son[a,b,*]

[a] Laboratory of Computational Biology & Bioinformatics, Institute of Health and Environment, Graduate School of Public Health, Seoul National Uniersity, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[b] Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

ARTICLE INFO

Keywords:
Cross-species infection
Discriminant analysis
Amino acid substitution
Evolutionary distance
Sequence similarity

ABSTRACT

Studies of host factors that affect susceptibility to viral infections have led to the possibility of determining the risk of emerging infections in potential host organisms. In this study, we constructed a computational framework to estimate the probability of virus transmission between potential hosts based on the hypothesis that the major barrier to virus infection is differences in cell-receptor sequences among species. Information regarding host susceptibility to virus infection was collected to classify the cross-species infection propensity between hosts. Evolutionary divergence matrices and a sequence similarity scoring program were used to determine the distance and similarity of receptor sequences. The discriminant analysis was validated with cross-validation methods. The results showed that the primary structure of the receptor protein influences host susceptibility to cross-species viral infections. Pair-wise distance, relative distance, and sequence similarity showed the best accuracy in identifying the susceptible group. Based on the results of the discriminant analysis, we constructed ViCIPR (http://lcbb3.snu.ac.kr/ViCIPR/home.jsp), a server-based tool to enable users to easily extract the cross-species infection propensities of specific viruses using a simple two-step procedure. Our sequence-based approach suggests that it may be possible to identify virus transmission between hosts without requiring complex structural analysis. Due to a lack of available data, this method is limited to viruses whose receptor use has been determined. However, the significant accuracy of predictive variables that positively and negatively influence virus transmission suggests that this approach could be improved with further analysis of receptor sequences.

1. Introduction

Over the past 50 years, new infections caused by pathogens such as human immunodeficiency virus (HIV), Ebola virus, severe acute respiratory syndrome coronavirus (SARS-CoV), H5N1 avian influenza virus, antibiotic-resistant *S. aureus*, and antibiotic-resistant *Mycobacterium tuberculosis* have emerged worldwide (Snowden, 2008; Jones et al., 2008). The majority of these new infections are caused by infectious agents crossing species barriers and completing their life cycle with expanded host ranges, a process that is influenced by diverse parameters (Domingo, 2010; Woolhouse and Gowtage-Sequeria, 2005). Given the necessity and importance of early detection and response to potential threats, experts from various fields have attempted to address this problem, although it has yet to be solved. For decades, emergent viruses have been studied using both classical methods of virology and genome-based technologies. Recently, computational approaches, including bioinformatics, have also been used, such as genome sequencing, construction of databases and analysis systems, and development of models and software to predict emerging infections (Rappuoli, 2004; Haagmans et al., 2009; Pepin et al., 2010; Morse et al., 2012; Woolhouse et al., 2012). There have also been studies investigating genomic patterns of receptor proteins or receptor-binding domains that influence host susceptibility to viral infections, which have enabled the discrimination of infection propensities based on the primary structure of receptors without requiring complicated structural analysis (Rogers et al., 1983; Matrosovich et al., 2000; Graham and Baric, 2010; Bae and Son, 2011; Imai and Kawaoka, 2012). The cell-surface proteins used as receptors by viruses have been identified (Schneider-Schaulies, 2000; Dales, 1973; Grove and Marsh, 2011). However, determining the biological parameters that influence virus–receptor interactions is problematic because the virus is different from the natural ligands or substrates of the receptors (Dimitrov, 2004;

* Corresponding author at: Laboratory of Computational Biology & Bioinformatics, Institute of Health and Environment, Graduate School of Public Health, Seoul National Uniersity, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea.
E-mail address: hss2003@snu.ac.kr (H.S. Son).

https://doi.org/10.1016/j.meegid.2019.04.016
Received 24 June 2018; Received in revised form 21 February 2019; Accepted 19 April 2019
Available online 23 April 2019

Skehel and Wiley, 2000). In addition, the details of the entry, penetration, and uncoating of some viruses are unclear. This is gradually being overcome by the increasing availability of virus genetic information and the ongoing advancements in computerized data-processing and sequence-analysis methods (Dales, 1973; Dimitrov, 2004). In the present study, the distance and similarity of receptor proteins that function as a species barrier to viral infection were calculated and used to predict the propensity for cross-species viral infections between host species. With the aim of estimating the capacity of viruses to affect the emergence of new viral diseases, we developed a computational framework to predict the cross-species infection propensities of viruses using receptor sequences. We postulated that a virus from a reservoir might be able to adapt to a new host only if the similarity between receptor proteins present in the potential host species is high enough for them to cross the species barrier. Evolutionary divergence matrices were used to calculate distance scores of target sequence pairs considering amino acid substitutions, and overall sequence similarities were computed using Java programming.

2. Materials and methods

2.1. Data sets

Receptor sequences of 98 species of source organisms were collected from the NCBI protein database. In total, 277 amino acid sequences for 18 types of receptor proteins were collected by choosing non-partial sequences. Accordingly, 18 receptor data sets consisting of orthologous amino acid sequences derived from different species were constructed (Table 1). Multiple sequence alignment was performed on collected sequences with MUSCLE using the default parameters for each receptor data set (Edgar, 2004). Using the alignment results, an original data set was constructed to train, validate, and test build a classifier. For each receptor data set, all possible sequence pairs were generated to construct the original data set, which is the source of variables for the classification model. Sequence variation in receptor proteins is relevant in the conformational differences of virus–host interaction interfaces and in protein expression, which may strongly influence viral infection propensity. With the aim of confirming the effect of receptor similarity on host susceptibility to cross-species infection, evolutionary distances and sequence similarities were calculated and used to classify groups of infection propensity from 50 sequence pairs of training sets. A test data set was constructed with 6 sequence pairs that were randomly split from the original data for each group ratio. Of the original datasets (consisting of 56 sequence pairs), 50 were used for model construction using discriminant analysis, and 6 were used for validation using the constructed models. Two groups were classified based on susceptibility to viral infection: Group 1, which contained sequence pairs with high similarity, which increases cross-species infection propensity, and Group 2, which contained sequence pairs with low similarity, which decreases cross-species infection propensity. The possibility of virus–host-cell interactions, which confer host susceptibility to viral infection, was determined using a database search and literature review for each virus with specific receptor used. Viruses with unknown or unspecified host receptors were excluded from the original data set, and both zoonotic and non-zoonotic viruses were included. Each receptor data set had at least one sequence derived from the natural host (primary reservoir). The numbers of sequence pairs for each group in the original data set were 36 and 20, respectively. The frequencies of the groups in the training data sets were 32 and 18, and those in the test data sets were 4 and 2, respectively.

Table 1
List of the 18 virus receptors used in this study.

