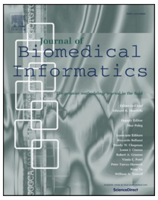
Journal of Biomedical Informatics 117 (2021) 103736

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Original Research

Reservoir hosts prediction for COVID-19 by hybrid transfer learning model

Yun Yang [a,1,*], Jing Guo [b,1], Pei Wang [b], Yaowei Wang [a], Minghao Yu [a], Xiang Wang [a], Po Yang [c], Liang Sun [d,e,*]

[a] School of Software, Yunnan University, Kunming, China
[b] School of Information Science and Engineering, Yunnan University, Kunming, China
[c] Department of Computer Science, Sheffield University, Sheffield, UK
[d] The MOH Key Laboratory of Geriatrics, Beijing Hospital, National Center of Gerontology, China
[e] The NHC Key Laboratory of Drug Addiction Medicine, Kunming Medical University, Kunming, China

ARTICLE INFO

Keywords:
COVID-19
Machine learning
Transfer learning
Ensemble learning
Hosts prediction
Virus origins

ABSTRACT

The recent outbreak of COVID-19 has infected millions of people around the world, which is leading to the global emergency. In the event of the virus outbreak, it is crucial to get the carriers of the virus timely and precisely, then the animal origins can be isolated for further infection. Traditional identifications rely on fields and laboratory researches that lag the responses to emerging epidemic prevention. With the development of machine learning, the efficiency of predicting the viral hosts has been demonstrated by recent researchers. However, the problems of the limited annotated virus data and imbalanced hosts information restrict these approaches to obtain a better result. To assure the high reliability of predicting the animal origins on COVID-19, we extend transfer learning and ensemble learning to present a hybrid transfer learning model. When predicting the hosts of newly discovered virus, our model provides a novel solution to utilize the related virus domain as auxiliary to help building a robust model for target virus domain. The simulation results on several UCI benchmarks and viral genome datasets demonstrate that our model outperforms the general classical methods under the condition of limited target training sets and class-imbalance problems. By setting the coronavirus as target domain and other related virus as source domain, the feasibility of our approach is evaluated. Finally, we show the animal reservoirs prediction of the COVID-19 for further analysing.

1. Introduction

With the rapid growing number of infected and dead people all over the world, COVID-19 is becoming one of the most serious epidemic situations for the humankind. The emerging viral infection, such as COVID-19, Ebola, SARS, MERS and Zika, is a kind of zoonosis which is caused by the cross-species transmission of virus [1]. Due to the pathogenicity, lethality and difficulty of prevention, these diseases have made a grievous threat to the health of mankind and livestock [2]. Virus is an acellular form, usually contains a long chain of nucleic acids and a protein shell, which has no metabolic machinery or enzyme system. Unlike most living things, virus survives mainly depending on the host cells that provide places for virus to replicate, transcribe and transfer to other organisms [3]. Once the zoonotic disease outbreak or new pathogenic virus is found, origin tracing has a great significance not only to figure out how the pathogens infect humans and develop the vaccine for human beings, but also can isolate the infectious animal hosts for further spreading. In the history of mankind, the explorations of the virus origins have made many progresses, for example, Hu et al. [4] take 5-year surveillance of SARSr-CoVs in a cave inhabited by multiple species of horseshoe bats in Yunnan Province, China, to identify the origin of SARS-CoV. Reusken et al. [5] collect 349 serum samples from all over the world to demonstrate the reservoir of MERS-CoV has high relationship with Omani camels. The researchers have made many studies for COVID-19 recently [6–9]. By the conclusion that the similar phylogenetic characteristics of genomes may have the similar hosts information, the researchers seek the similar genome sequences from the previous virus database and find that the COVID-19 and the virus (Bat CoV RaTG13) have a full gene level consistency up to 96%, and the viruses (bat-SL-CoVZC45 and bat-SL-CoVZXC21) have the consistency

[*] Corresponding authors at: School of Software, Yunnan University, Kunming, China (Yun Yang); The MOH Key Laboratory of Geriatrics, Beijing Hospital, National Center of Gerontology, China and The NHC Key Laboratory of Drug Addiction Medicine, Kunming Medical University, Kunming, China (Liang Sun).

E-mail addresses: yangyan19@hotmail.com (Y. Yang), sunbmu@foxmail.com (L. Sun).

[1] Yun Yang and Jing Guo contributed equally to this work.

https://doi.org/10.1016/j.jbi.2021.103736
Received 23 August 2020; Received in revised form 4 February 2021; Accepted 27 February 2021
Available online 9 March 2021

nearly 90% [9]. According to these researches, the mainstream of hypothesizes present that the most probable host for COVID-19 comes from the Chinese Horseshoe Bat. However, traditional methods of virus origin need huge evidences from field surveillance, phylogenetic, laboratory experiments, and real-world interventions [10] which consume lots of time and labor. Thus, it is crucial to develop a more precise, celerity and comprehensive method not just for COVID-19, but also for other newly discovered viruses in the future.