Virus	Receptor
HIV	CD4
Hantaviruses, foot-and-mouth disease virus	Integrin αvβ3
SARS coronavirus	ACE2
Rabies virus	nAchR
Echovirus (E-6, E-7, E-11, E-12, E-20, E-21 and E-70), 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, Ephrin B3
Lassa virus, lymphocytic choriomeningitis virus	DAG1
Junin arena virus, Machupo virus	TRFC
Sendai virus	ASGR2

The amino-acid sequences of 18 receptors for 20 viruses were used to construct an original data set consisting of receptor sequence pairs. CD4: cluster of differentiation 4; CD61: cluster of differentiation 61; ACE 2: angiotensin converting enzyme 2; nAchR: nicotinic acetylcholine receptor; CD55: Complement decay-accelerating factor; aminopeptidase N; PS: Phosphatidylserine; VCAM1: vascular cell adhesion molecule 1; HAVCR1; Hepatitis A virus cellular receptor 1; CD46: cluster of differentiation 46; SLAM: signaling lymphocytic activation molecule; DPP4: Dipeptidyl peptidase-4; DAG1: Dystroglycan1; TRFC: transferrin receptor; ASGR2: asialoglycoprotein receptor 2.

2.2. Calculation of distance scores and sequence similarities

Classification of host pairs into infectivity groups was performed using evolutionary distance and sequence similarity based on sequence alignment. Multiple sequence alignments and optimal pairwise sequence alignments were performed on the collected sequences of each of the 18 receptor protein sets using MUSCLE (Edgar, 2004). A scoring matrix was used to represent the evolutionary distance of a sequence from the other sequence within the pair. MEGA6 software was used to calculate the distance (Tamura et al., 2013). The distance of a residue within a sequence was measured as the substitution score from the amino acid of the relevant column in a matched sequence. The matrix showed substitution scores for all possible sequence pairs within a receptor data set. Total sequence similarity was calculated using Java programming based on the alignment results. As a result of the distance and similarity analysis, we parameterized the absolute distance (pairwise distance), relative distance, and overall sequence similarity for each host pair in the data set. All of the host-pair data used to generate predictive variables were examined for infection characteristics by literature review. The absolute distance (pairwise distance, ${}^{g}S_{i,1}$) is an estimate of the evolutionary divergence between sequences and is defined as the number of amino acid substitutions between aligned sequences. The Poisson correction model was used as a substitution model and the complete deletion method was used to process gaps/missing data (Zuckerkandl and Pauling, 1965). The relative distance (${}^{g}S_{i,2}$) is the ratio of the pairwise distance to the maximum distance value calculated from the distance analysis results (pair-wise distance ÷ maximum distance within datasets). The total similarity (${}^{g}S_{i,3}$) is the result of a similarity analysis of all possible host-pairs in the dataset, which is the number of matched amino acids in an orthologue sequence to the total length of the aligned sequence including indels. These three independent variables were used to classify group members into the cross-species infectious or non-infectious group, and the decision coefficients obtained from the discriminant analysis were used to build a prediction model. In our study, similarity of predicted interaction hotspots with some amino acid residues in the receptor sequence had been considered a candidate predictive variable, but the significance of the discriminant analysis was low, and it was excluded from the prediction model. The scores for the three variables were defined as:

$$^{g}S_{i,1} = \textit{The number of amino acid substitutions per site from between sequences.}$$

$$^{g}S_{i,2} = \frac{pair - wise\ distance}{maximum\ distance\ within\ datasets}$$

$$^{g}S_{i,3} = \frac{n_{tot}}{N_{tot}} = \frac{total\ number\ of\ matched\ amino\ acids\ in\ the\ sequence}{total\ number\ of\ amino\ acids\ in\ one\ sequence\ string}$$

where $^{g}S_{i,1}$ is the score for the distance of the i^{th} (the number of variable sets) row of infectivity group g (), $^{g}S_{i,2}$ is the score for the relative distance of the i^{th} row of infectivity group g, and $^{g}S_{i,3}$ is the score for similarity of the i^{th} row of infectivity group g. The distance scores, which categorized absolute and relative scores, and the similarity scores of collected sequence pairs were stored in a MySQL database. All decision coefficients for three independent variables were calculated and used to construct a model for classifying group members using a discriminant function.

2.3. Discriminant analysis

IBM SPSS Statistics 24.0 (Cor, 2016; Green et al., 1996) and XLSTAT (Addinsoft, 2017) software were used for discriminant analysis to classify susceptibility to cross-species infection based on receptor similarity. Discriminant analysis is a form of multivariate analysis in which distinct sets of observations are classified according to previously defined groups, and a model is built from predictive variables and objects to allocate new observations to pertinent groups (Mika et al., 1999). The discriminant analysis method classifies an individual object into the group using the discriminant scores from linear discriminant function or the Mahalanobis distance from pooled covariance matrices of non-normally distributed sets (Mika et al., 1999). In this study, we used discriminant scores to estimate infection propensities considering that the differences in probability of cross-species infection of viruses between the two groups (infectious and non-infectious) should be maximally reflected and quantitatively expressed through the same index. Considering that the adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates (Rosenbaum and Rubin, 1983), we calculated propensity scores of all sequence pairs to estimate cross-species infection propensities between host species through the adjustment of observed covariates and determinant scores. The adjustment was performed by calibrating determinant scores of all sequence pairs using the first order function formula derived using the determinant score range and the centroid for each group. The z-score was used as a cut-off for classifying each group member into the infectious or non-infectious group. The coefficient of determination (discriminant coefficient) is estimated to maximize the difference between the groups, and is multiplied by each independent variable as a weighted score. The discriminant scores of all sequence pairs were calculated and the average discriminant scores were used to obtain the group centroid. The discriminant and propensity scores were stored in a MySQL database and used to generate a web application system.

2.4. Accuracy evaluation

In cross validation, each case was classified by the functions derived from all cases. Leave-one-out cross-validation was performed to analyze the accuracy of discriminant analyses, and the erroneous classification rates were verified (Lachenbruch and Mickey, 1968). A single object was first omitted from training, and the discriminant model was built with leave-one-out cross-validation. The omitted subject members were classified after training based on the built model. A total of 50 rounds of class predictions were performed, and the calculated confusion matrices were assessed. For the original and cross-validated grouped cases, the ratio of the number of correctly grouped members to total members was calculated, respectively. The performance of the discriminant model can be determined by calculating the sensitivity, specificity, total classification accuracy, and area under the receiver operating characteristic (ROC) curve (AUC). Performance scores were calculated from the confusion matrices for the training sample, validation samples, and cross-validation results (Foody, 2002). The sensitivity, specificity, total classification accuracy, and AUC were defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Total classification accuracy} = \frac{\text{The number of correctly predicted sequence pairs}}{\text{The total number of sequence pairs}}$$

$$\text{AUC} = \int_a^b f(x)dx$$

The accuracies of test data sets are presented as the ratio of the number of correctly predicted sequence pairs to the total number of test sequence pairs. Seven rounds of class predictions were performed in the test phase.

2.5. Design and implementation of the web application system

2.5.1. Data resources

The propensity scores calculated from the discriminant analysis using the original data set were used for the implementation of our web application system. We named the propensity score, which indicates the probability of viral infection between species, Infectindex. The calculated Infectindex values ranged from 0.001 to 99.994% in the total data set, from 0.001 to 36.946% in the non-infectious group, and from 68.294 to 99.994% in the infectious group.

We also collected and stored reference sequences for 24 virus genomes to generate the target sequence resources required to execute the Search Engine (Sequence Similarity Scoring System) according to the virus sequence query input. Accession numbers and viral species for reference genome sequences used in the study were as follows: NC_001479.1 (encephalomyocarditis virus), AF465516.1 (echovirus E7, human echovirus 7 strain Wallace), NC_004004.1 (foot-and-mouth disease virus), NC_005219.1 (Hantaan virus), NC_001489.1 (hepatitis A virus), NC_002645. 1 (human coronavirus 229E), NC_001802.1 (human immunodeficiency virus 1), NC_005080.1 (Junin virus segment L), NC_005081.1 (Junin virus segment S), NC_004297.1 (Lassa virus segment L), NC_004296.1 (Lassa virus segment S), NC_004291.1 (lymphocytic choriomeningitis virus segment L), NC_004294.1 (lymphocytic choriomeningitis virus segment S), NC_005079.1 (Machupo virus segment L), NC_005078.1 (Machupo virus segment S), NC_001498.1 (measles virus), NC_019843.3 (Middle East respiratory syndrome coronavirus), NC_002728.1 (Nipah virus), NC_001542.1 (rabies virus), NC_003436.1 (porcine epidemic diarrhea virus), NC_004718.3 (SARS coronavirus), NC_001552.1 (Sendai virus), NC_002306.3 (feline infectious peritonitis virus), and NC_001560.1 (vesicular stomatitis Indiana virus) (Table 2).

2.5.2. Database schema and web interface implementation

As a web server-based analysis tool, the web interface of ViCIPR was programmed to operate with a MySQL-based database to search information efficiently according to query input and to output calculation results. The data fields are divided into class (classified group), casenum (case number), receptor, virus, reservoir, host1 (primary/donor host), host2 (secondary/receipt host), pairwise_distance, relative_distance, total_similarity, g_verified (verified group), disg_predicted (predicted group by discriminant score), disds_calculated (calculated discriminant score), disds_trimmed (adjusted discriminant

Table 2
List of reference sequences of 24 virus genomes collected and stored for system construction.

Virus name	Organism	Strain	ncbi accession	bp length
EMCV	Encephalomyocarditis virus	Ruckert	NC_001479.1	7835
Human echovirus 7	echovirus E7	Wallace	AF465516.1	7427
FMDV	Foot-and-mouth disease virus - type O	Unknown	NC_004004.1	8134
Hantavirus	Hantaan orthohantavirus	Unknown	NC_005219.1	3616
Hepatitis A virus	Hepatovirus A	Unknown	NC_001489.1	7478
HCoV-229E	Human coronavirus 229E	229E	NC_002645.1	27,317
HIV (SIV; Lentivirus)	Human immunodeficiency virus 1	Unknown	NC_001802.1	9181
Junin virus	Junin mammarenavirus	Unknown	NC_005080.1NC_005081.1	7114 3411
Lassa virus	Lassa mammarenavirus	Josiah	NC_004297.1NC_004296.1	7279 3402
LCMV	Lymphocytic choriomeningitis mammarenavirus	Unknown	NC_004291.1NC_004294.1	6680 3376
Machupo virus	Machupo mammarenavirus	Carvallo	NC_005079.1NC_005078.1	7196 3439
Measles virus	Measles morbillivirus	Ichinose-B95a	NC_001498.1	15,894
MERS-CoV	Middle East respiratory syndrome-related coronavirus	HCoV-EMC	NC_019843.3	30,119
Nipah virus	Nipah henipavirus	Unknown	NC_002728.1	18,246
Rabies virus	Rabies lyssavirus	Unknown	NC_001542.1	11,932
SARS-associated CoV	Porcine epidemic diarrhea virus	CV777	NC_003436.1	28,033
SARS-CoV	SARS coronavirus	Unknown	NC_004718.3	29,751
Sendai virus	Murine respirovirus	Ohita	NC_001552.1	15,384
TGEV	Feline infectious peritonitis virus	Unknown	NC_002306.3	29,355
VSV	Vesicular stomatitis Indiana virus	Unknown	NC_001560.1	11,161

A reference sequences for 24 virus genomes were collected and stored to generate a target sequence resource for performing a similarity search engine (Sequence Similarity Scoring System) in ViCIPR according to a virus sequence query input. In constructing the gene database and the protein database, 118 genes and protein sequences were collected and processed to construct target sequence resources by parsing cds regions of 24 reference genome sequences.

Table 3
List of data in MySQL database.

Field name	Type	Description
Class	varchar (Graham and Baric, 2010)	Data set class of each case: 50 class for training data sets and 6 class for test data sets are designated for each case
Casenum	varchar(Graham and Baric, 2010)	Index number for database primary key: a1-a56, b1-b56 used to build training datasets and test datasets contain all data items with eliminating duplicate values
Receptor	varchar(50)	Receptor proteins: ACE2, APN, ASGR2, CD4, CD46, CD55, CD61, DAG1, DPP4, Ephrin B2, Ephrin B3, HAVCR1, Integrin alpha 5, NAchR, PS receptor, SLAM, TRFC, VCAM1
Virus	varchar(50)	Virus name: EMCV, human echovirus 7, FMDV, hantavirus, Hepatitis A virus, HCoV-229E, HIV (SIV; Lentivirus), junin virus, lassa virus, LCMV, machupo virus, measles virus, MERS-CoV, nipah virus, rabies virus, SARS-associated CoV, SARS-CoV, sendai virus, TGEV, VSV
Reservoir	varchar(50)	Information of reservoir hosts of viruses including *S. scrofa, B. Taurus, M. musculus, P. alecto, H. sapiens, P. troglodytes, C. dromedaries, M. brandtii, C. lupus familiaris, R. ferrumequinum, F. cattus, R. norvegicus, C. quinquefasciatus*
Host1	varchar(50)	29 species of donor host organisms
Host2	varchar(50)	29 species of recipient host species
pairwise_distance	Double	Pairwise distance score: 0.008–1.712 for the original data set, 0.375–1.712 for non-infectious group, and 0.008–0.36 for infectious group
relative_distance	Double	Relative distance score: 0.004–0.994 for total dataset, 0.433–0.994 for non-infectious group, and 0.004–0.297 for infectious group
total_similarity	Double	Overall sequence similarity: 0.156–0.981 for total dataset, 0.156–0.643 for non-infectious group, and 0.608–0.981 for infectious group
g_verified	int(Green et al., 1996)	Predetermined group 1 for cases verified as non-infectious group through literature review Predetermined group 2 for cases verified as infectious group through literature review
disg_predicted	int(Green et al., 1996)	Score-based predicted group 1 for cases predicted as infectious group by discriminant model Score-based predicted group 2 for cases predicted as non-infectious group by discriminant model
disds_calculated	Double	Calculated discriminant z-scores which range from −5.804 to 3.526
disds_trimmed	Double	z′-scores converted from z-score based on the group centroids, which range from −4.316 to 2.326
Infectindex	Double	Propensity scores for total data set which range from 0.001 to 99.994%

This table shows the data items, types and the values with description of each data item stored in MySQL DBMS for interworking with the web server ViCIPR.

score), and Infectindex (cross-species infection propensity) (Table 3). The values corresponding to each data item (sequence pair) were classified, assigned, calculated, and stored. Table 3 shows the stored data items, their types, and the values stored in each item.