Recent studies have made a great development in discovering the information of genome sequences via machine learning [11] which offer a new research orientation for virus origins. Through building a mathematical model based on training datasets, machine learning can dig out the latent information from dataset and make predictions for specific tasks [12]. Since a large amount of genetic information is recorded in genome sequences, knowledge about the viral hosts can be mined through machine learning [13–16]. Compared with other application fields in machine learning, although the overall base of biological data is large, the scarcity of annotated virus and imbalanced hosts information restrict the efficiency of traditional machine learning in hosts prediction. Specifically, due to the difficulty of tracing the virus origins, the number of annotated virus sequences are limited, especially for the newly discovered virus. The variances of data distribution among different virus species will further constrain the performance of traditional machine learning methods in origin predictions. In addition, some hosts categories, like bat and poultry, are the most frequently discovered carriers of virus, which lead to the prediction results bias to the majority categories.

Transfer learning breaks the assumption of traditional machine learning that the distribution of training data and testing data must be consistent [17,18]. It can effectively use the knowledge obtained from the related virus domain to improve the capability for the task of hosts prediction when there is limited data in specific virus species. Transfer learning can be divided into four categories [17]: instance-based transfer learning [19–21], feature-based transfer learning [22–25], model-based transfer learning [26,27] and relation-based transfer learning [28,29]. Although transfer learning can remarkably decrease the distribution variance between different virus species, it cannot fix the imbalance problem and lead to low generalization when applying into newly discovered virus. While ensemble learning completes classification tasks by constructing and combining multiple weak classifiers to achieve better results [30,31]. To obtain a superior performance in hosts prediction, ensemble learning can be used for three main purpose: (1) Fix the imbalanced class; (2) Utilize the tri-training strategy to give the predictions from different views; (3) Ensemble the results in each iteration to acquire better generalization. Our previous studies have illustrated the advantages of ensemble learning algorithms [32–36].

In this paper, we provide an integrated workflow to make a fast, accurate and reliable host tracing for newly discovered virus (e.g., COVID-19). All the adopted genome sequences and the hosts information are downloaded from GenBank. We preprocess and annotate the collected sequences to construct a benchmark dataset of reference sequences (refseqes) [37]. To overcome the fundamental weakness in traditional machine learning methods in hosts prediction, we extend transfer learning and ensemble learning to present a novel hybrid transfer learning model (HTL) that takes the advantages of the information gained from the related virus domain to enhance the prediction reliability on target virus domain. Specifically, considering the scarcity of annotated virus dataset, we utilize transfer learning to decrease the distribution variance between different virus species and leverage the well annotated virus domain as auxiliary to improve the prediction reliability for newly discovered virus. A bootstrap-based ensemble strategy is proposed to unravel the class imbalance problem and reconcile the outputs to obtain a final transfer result. In general, the main contributions of our work are summarized as follows:

- Considering the traditional virus origin methods consume lots of time and labor, our approach provides a new research reference for hosts tracing from the perspective of data mining and offers an intelligent workflow to accurately and quickly predict the host of newly discovered virus.
- By collecting the genome sequences and annotating the hosts information, we build a benchmark of viral reference sequences to help researchers discover and develop the fight against virus.
- To solve the problems of few annotated data and class-imbalance in virus datasets, we utilize the advantages of transfer learning and ensemble learning to conduct a HTL model. Being compared with other classical machine learning algorithms, we demonstrate the superiority of HTL on UCI benchmarks and virus datasets.
- We utilize HTL to give the host prediction for newly discovered virus (e.g., COVID-19). Being compared with traditional virus origin methods such as sequence identity and phylogenetic tree, we demonstrate the reliability of our method.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 describes the workflow of predicting the origins of virus. Section 4 reports the simulation results on UCI benchmarks and viral dataset, furthermore, gives the prediction results of COVID-19. Then we provide further discussion of the results and the future work. Conclusions are drawn in the last section.

2. Related work

Several previous works about viral origins prediction via machine learning methods are illustrated in this section. Christine et al. [13] propose a methodology to construct the computational models for 11 influenza proteins. By adopting the machine learning algorithm random forest [38], high accurate prediction models are trained to predict the host tropism of influenza A virus. Being compared with other 5 classical machine learning methods, random forest reveals the better performance than others. Babayan et al. [14] utilize gradient boosting machines (GBMs) [39] to demonstrate the superiority of machine learning in predicting the virus hosts and arthropod vectors from viral genome sequences. With the hypothesis of related viruses have closely related hosts, they design a phylogenetic neighbourhood (PN) model to predict host associations from viral phylogenetic connection [40] and combine these phylogenetic information with algorithm to maximize the prediction accuracy. Furthermore, Babayan provides a suitable way to quantify the genome sequences into 4229 traits, including codon pair, dinucleotide, codon, and amino acid biases. Through this way, the latent information of sequences can be extracted into a more interpretative form for machine learning to construct the classification models. Zhang et al. [15] present a method which uses the word frequencies of viral sequences to predict the virus-host infectious associations. Under the hypothesis of the word pattern between viruses and their hosts have more similarities [41], Zhang conducts four different feature vector representations [42–44] and evaluates them by various supervised machine learning methods. The latest work of viral host prediction, Mock et al. [16] provide a general workflow consists of several steps, including data collection, subsets creation, model construction and output the final prediction. Unlike the work illustrated before, they directly use one hot encoding to alter the sequences into a list of numerical vectors rather than make a sequence quantification. The deep neural network architecture is constructed by the CNN layers [45], LSTM layers [46] and two dense layers.