In this study, web programming was performed to implement the functionality provided by ViCIPR. To implement the sequence input window and an upload function for the user's sequence to perform a sequence homology search, sequence databases were generated as a target sequence resource. All collected sequences were stored in a web project as a FASTA file. We also developed a 'Sequence Similarity Scoring System' based on the Java programming language. HTML, JSP, and JAVA scripts were also used for web development that includes the functions of the program. The main analysis page was designed to enable users to input or upload a query sequence and select a primary (donor) host organism from the species available in the database. To present a selectable secondary (recipient) host species based on the virus genome sequence with the highest percent identity (%) among the target sequences (virus reference genome sequences), the web interface

was designed to include a dynamic search function for multiple conditions, such as virus species and primary host species, in conjunction with the MySQL database. The information for potential secondary host species is provided on the web page, and a dynamic selection box is presented for confirmation. The web interface was configured to search, store, and output all information corresponding to virus name, primary and secondary host species, and Infectindex under a series of analysis processes to output the calculated infection propensities corresponding to the selected data item. For reliable data management and future updates, the pairwise distance, relative distance, total similarity, discriminant scores, and Infectindex values of all sequence pairs were stored in the MySQL database.

2.5.3. System development environment

We organized a system development environment to use our analysis system as an open web resource. For the construction of a server-based calculation program and web interface, we built a server with Intel Xeon E5-240 v2 2.40GHz CPU, 8Gb RDIMM, 1600MT/s Memory, 300GB 15K RPM, and 6Gbps HDD specification. We used CentOS 6.6 as the operating system and MySQL 5.5.40 as the data management system for data storage in a Linux server environment. Java programming language was used for data parsing and computing program development, and JSP, HTML, and Java script programming languages were used for the web pages. The web server program was based on Apache, and Tomcat v7.0.55 was used as a web container.

3. Results

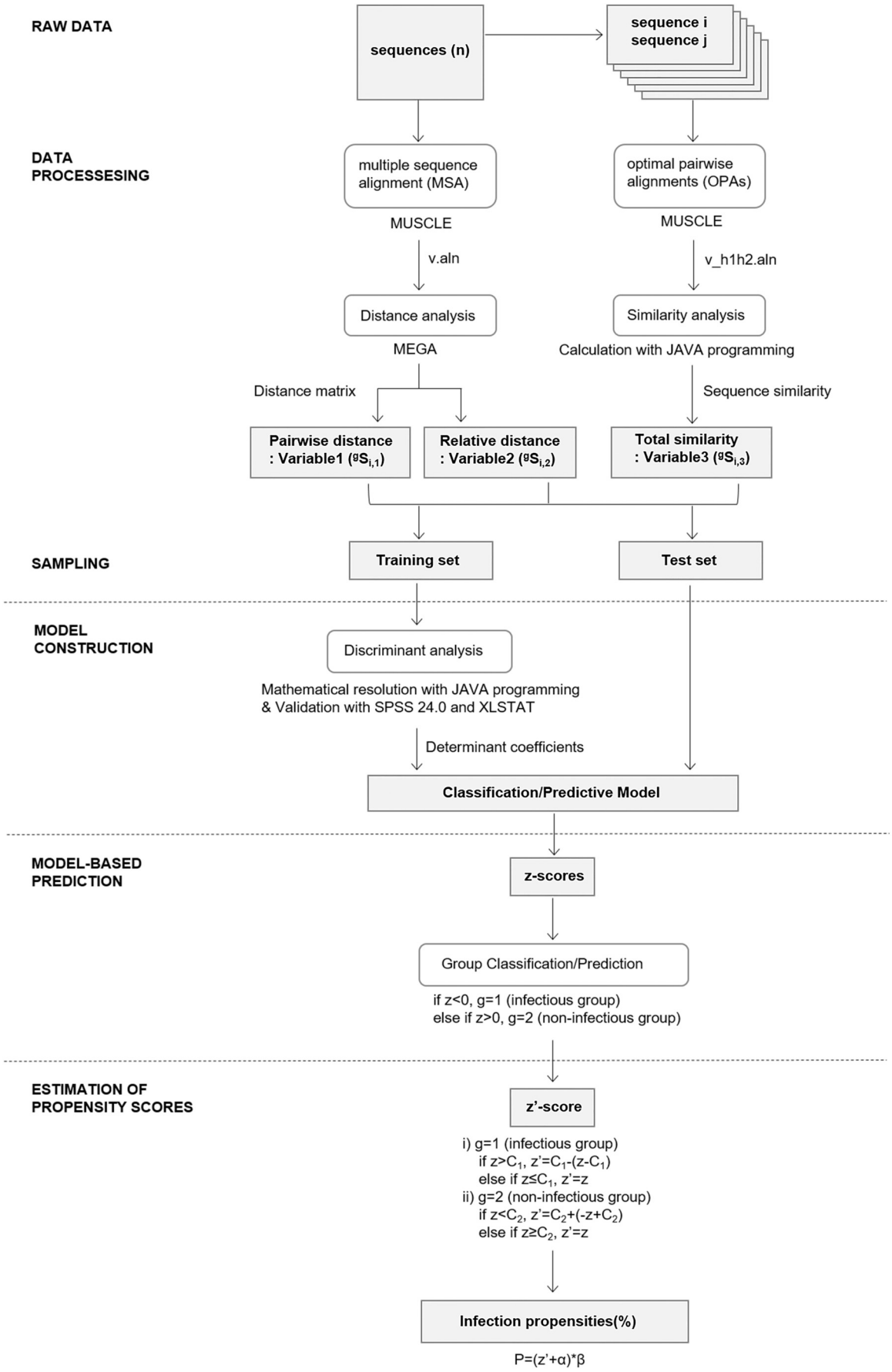
3.1. A computational framework for estimation of virus infection risk based on sequence data

We propose a computational framework to estimate propensities of virus infection between host species. The framework largely consists of four stages: 1) amino acid sequence analysis and predictive variable selection, 2) construction of the classification model, 3) model-based prediction, and 4) calculation of propensity scores (Fig. 1). Based on the results of the discriminant analysis, three characteristics, which quantitatively indicate the evolutionary distance and similarity between two sequences, were selected as predictive variables and used to construct the model. The similarity of the interaction hotspots was excluded from the model structure due to low significance. We derived covariant matrices and pool-within-class covariant matrices for discriminant analysis. SPSS 24.0 was used to calculate inverse matrix and discriminant coefficients to derive the discriminant model and evaluate the contributions of predictive variables. A training data set consisting of 50 sequence pairs was used for model construction, and the constructed classification model was used in model-based prediction using the test data set consisting of six sequence pairs. The z-scores for all sequence pairs were used to classify or predict the group of individual sequence pairs and to estimate the probability of infection between host species in each data set. In the final step, the propensity score for each sequence pair was calculated according to each centroid and z-score transformation equation, and the predicted infection propensity value was presented as the Infectindex.