To the best of our knowledge, our work is among the first attempt to utilize the transfer learning to solve the host origin problems. In such way, the fundamental weakness of traditional methods can be reduced, hence a satisfactory result can be achieved.

3. Methods

In this section, we first describe the proposed workflow of predicting the hosts for newly discovered virus, and then present the way of data preprocessing and viral datasets construction. The hybrid transfer learning model is described as the main component of our proposed method.

3.1. AI-based viral origins prediction

The AI-based viral origins prediction method consists of four components: (I) Use the detection technology to discover the new virus; (II) Detect the genome sequences of the viruses by using nucleic acid testing, such as high-throughput sequencing and reverse transcription-polymerase chain reaction; (III) Collect the genome sequences and annotate the datasets with the hosts information; (IV) Train machine learning based classification model with the collected datasets, then utilize the model to make the final prediction. The workflow of AI-based viral origins prediction is illustrated in Fig. 1.

In this paper, we focus on the last two processes and summarize them into two major perspectives. First, for the data pre-processing and virus dataset construction, we collect the genome sequences and annotate the sequences with the information from GenBank, which is an online gene sequence database collects and annotates all published nucleic acid and protein sequences. Then the feature quantification methods are used to change the collected genome sequences into a numerical form. We build a phylogenetic tree to further demonstrate the variances exist among different virus species and show the similarity between COVID-19 and coronavirus in Fig. 2. Thus, the collected datasets are divided into the target domain with coronavirus and COVID-19 and the source domain with other related virus sequences. Second, we construct the hybrid transfer learning model to make the final prediction. We utilize the domain adaptation method [23] to map the source domain and target domain into a Reproducing Kernel Hilbert Space (RKHS) [47] in which the variances between the source and target domain should be as similar as possible. After that, ensemble learning methods are used to construct the classification model. With the undersampling [48] and the tri-training strategy [31], our method can refine the class-imbalance problems and get a better performance on target domains.

3.2. Data Pre-processing and virus dataset construction

For predicting the hosts for newly discovered virus (e.g., COVID-19), we focus on the RNA viruses since they are the primary pathogen group responsible for emerging human diseases. The detail information of concerning viral datasets are shown in Table 1. We download all the viral genome sequences from NCBI viral genome database, which consists of 308 samples of COVID-19, including 307 complete genome sequences and 1 refseq, and 775 annotated refseqes with 11 taxonomic groups (Arenavirus, Astrovirus, Bunyavirus, Calicivirus, Coronavirus, Filovirus, Flavivirus, Hepevirus, Paramyxovirus, Picornavirus, Rhabdovirus). The selected viral species cover most of the virus species that can infect human diseases. The reason for we adopt the refseqes to construct the training datasets is that the complete genome sequences contain a large number of analogous sequences from same virus types, resulting in the overfitting when building the classification model. For each genomic sequence, corresponding coding sequence (CDS) encodes the specific protein products for the virus which provides a significant amount of important information about gene expression [49]. Therefore, we only consider the sequences that encodes the protein product and extract the corresponding CDS to form the final data collection. With regard to the multi coding sequences exist in each genomic sequence caused by the different synthesis of proteins, we quantify each coding sequence and superimpose the results to obtain the characteristic quantization for this sequence. After that, the relevant hosts information is annotated from GenBank. According to the authoritative literatures of viruses, we annotate most of the hosts information based on organisms that have long helped the virus survive. On account of the dense granularity of hosts information given in GenBank, we map the collected hosts information into 10 categories, including Artiodactyl, Carnivore, Fish, Galloanserae, Neoaves, Plant, Primate, Pterobat, Rodent and Vespbat. It is worth noting that the bats (order Chiroptera) were split into Pteropodiformes (families Pteropodidae, Rhinolophidae, Hipposideridae, Megadermatidae, and Rhinopomatidae, here abbreviated “Pterobats”) and Vespertilioniformes (remaining microbat families, here abbreviated “Vespbats”). These hosts categories include mammals, rodents and birds, which are common linked to the virus that can infect humans and spread from species.

Inspired by [9], which provides a Blastn research for the complete genomes of COVID-19 and shows the similarities with other coronavirus in sequence identity and query coverage, we speculate that coronavirus is the most suitable virus group to construct the classification model for predicting the hosts for COVID-19 and the discrepancy among different virus species will restrain the accurate of the final prediction. To further demonstrate the variances between different viral species and provide a better visualization, we randomly sample from the collected viral refseqes and the genome sequences of COVID-19 to construct a phylogenetic tree. The MAFFT is a multiple sequence alignment program which is used to align the viral sequences [50]. With the processed sequences, we create the tree by iTOL [51] which is an online tool for the display and manipulation of phylogenetic trees. As shown in Fig. 2, we variegate the leaf nodes with three types of colors to represent the coronavirus group, the COVID-19 virus group and other related virus group, the inside circle denotes the GenBank accession of the viruses. We use 10 kinds of colors to denote 10 reservoir hosts and blank for COVID-19. The branch in the radial direction represents the viral evolution of the evolutionary lineage with time, which means that the nearer the viral species on the branch, the closer the kinship of virus will be and the genome sequences will have the similar expressions. Note a fact that, the COVID-19 virus has the similar branch length with most of coronavirus in depth. In other words, the COVID-19 viruses have the large dissimilarities with other related virus groups, but the similar depth and breadth with most of the coronavirus. However, as shown in Table 1, the annotated training samples are limited and the classes of reservoir are imbalanced, only using coronavirus dataset is insufficient to build a highly generalized model for COVID-19. Consequently, the other related 716 viral sequences with well annotated hosts information can be used as an auxiliary to optimize the task of predicting the hosts of COVID-19.

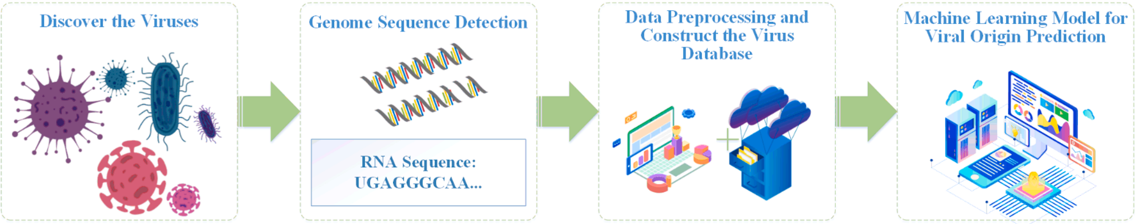

Discover the Viruses
Genome Sequence Detection
RNA Sequence: UGAGGGCAA...
Data Preprocessing and Construct the Virus Database
Machine Learning Model for Viral Origin Prediction

Fig. 1. An overview of the proposed AI-based viral origins prediction.

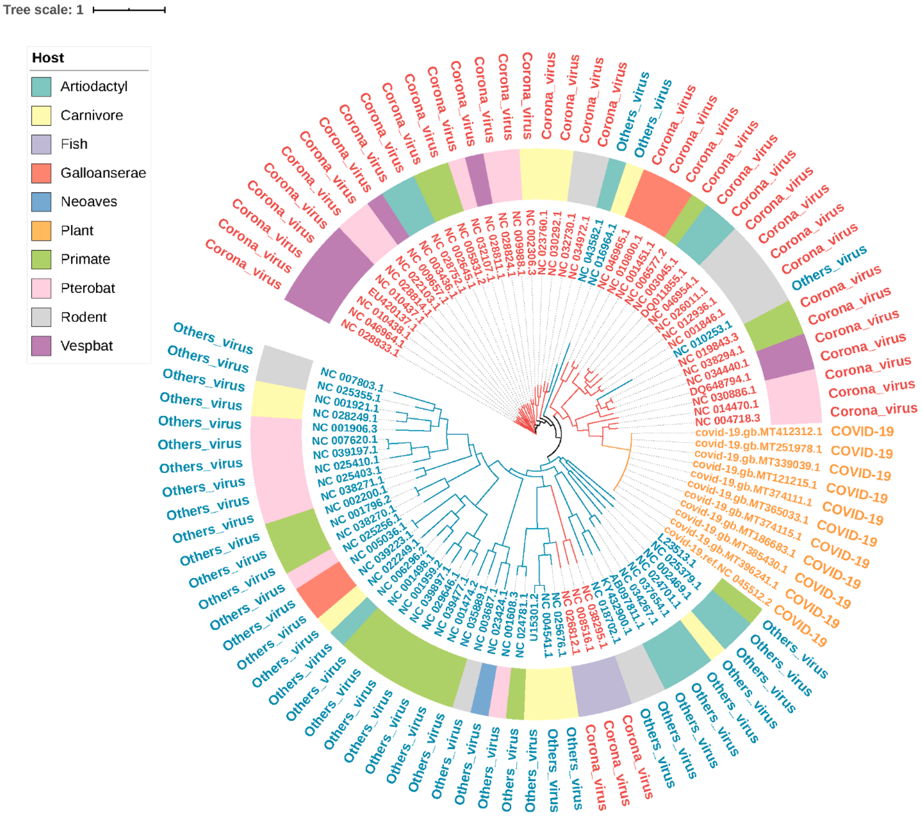


Fig. 2. The phylogenetic tree of the viral genome sequences. By proportional sampling from the collected datasets, the phylogenetic tree is built to demonstrate the variances among different virus species. The blue leaf nodes denote the other related virus species, the red denote the coronavirus and we highlight the COVID-19 and colored with orange.