3.2. Receptor similarities and propensity scores conferring host susceptibility to cross-species infection of viruses

Based on the Infectindex, the infection properties were classified into two groups in the training data set: 32 infectious and 18 non-infectious properties. In the test data set, the infection properties were also classified into two groups: four infectious and two non-infectious properties. The discriminant model had 100% sensitivity, specificity, and total classification accuracy using the training and validation samples. The AUC was calculated as 1. The accuracy of the six test data sets was calculated as the ratio of the number of correctly predicted sequence pairs to the total number of test sequence pairs, and confirmed to be 100%. Despite the low level of mutations among the orthologous receptor sequences used as the current input data, the high sensitivity and specificity obtained in our results may be significant for prediction results produced by the statistical model based on evolutionary and genetic similarities between potential host species. Variables derived from evolutionary distance and sequence similarity appear to have acted as positive or negative factors influencing species-specific susceptibility. However, considering that this approach was limited to the currently available data, our results may be insufficient to obtain highly accurate predictions of cases where detailed and specific infection characteristics must be considered. This limitation serves as a disadvantage in predicting new cases, especially when the method is applied to exceptional cases where the similarity to the current input data is very low. As shown by the relationship among the predictive parameters and resulting propensity index (Table 4), input data with close evolutionary distance and high sequence similarity produce an output result of an infectious group and high Infectindex value. In this regard, the present model should be cautiously applied to new cases that may confer more diverse and subtle host specificities under the assumption that, as evolutionary distance increases and sequence similarity decreases, the probability of sharing sequence properties associated with susceptibility to cross-species infections also decreases. Furthermore, the level of variation among receptor sequences may have a significant impact on host susceptibility, but exceptional patterns can be observed, for example new cases that do not follow linear species distances due to subtle differences in host specificity. In this study, substitution matrices were calculated from target sequences, and the discriminant model consisting of the combined variables was evaluated. In the leave-one-out cross-validation, all original grouped cases were correctly classified. However, the cross-validation result for the test dataset (83.3%) suggests that incomplete coverage of potential polymorphisms in receptor sequences, which may affect infection propensities, can compromise prediction accuracy. Therefore, it remains difficult to predict risk of infection based on the primary structure information of the receptor proteins. However, the discriminant coefficient (7.026) of similarity within the sequence pair, which had the most positive effect on host susceptibility to viral infection, confirmed the importance of high-accuracy sequence polymorphism in improving classification and prediction accuracy. These results suggest that our approach could be improved by including more receptor sequence data. Both SPSS 24.0 and XLSTAT software were used to conduct the discriminant analysis and accuracy evaluation. The z-scores for all cases in the training set were −5.804–3.526, and the propensity scores were 0.001–99.993%. The propensity score ranges were 0.001–36.946%, and 68.294–99.993% for the non-infectious and infectious groups, respectively. These results confirmed that the model correctly classified the original cases into infectious and non-infectious groups with the pertinent propensity score range. From the constructed discriminant model, we devised a simple calculation process that can predict the infection propensity for new individual cases by deriving the first-order function formula using the z′-value, which is converted from the discriminant score using the group centroids.

Table 4 shows cases where there is a high importance of cross-species host susceptibility to viruses in the classification results of the discriminant analysis with distance and similarity scores. The representative zoonotic viruses MERS-CoV and SARS-CoV have an Infectindex of 98.759% for *C. dromedaries–H. sapiens* and 98.495% for *F. catus–H. sapiens*. Both cases show a high probability of interspecies infection. Considering the actual host range of SARS-CoV (Martina et al., 2003; Holmes, 2005), these findings confirm that the difference in Infectindex value reflects the infection properties of the virus. In the case of Nipah virus, the interspecies infection propensity of *S. scrofa–H. sapiens* was 88.572%, while that of *S. scrofa–I. punctatus* was 4.769%. The results show that the infection properties of the viruses differed significantly among the host-pair data sets (Chua et al., 1999; AbuBakar



(caption on next page)

Fig. 1. A computational framework to estimate propensities for virus infection between host species. The framework largely consists of four stages: 1) amino acid sequence analysis and predictive variable selection, 2) construction of the classification model, 3) model-based prediction, and 4) calculation of propensity scores. Based on the results of the discriminant analysis, three measures were selected as predictive variables and used to construct the model. We derived covariant matrices and pool-within-class covariant matrices for the discriminant analysis. SPSS 24.0 software was used to calculate inverse matrix and discriminant coefficients, to derive the discriminant model, and to evaluate the contributions of predictive variables. In this figure, C_1 and C_2 indicate the group centroid of each group used for z′-value computation, and α and β are the coefficients used to transform the z′-score for propensity estimation.

Table 4
Scores for distance, similarity, discrimination and cross-species infection propensity of receptor sequence pairs.

Virus	Host1	Host2	${}^{g}S_{i,1}$	${}^{g}S_{i,2}$	${}^{g}S_{i,3}$	Group	DS	Infect-index
Sendai virus	*R. norvegicus*	*M. musculus*	0.130	0.068	0.837	1	2.326	99.993
MERS-CoV	*C. dromedarius*	*H. sapiens*	0.084	0.072	0.888	1	2.612	98.759
VSV	*B. taurus*	*H. sapiens*	0.025	0.066	0.893	1	2.616	98.700
SARS-CoV	*F. catus*	*H. sapiens*	0.131	0.106	0.852	1	2.226	98.495
SARS-CoV	*F. catus*	*M. putorius furo*	0.085	0.069	0.897	1	2.693	97.546
HIV (SIV;Lentivirus)	*P. troglodytes*	*C. aethiops*	0.091	0.045	0.904	1	2.879	94.749
HIV (SIV;Lentivirus)	*C. aethiops*	*H. sapiens*	0.088	0.044	0.906	1	2.895	94.505
Hantavirus	*M. musculus*	*R. norvegicus*	0.038	0.094	0.962	1	2.963	93.474
Lassa virus, LCMV	*R. norvegicus*	*H. sapiens*	0.071	0.040	0.927	1	3.046	92.233
Hantavirus	*R. norvegicus*	*H. sapiens*	0.099	0.244	0.901	1	1.788	91.898
Lassa virus, LCMV	*M. musculus*	*H. sapiens*	0.064	0.036	0.933	1	3.102	91.386
Measles virus wild-type strains	*H. sapiens*	*M. mulatta*	0.045	0.057	0.967	1	3.207	89.812
Junin virus, machupo virus	*M. musculus*	*C. griseus*	0.161	0.198	0.836	1	1.647	89.784
Nipah virus	*S. scrofa*	*H. sapiens*	0.025	0.043	0.971	1	3.289	88.572
Nipah virus	*M. brandtii*	*H. sapiens*	0.022	0.038	0.974	1	3.334	87.895
Rabies virus	*C. lupus familiaris*	*H. sapiens*	0.027	0.024	0.963	1	3.338	87.835
Human echovirus 7	*P. vampyrus*	*H. sapiens*	0.797	0.433	0.420	2	−1.862	36.946
EMCV	*H. sapiens*	*D. rerio*	1.335	0.994	0.234	2	−5.628	19.752
Sendai virus	*R. norvegicus*	*X. tropicalis*	1.075	0.560	0.300	2	−3.093	18.411
Hepatitis A virus	*H. sapiens*	*L. crocea*	1.050	0.822	0.160	2	−5.524	18.180
Measles virus wild-type strains	*H. sapiens*	*M. davidii*	0.405	0.512	0.360	2	−3.135	17.778
Rabies virus	*B. taurus*	*H. sapiens*	1.126	0.986	0.294	2	−5.389	16.151
Lassa virus, LCMV	*M. musculus*	*D. melanogaster*	1.671	0.941	0.200	2	−5.217	13.559
Measles virus vaccine strains	*H. sapiens*	*S. scrofa*	0.776	0.791	0.431	2	−3.748	8.557
HIV (SIV;Lentivirus)	*P. troglodytes*	*G. gallus*	1.359	0.676	0.250	2	−3.766	8.275
VSV	*H. sapiens*	*A. darlingi*	0.375	0.989	0.630	2	−3.856	6.931
Nipah virus	*S. scrofa*	*I. punctatus*	0.625	0.819	0.44	2	−3.999	4.769
HCoV-229E	*P. alecto*	*C. quinquefasciatus*	1.159	0.840	0.289	2	−4.597	4.230
Hepatitis A virus	*H. sapiens*	*M. brandtii*	0.783	0.613	0.186	2	−4.497	2.718
Nipah virus	*M. brandtii*	*D. rerio*	0.570	0.991	0.536	2	−4.316	0.001