Table 1
The descriptions of the virus datasets.

Reservoir groups	*Samples*	*Source domain*	*Target domain*	
		Other related virus	*Coronavirus*	*COVID-19*
Artiodactyl	79	73	6	–
Carnivore	36	33	3	–
Fish	35	32	3	–
Galloanserae	78	72	6	–
Neoaves	47	40	7	–
Plant	111	107	4	–
Primate	146	140	6	–
Pterobat	62	53	9	–
Rodent	147	141	6	–
Vespbat	34	25	9	–
Total	775	716	59	308

We divide the datasets into target domain with 59 coronavirus refseqes and COVID-19 sequences, and source domain with other related virus sequences as illustrated in Table 1.

3.3. Quantification of viral genomic traits

Following the method in [14], we quantify each viral genomic sequences into 4229 genomic traits, including all possible codon pair, dinucleotide, codon, and amino acid biases. The detail calculation methods are defined as follows:

3.3.1. Dinucleotide bias

The dinucleotide bias is calculated across all coding genomic sequences by Eq. (1).

$$Dinucleotide\ bias = \frac{\left(\frac{N_{XY}}{DN_{all}}\right)}{\left(\frac{N_X}{N_{all}} \times \frac{N_Y}{N_{all}}\right)} \tag{1}$$

in which, the N denote the count number, the N_X,N_Y, N_{all}, N_{XY} and DN_{all} represent the number of nucleotides X and Y, the total nucleotides numbers, the total count of the dinucleotide XY, and the total number of dinucleotides across the whole sequences. According the [52], dinucleotide bias on viral fitness are reported to be strongest at the bridge between neighbouring codons, an extra calculation is made for "bridge" and "non-bridge" codon positions. With each of the 16 possible dinucleotides for nucleotide, it generates 64 traits in this way.

3.3.2. Codon pair score (CPS)

Codon pair bias was measured as the codon pair score (CPS) for each of the 4096 (64 × 64) possible codon-codon pairs. Codons are the three adjacent bases that determine amino acids on mRNA. There are 64 codons, including 61 amino acid codons (including the start codon) and 3 stop codons. In this work, the stop codons are included when count the codon pairs. CPS is calculated by Eq. (2).

$$CPS = ln\left(\frac{AB}{\frac{A\times B}{X\times Y} \times XY}\right) \tag{2}$$

where A and B represent the number of codons A and B, respectively, and AB denotes the codon pair which is consisted by codons A and B. Similarly, X and Y represent the number of the corresponding amino acids X and Y, and XY as the amino acid pair across the genome sequences. To count the number of the codon pairs and the amino acids, the CPS determines if a given codon pair is over-represented or under-represented. For the given codons A and B, if the number of codon pair AB is larger than the amino acid pair XY, the CPS is a positive number represents the over-represented, on the contrast, a negative number as under-represented.

3.3.3. Codon bias

Codon bias is calculated for the 64 codons separately by dividing the number of each codon N_c by the total number of all codons across the complete sequences. The formulation is presented as Eq. (3)

$$Codon\ bias = \frac{N_c}{\sum N_c} \tag{3}$$

3.3.4. Amino acid bias

Same as the calculate way of codon bias, the amino acid bias for 21 amino acids (stop is also considered as an amino acid here) is calculated by dividing each amino acid N_a by the total number of amino acids in the selected sequences.

$$Aminoacid\ bias = \frac{N_a}{\sum N_a} \tag{4}$$

According to the quantification methods illustrated above, the genome sequences are transformed into 4229 traits (CPS = 4096, dinucleotide biases = 48, codon biases = 64, amino acid biases = 21).

3.4. Hybrid transfer learning model

In this section, we will give a description of our hybrid transfer learning model. As illustrated in Fig. 3, the proposed method can be further divided into two modules: domain adaptation and ensemble learning model.

3.4.1. Domain adaptation

In domain adaptation module, domain invariant feature is extracted by transfer learning method. Domain Adaptation is an important branch of feature-based transfer learning that improves the performance of classifier by adopting one or more source domains for the purpose of transferring information [18] and it is one of the most actively researched transfer learning methods [22,24,25]. The Maximum Mean Difference (MMD) [53] is a commonly used measurement to evaluate the distribution discrepancy across domains. By minimizing the MMD, the optimal mapping space can be found. Assume that source domain $\mathscr{D}_{\mathrm{S}} = \{\mathrm{X}_{\mathrm{S}}, \mathrm{Y}_{\mathrm{S}}\}$, where $X_S = \left\{x_j^S\right\}_{j=1}^{n}$ is the feature of the source domain and $Y_S = \left\{y_j^S\right\}_{j=1}^{n}$ is the corresponding label. Similarly, the target domain can be denoted as $\mathscr{D}_{\mathrm{T}} = \{\mathrm{X}_{\mathrm{T}}, \mathrm{Y}_{\mathrm{T}}\} \cup \{\mathrm{X}_{\mathrm{T}}^{\mathrm{U}}\}$, where the feature of the labelled data is $X_T = \left\{x_j^T\right\}_{j=1}^{l}$ and $Y_T = \left\{y_j^T\right\}_{j=1}^{l}$ is the corresponding label, $X_T^U = \left\{x_j^U\right\}_{j=l+1}^{l+u}$ is the feature of the unlabelled data. The distribution issues that occur between the source and target domains can be largely summarized in two different aspects: marginal distribution, namely $P(X_S) = P(X_T)$ and conditional distribution, namely $P(Y_S|X_S) = P(Y_T|X_T)$ [18].

Regarding the complex variances between different viral species, it is rare to tell the variance types for the given viral datasets. Therefore, Joint Domain Adaptation (JDA) [23] is adopted to reduce the variance through the perspective of the joint probability distribution, which takes both the marginal and conditional distributions into consideration. In addition, when mapping the source and target domain into the new feature space, JDA provides a method of dimensionality reduction which can effectively reduce the redundancy in our datasets. Unlike the traditional method which the target domain has no annotated datasets in JDA, we can use the limited label information from the target domain to give the pseudo labels for target testing sets and provide a refine adaptation result. The analysis is based on the conceptual framework of MMDE [27], which first takes advantage of kernel tricks to change the way of learning the mapping function φ into an optimal adaptation matrix $\mathbf{A}$. The detail computation methods for JDA is reported in [23]. After mapping into the new feature space, the new dataset can be notated as $\mathscr{D}_{\mathrm{S}} = \{\mathrm{A}^{\mathrm{T}}\mathrm{X}_{\mathrm{S}}, \mathrm{Y}_{\mathrm{S}}\}$ in the source domain, and $\mathscr{D}_{\mathrm{T}} = \{\mathrm{A}^{\mathrm{T}}\mathrm{X}_{\mathrm{T}}, \mathrm{Y}_{\mathrm{T}}\} \cup \{\mathrm{A}^{\mathrm{T}}\mathrm{X}_{\mathrm{T}}^{\mathrm{U}}\}$ in the target domain.

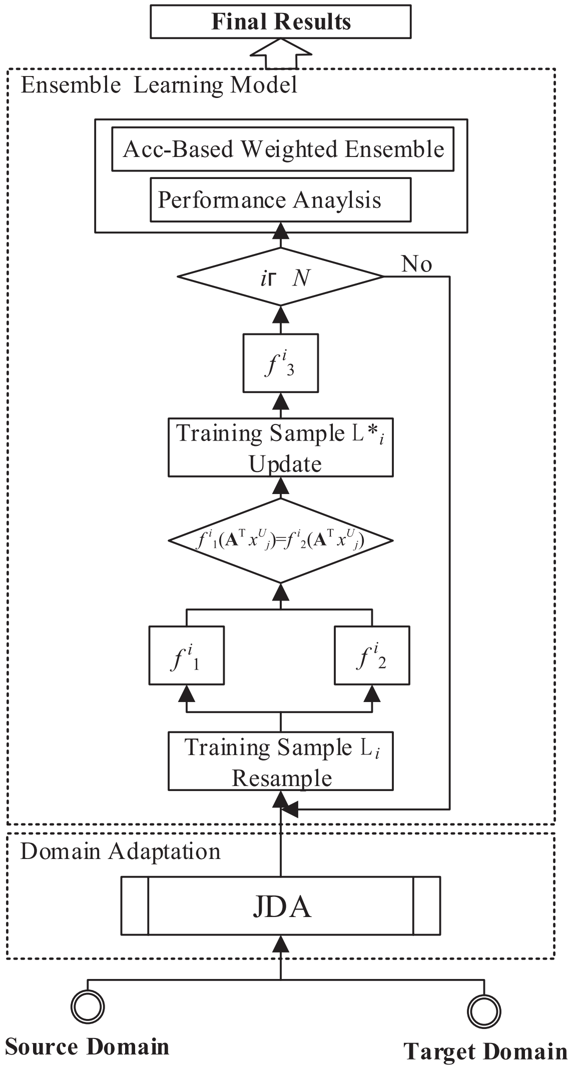

Final Results
Ensemble Learning Model
Acc-Based Weighted Ensemble
Performance Anaylsis
iɾ N
No
f^i_3
Training Sample L^*_i Update
$f^i_1(\mathbf{A}^T x^U_j)=f^i_2(\mathbf{A}^T x^U_j)$
f^i_1
f^i_2
Training Sample L_i Resample
Domain Adaptation
JDA
Source Domain
Target Domain

Fig. 3. The flow chart of hybrid transfer learning method.