Host1, original/donor host species; host2, alternative/recipient host species. ${}^{g}S_{i,1}$, ${}^{g}S_{i,2}$, and ${}^{g}S_{i,3}$, pair-wise distance, relative distance, and total similarity, respectively. Groups were classified based on the discrimination scores (DSs) (1, infectious group; 2, non-infectious group). The DS was calibrated for correct classification and propensity calculation in the dataset. The group centroid of the discriminant function was 2.428 for the infectious group and −4.316 for the non-infectious group, and was used to calculate the Infectindex.

et al., 2004) (Table 4). In the present study, the maximum calculated propensity value, which represents the greatest risk of cross-species infection, was used as the InfectIndex for viruses that recognize different receptors. For example, the InfectIndex between *M. brandtii* and *H. sapiens* for Nipah virus was 88.310 for ephrin B2 and 88.572 for ephrin B3; we defined the InfectIndex as the highest index value, i.e. 88.572.

3.3. Web resource for cross-species infection propensities

In this study, we constructed ViCIPR, a server-based tool to enable users to easily extract cross-species infection propensities of specific viruses using a simple two-step procedure. ViCIPR can be accessed from a web server at http://lcbb3.snu.ac.kr/ViCIPR/home.jsp.

ViCIPR provides search functions for homology and calculates results based on query sequences with an interface that allows users to select from a set of databases, including sequence data sets uploaded by the user. The target sequence resource for calculating the Infectindex was the sequence database collected, processed, and stored in the NCBI protein database. The similarity of a query sequence to database reference sequences is calculated, and the result is presented via the 'Sequence Similarity Scoring System'. Users can enter a query sequence by pasting directly into the query box or by uploading a sequence as a FASTA file from a local computer. Currently, our sequence similarity scoring system is performed based on three databases: ViCIPR, all nucleotides (genomes); ViCIPR, all cds nucleotides (genes); and ViCIPR, all cds proteins (proteins). Fig. 2 shows the main components and data flow of ViCIPR. ViCIPR presents a selectable secondary (recipient) host species based on similarity search results for the input data corresponding to the query sequence (gene or protein) and the host species, and presents the Infectindex calculation result at the same time as selecting the options (Fig. 2). Fig. 3 shows the progress and results of extracting predicted infection propensities of SARS-CoV using ViCIPR. As shown in Fig. 3, a simple two-step procedure makes it easy to obtain the Infectindex of two potential hosts for cross-species infection of a particular virus (Fig. 3).

4. Discussion

Infections caused by newly emerging viruses are serious threats to public health and have become a global concern. A variety of factors and their interactions may contribute to disease emergence. In this study, we focused on how viruses can be transmitted between an established reservoir species and a new host species and on what determines the potential of a virus to cross the barrier to a previously uninfected species. Based on the hypothesis that the major barrier to interspecies virus infection is the difference between cell-receptor sequences, we evaluated the genetic risk for cross-species infections