3.4.2. Ensemble learning model

Although the distribution discrepancy between the source domain and target domain are substantially reduced, the class-imbalance problem and low generalization of model still constrains the performance of classifier on the target domain. To improve the performance on the target domain, we construct the ensemble learning model in three steps: First, during each iteration, the training sets $\mathscr{L}_{\mathrm{i}}$ is obtained using undersampling [48]. Specifically, for each process, we keep the target training sets remained and use a subset of the majority class in source domain. In this way, the class-imbalanced problem will be fixed, and the model can be robust to noisy samples. Second, based on the assumption that if two heterogeneous classifiers have consistency for labelling the testing sample, it can be considered as high confidence, we propose a tri-training strategy to exploit unlabelled data in target testing sets. These two heterogeneous classifiers are trained with training sets $\mathscr{L}_{\mathrm{i}}$. If the testing sample x_j^U fits the condition $\widehat{y}_j^U = f_1^i\left(\mathrm{A}^{\mathrm{T}}\mathrm{x}_{\mathrm{j}}^{\mathrm{U}}\right) = f_2^i\left(\mathrm{A}^{\mathrm{T}}\mathrm{x}_{\mathrm{j}}^{\mathrm{U}}\right)$, then the x_j^U is considered as the high confidence sample and added into the training sets $\mathscr{L}_{\mathrm{i}}$ with its corresponding pseudo label $\widehat{y}_j^U$ to form the new training set $\mathscr{L}_{\mathrm{i}}^*$. Then, the new training sets $\mathscr{L}_{\mathrm{i}}^*$ are utilized to train the classifier f_3^i which is used to give the prediction for all testing sets. We record the accuracy on the target training samples for the final weighting strategy. Repeat the former two steps until the number of iterations is reached. Third, all the outputs of f_3^i from each iteration are ensemble into a final output. The differences of each output mainly come from the sampling result and the quality of the high confidence samples in each iteration. To cut down the impact in each iteration and give a promising final output, we measure the performance of f_3^i on the accuracy of the target training sets Acc_i and define the weights w_i as follow:

$$w_i = \frac{Acc_i}{\sum_{i=1}^{N} Acc_i} \tag{5}$$

where $w_i > 0$ and $\sum_{i=1}^{N} w_i = 1$. The final output of hybrid transfer learning method is constructed by a linear combination of the output in each iteration which is shown in Eq. (6).

$$F(x_j^U) = \sum_{i=1}^{N} w_i f_3^i\left(\mathbf{A}^{\mathrm{T}}\mathrm{x}_{\mathrm{j}}^{\mathrm{U}}\right) \tag{6}$$

4. Simulation

Before giving the prediction for the hosts origins of COVID-19, some simulations need to be made for validation. We apply our hybrid transfer learning method to a collection of UCI medical benchmarks and our collected viral datasets. The experimental settings and the results are reported in Section A and B. After demonstrate the effectiveness of our method, the prediction results for COVID-19 are illustrated in Section C.

4.1. UCI medical benchmarks

To evaluate the properties of our approach in handling the large variances datasets, 13 representative medical datasets from the UCI repository [54] are selected to conduct the experiments. For the UCI datasets do not have any hierarchy, when splitting the data to generate the source and target domains, we use the methods in [19] and apply binary feature values in the datasets, such as age group, area to split the datasets and simulate inconsistent data distributions. In the last three groups of datasets, the attributes are all continuous variables. One way to separate the source domain and target domain is to select a given attribute and use K-means to cluster them into two partitions. Intuitively, these two partitions will have different data distributions. The information of concerning datasets is shown in Table 2 in detail.

Table 2
The descriptions of the UCI medical benchmarks.

Datasets	*Feature*	*Sample*	*Source domain*		*Target domain*	
			Pos.	*Neg.*	*Pos.*	*Neg.*
Autism	20	346	126	122	62	36
Heart Disease	12	270	83	100	67	20
Brest Cancer	8	277	32	30	164	51
Diabetic	18	1151	194	193	417	347
Sani	54	303	129	40	87	47
Thoracic	15	470	55	268	15	132
Lung	7	365	149	100	56	60
Colic	16	368	160	110	72	28
Cervical	32	668	37	535	8	88
Sick	27	2643	1633	131	798	81
Parkinson	22	195	31	19	116	29
Mammography	5	830	23	44	380	383
wdbc	14	569	185	93	172	119

For the main parameters in the domain adaptation method, the adopted experiments settings are showed as follows. We simulate the parameter adjustment methods in JDA [23] and seek the optimal parameters through an empirical approach. The iteration number of JDA is set as 10. The optimal λ is obtained from $\lambda \in \{0.01, 0.1, 1, 10\}$. The domain adaptation methods involve dimensionality reduction and we select $\{1/4, 1/2 \text{ and } 3/4\}$ of the initial data dimensions for dimensionality reduction and choose the best parameter for each model. Regarding kernel selection, the RBF kernel is chosen for the datasets *Autism*, *Lung* and *Cervical*, while the linear kernel is used for the rest of the datasets. We set f_1 as Logistic Regression (LR), f_2 as SVM, f_3 as Multilayer Perception (MLP), and the iteration number of the ensemble learning model is $N = 10$. The average accuracy and standard deviation of 10 times of experiments are used for evaluation, furthermore, we select 10% target samples as target training sets in each experiment and set the same random seed to guarantee the fairness.