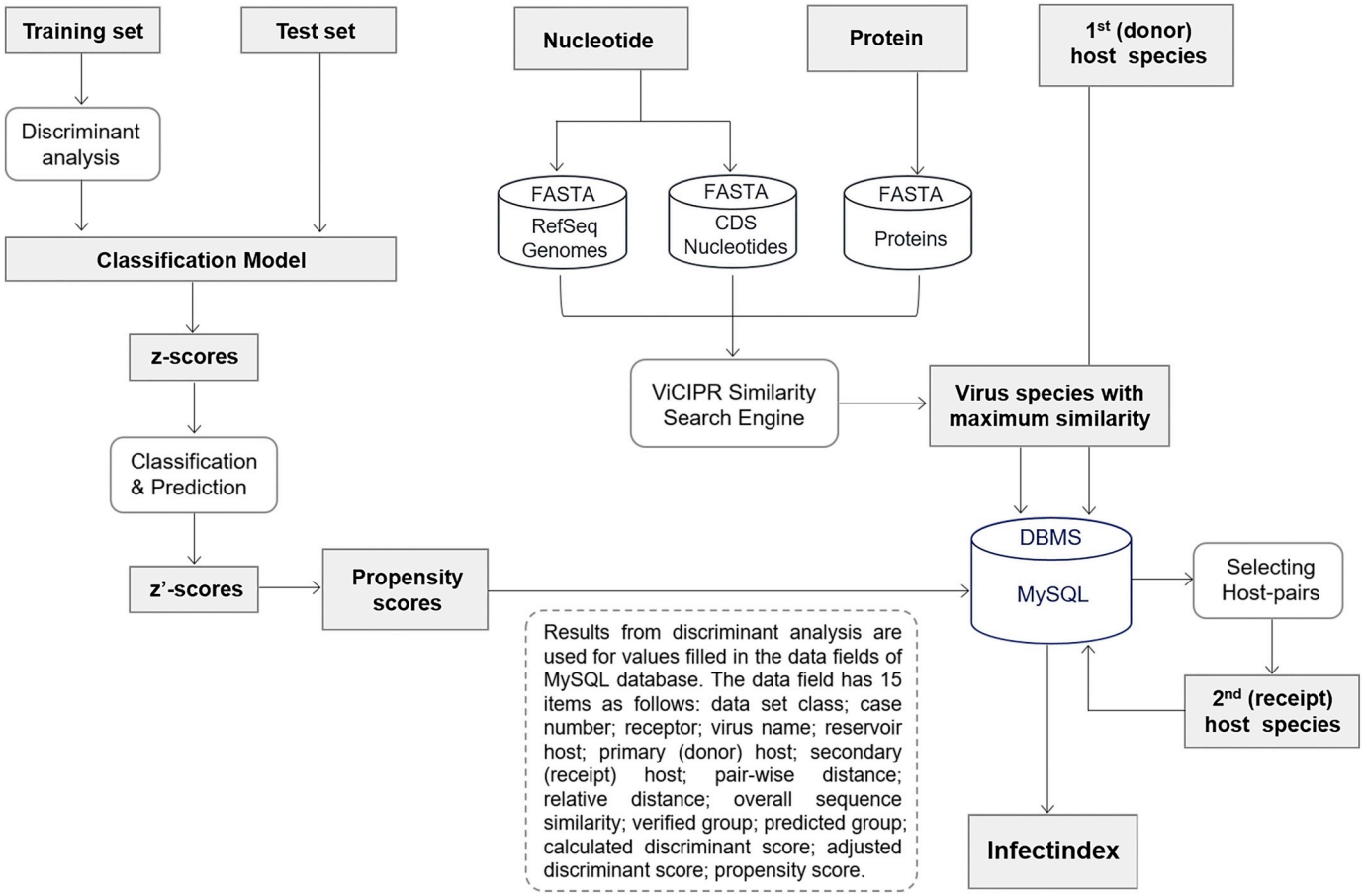


Fig. 2. Main components and data flow of ViCIPR, a web-based prediction system. The data flow is shown with the process of establishing a discriminative model and predictive protocol, a database capable of dynamic interaction, and a user-friendly web interface as the key components for the operation of analytical systems in ViCIPR (http://lcbb3.snu.ac.kr/ViCIPR/home.jsp). As shown, we tried to construct ViCIPR based on our own statistical protocol. The results of the protocols and prediction studies were stored in a database that can be used in ViCIPR. Operation of the ViCIPR analysis system is initiated by the input of query sequences (nucleotides or proteins) and selection of the primary (donor) host species. Next, a similarity search is performed on the query sequence and host information with the user's input. Based on the similarity search result, a virus species with maximum similarity is given as the output, and a selectable secondary (recipient) host species is presented in connection with the MySQL database. The results show a calculated value for the Infectindex of the selected host pair of the corresponding virus species at the same time as the selection of the secondary host species.

between potential host species based on evolutionary distance and similarity of receptor sequences. Correct results from both the training and test data sets may have been a result of accurate measurement of possible sequence variations that affect cross-species infection susceptibility. The accuracy of predictive variables that positively and negatively influence virus transmission suggests that this approach could be improved with additional receptor sequence data. Among the three variables, sequence similarity was the most important for classification and prediction. The possibility of identifying virus transmission between hosts without requiring complex structural analysis suggests the importance of this sequence-based approach. Due to the lack of available data, this method is limited to viruses whose receptor use has been determined. In our method, it is necessary to calculate distance matrices and perform MSA for parameterization of the characteristics obtained from protein sequence analyses. Thus, accurate protein sequence data comprising sufficient sequence lengths should be available for various species. As a result, the Infectindex was estimated for ephrin B2 and ephrin B3 of the Nipah virus, respectively, by excluding other receptor proteins whose sequence data for various species were insufficient from the analysis. For the receptors used by other viruses, further analyses will enable calculation of the Infectindex as soon as sufficient data are available. We will continue to conduct additional studies to update the relevant data and prediction tools.

In addition, our results are based on primary structural information from full-length protein sequences and are limited in that they do not reflect the detailed mechanisms involved in virus–host-cell interaction. These limitations can make it difficult to explain subtle differences in the susceptibility of host organisms to viral infections, which exhibit differences in host range at the strain level. Therefore, the statistical model and model-based discriminant values should be cautiously accepted at the viral species level, and further experimental validation is required to improve applicability. For example, in the case of influenza virus, the type of sialic acid (2,3- or 2,6-linkage), which is determined by the biochemical repertoire of the host cell, is identified as an important factor involved in preferentially recognizing and binding to avian or human cells. In this study, these subtype-level properties were not applied to generate parameters for model construction. To take these detailed features into account, additional problems should be solved such as adjustment to the taxonomic level of other virus species constituting the data sets and weighting the values of specific amino acid residues that affect multiple receptor types. In considering all of these problems, the accuracy of multivariate-based classification and prediction can be impaired if a sufficient amount of reliable data cannot be guaranteed. Therefore, we believe that further research should be conducted based on our findings to discover more advanced methods that can be applied in special cases, such as influenza virus. As discussed above, the propensity for interspecies infection, which is estimated statistically for each virus species, has limitations in reflecting

Step 1. Identify Similar Sequences

<Query page>

Sequence type nucleotide ▼ Database ---------- ▼

ViCIPR all nucleotides
ViCIPR all cds nucleotides
ViCIPR all proteins

Enter query sequence here: (70~7000 nt, 30~7

>Sample sequence: SARS coronavirus sars4 E g

ATGTACTCATTCGTTTCGGAAGAAACAGGTACGT TTGCTTTCG
TGGTATTCTTGCTAGTCACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCTGCAATATTGT
TAACGTGAGTTTAGTAAAACCAACGGTTTACGTCTACTCGCGTGTTAAAAATCTGAACTCTTCTGAAGGA
GTTCCTGATCTTCTGGTCTAA

※Important note: This tool allows for similarity search for the query sequence of up to 7,000 lengths

Reset Search

<Result page>

═ Results Summary ═

- Maximum within-group matching score: 231
- Virus species with maximum identity: SARS coronavirus region4 (100%)
- Virus species producing significant sequence alignments [selectable hosts]: SARS coronavirus region3-4 (100%) [R. ferrumequinum, M. brandtii, M. putorius furo, H. sapiens, F. catus, P. larvata], SARS coronavirus region4 (100%) [R. ferrumequinum, M. brandtii, M. putorius furo, H. sapiens, F. catus, P. larvata]

→ See similarity search results for the entire target sequence database

※Select the items required in the next step based on the results of virus species exhibiting significant sequence alignment and selectable hosts

Step 2. Select Virus & Primary (donor) host species

Virus species SARS-CoV ▼ Primary (donor) host species ---------- ▼ Submit

----
R. ferrumequinum
M. brandtii
M. putorius furo
H. sapiens
F. catus
P. larvata

Infectindex Calculation

1> Virus	SARS-CoV
2> Primary host	F. catus
3> Secondary host	H. sapiens ▼
Infectindex	98.495 %

H. sapiens
M. putorius furo

(caption on next page)

Fig. 3. A simple two-step procedure in the ViCIPR web interface. The process and results of the extraction of Infectindex for the SARS-CoV are shown. In the first step, a similarity search of the viral genome sequence data library among the target sequence resources in the ViCIPR genomic database was performed, and the results were output. Using the built-in search function of ViCIPR, the maximum matching score, the virus species with the best and most relevant hits, and the virus species with hits (%) and selectable hosts corresponding to the results of significant sequence alignments were retrieved. In the second step, a list of selectable primary host species is presented by the user's selection of the virus, which is based on the information of the virus species with the maximum percent identity among the viruses corresponding to the target sequences. Finally, the cross-species infection propensity of the host pair determined according to the user's selection is calculated and output to the result window. As shown, ViCIPR database similarity search results indicated that the SARS-CoV was the virus species most similar to the query sequence. We can confirm that the selected secondary hosts *H. sapiens* and *M. putorius furo* for the primary host *F. cattus* are presented in the select box. Simultaneously with selection of the secondary host species, the results of the Infectindex calculation were output to the box.

subtle differences that may occur at lower virus classification levels. However, under the assumption that the variable sequence similarity itself comprehensively includes variation in subtype-specific virus–cell interactions, which confer the receptor binding affinity that causes changes in host tropism, it is expected that polymorphisms of receptor sequences in a wide variety of potential host species can be accurately measured, and the resulting information about relevant residues utilized. In the present study, interaction hotspot similarity, which we thought would serve as an important variable following preliminary analysis, was excluded from the model due to its low significance. This seems to be due to the limitations of the use of predicted data resulting from the lack of information on protein tertiary structure and amino acid residues important in viral attachment to host cells. The identification of specific amino acid residues that participate in the virus–host interaction would enable more realistic simulation of the infection mechanism, which would lead to a more precise prediction of the host ranges and infection propensities of newly emerging viruses. Consideration of the effects of the secondary structure, physical properties and chemical properties of proteins, which are involved in virus–host interactions, on host susceptibility to cross-species viral infections may improve this method. In this study, predictive variables were determined based on receptor sequence analyses, but we believe that the receptor-binding domains of each virus will provide significant information to improve the predictive power of the model. Further effort should be made to improve the flexibility of the method in handling subtle differences in host specificity that do not follow linear species distances while ensuring high accuracy. In this regard, we are currently parameterizing the various properties of proteins to improve model prediction accuracy, and are working on docking simulations using virus and host protein information to produce reliable data that can be usefully applied in our future prediction studies. Complex factors that determine viral host tropism, such as a variety of viral transmission modes, the action of animal vectors or carriers, and the process of host immune response, could be included for better prediction. Although further refinements are needed, this approach may be useful as a basic tool for prior studies in accurately predicting host susceptibility to new viral infections.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (No. 2016R1C1B2015511) and the Ministry of Education (No. 2017R1D1A1B03033413).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2019.04.016.

References

AbuBakar, S., Chang, L.Y., Ali, A.M., Sharifah, S.H., Yusoff, K., Zamrod, Z., 2004. Isolation and molecular identification of Nipah virus from pigs. Emerg. Infect. Dis. 10, 2228.
Addinsoft, 2017. XLSTAT V. 2017: Data Analysis and Statistics Software for Microsoft Excel.
Bae, S.E., Son, H.S., 2011. Classification of viral zoonosis through receptor pattern analysis. BMC Bioinform. 12, 96.
Chua, K.B., Goh, K.J., Wong, K.T., Kamarulzaman, A., Tan, P.S.K., Ksiazek, T.G., Zaki, S.R., Paul, G., Lam, S.K., Tan, C.T., 1999. Fatal encephalitis due to Nipah virus among pig-farmers in Malaysia. Lancet 354, 1257–1259.
Cor, I.S., 2016. IBM Spss Statistics for Windows, Version 24.0. IBM Corp, Armonk (NY).
Dales, S., 1973. Early events in cell-animal virus interactions. Bacteriol. Rev. 37, 103.
Dimitrov, D.S., 2004. Virus entry: molecular mechanisms and biomedical applications. Nat. Rev. Microbiol. 2, 109.
Domingo, E., 2010. Mechanisms of viral emergence. Vet. Res. 41, 38.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.
Graham, R.L., Baric, R.S., 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J. Virol. 84, 3134–3146.
Green, S.B., Salkind, N.J., Jones, T.M., 1996. Using SPSS for Windows; Analyzing and Understanding Data. Prentice Hall PTR.
Grove, J., Marsh, M., 2011. The cell biology of receptor-mediated virus entry. J. Cell Biol. 195, 1071–1082 (jcb-201108131).
Haagmans, B.L., Andeweg, A.C., Osterhaus, A.D., 2009. The application of genomics to emerging zoonotic viral diseases. PLoS Pathog. 5, e1000557.
Holmes, K.V., 2005. Adaptation of SARS coronavirus to humans. Science 309, 1822–1823.
Imai, M., Kawaoka, Y., 2012. The role of receptor binding specificity in interspecies transmission of influenza viruses. Curr. Opin. Virol. 2, 160–167.
Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990.
Lachenbruch, P.A., Mickey, M.R., 1968. Estimation of error rates in discriminant analysis. Technometrics 10, 1–11.
Martina, B.E., Haagmans, B.L., Kuiken, T., Fouchier, R.A., Rimmelzwaan, G.F., Van Amerongen, G., Malik Peiris, J.S., Lim, W., Osterhaus, A.D., 2003. Virology: SARS virus infection of cats and ferrets. Nature 425, 915.
Matrosovich, M., Tuzikov, A., Bovin, N., Gambaryan, A., Klimov, A., Castrucci, M.R., Donatelli, I., Kawaoka, Y., 2000. Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. J. Virol. 74, 8502–8512.
Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R., 1999. Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop. Ieee, pp. 41–48.
Morse, S.S., Mazet, J.A., Woolhouse, M., Parrish, C.R., Carroll, D., Karesh, W.B., Zambrana-Torrelio, C., Lipkin, W.I., Daszak, P., 2012. Prediction and prevention of the next pandemic zoonosis. Lancet 380, 1956–1965.
Pepin, K.M., Lass, S., Pulliam, J.R., Read, A.F., Lloyd-Smith, J.O., 2010. Identifying genetic markers of adaptation for surveillance of viral host jumps. Nat. Rev. Microbiol. 8, 802.
Rappuoli, R., 2004. From Pasteur to genomics: progress and challenges in infectious diseases. Nat. Med. 10, 1177.
Rogers, G.N., Paulson, J.C., Daniels, R.S., Skehel, J.J., Wilson, I.A., Wiley, D.C., 1983. Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. Nature 304, 76.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Schneider-Schaulies, J., 2000. Cellular receptors for viruses: links to tropism and pathogenesis. J. Gen. Virol. 81, 1413–1429.
Skehel, J.J., Wiley, D.C., 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. 69, 531–569.
Snowden, F.M., 2008. Emerging and reemerging diseases: a historical perspective. Immunol. Rev. 225, 9–26.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.
Woolhouse, M.E., Gowtage-Sequeria, S., 2005. Host range and emerging and reemerging pathogens. Emerg. Infect. Dis. 11, 1842.
Woolhouse, M., Scott, F., Hudson, Z., Howey, R., Chase-Topping, M., 2012. Human viruses: discovery and emergence. Phil. Trans. R. Soc. B. 367, 2864–2871.
Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins, pp. 97–166.