To demonstrate the interpretability of HTL and figure out which part makes a major contribution in our method, we split every component of HTL and make a detail comparation for each output results. Firstly, we utilize the domain adaptation method (e.g., JDA) to decrease the variance between source and target domain, and use the data after feature mapping as input to train three heterogenous classifiers (e.g., LR, SVM, MLP). Secondly, the main purpose of constructing the ensemble learning model is to solve the class imbalance problems and enhance the reliability on testing datasets. Therefore, we compare the performance of HTL with the ensemble learning model without undersampling. Table 3 collectively lists all the experiment results. As shown in Table 3, our approach achieves significantly better performance in terms of both classification and standard deviations and wins on 10 out of 13. More specifically, being compared with the results of only adopting the domain adaptation method, the performances of our ensemble strategy show superior results which means by adding the pseudo labels, the classifier can be iteratively influenced to reach a better fitting for testing datasets. What's more, without the processing of undersampling, the prediction results may bias to the majority class which may cause the accuracy value higher than HTL. However, the bootstrap strategy could

Table 3
Classification accuracy of different components of HTL on benchmarks.

Datasets	*Domain adaptation*			*Ensemble learning model*	*HTL*
	LR	*SVM*	*MLP*	*Without undersampling*	
Autism	81.5 ± 3.5	79.1 ± 2.3	79.3 ± 3.8	87.3 ± 2.1	**89.1 ± 1.3**
Heart Disease	80.3 ± 3.1	79.8 ± 2.8	81.4 ± 3.6	82.1 ± 2.3	**83.8 ± 2.8**
Brest Cancer	72.3 ± 1.2	74.1 ± 2.3	73.1 ± 3.5	78.1 ± 2.2	**79.4 ± 1.2**
Diabetic	66.1 ± 0.6	65.2 ± 0.7	68.2 ± 1.3	69.4 ± 1.0	**70.6 ± 0.9**
Sani	70.2 ± 2.3	72.4 ± 1.9	71.5 ± 2.1	74.9 ± 1.8	**75.8 ± 2.4**
Thoracic	85.1 ± 1.1	83.2 ± 1.2	84.5 ± 2.3	**86.7 ± 1.5**	80.1 ± 1.7
Lung	72.1 ± 2.4	75.9 ± 3.5	76.3 ± 2.1	76.2 ± 1.4	**79.4 ± 2.2**
Colic	69.5 ± 1.9	70.8 ± 1.3	72.1 ± 1.5	76.3 ± 2.1	**78.7 ± 2.2**
Cervical	88.4 ± 2.2	87.5 ± 2.1	87.1 ± 3.6	**89.6 ± 2.7**	87.6 ± 1.2
Sick	93.2 ± 1.3	92.4 ± 2.5	93.2 ± 1.6	**94.1 ± 2.3**	88.1 ± 0.3
Parkinson	76.8 ± 3.1	78.5 ± 4.3	77.9 ± 3.5	79.3 ± 3.9	**80.5 ± 3.4**
Mammography	71.3 ± 2.3	72.4 ± 3.2	72.1 ± 2.7	73.5 ± 2.8	**74.1 ± 1.8**
wdbc	90.1 ± 0.4	91.6 ± 0.8	91.8 ± 1.2	91.3 ± 0.8	**92.3 ± 0.6**

Table 4
Classification accuracy of different transfer learning methods on benchmarks

Datasets	*TrAdaBoost*	*MTLF*	*CODA*	*HTL*
Autism	83.6 ± 3.3	87.6 ± 2.6	80.7 ± 2.3	**89.1 ± 1.3**
Heart Disease	70.8 ± 1.3	78.9 ± 1.1	75.4 ± 1.2	**83.8 ± 2.8**
Brest Cancer	69.1 ± 3.2	71.4 ± 3.6	73.1 ± 2.3	**79.4 ± 1.2**
Diabetic	62.8 ± 1.3	64.4 ± 1.2	65.7 ± 1.3	**70.6 ± 0.9**
Sani	72.5 ± 2.8	74.8 ± 1.7	73.4 ± 2.9	**75.8 ± 2.4**
Thoracic	**86.1 ± 1.8**	80.2 ± 2.4	82.1 ± 1.7	80.1 ± 1.7
Lung	67.2 ± 2.9	68.8 ± 1.7	77.4 ± 2.1	**79.4 ± 2.2**
Colic	74.7 ± 3.4	76.1 ± 1.0	75.7 ± 1.6	**78.7 ± 2.2**
Cervical	91.4 ± 0.9	**93.3 ± 0.5**	90.3 ± 0.7	87.6 ± 1.2
Sick	92.5 ± 0.7	**95.9 ± 0.1**	90.6 ± 0.2	88.1 ± 0.3
Parkinson	**83.4 ± 1.7**	81.5 ± 0.5	81.2 ± 1.2	80.5 ± 3.4
Mammography	70.4 ± 1.3	71.2 ± 1.6	69.8 ± 2.3	**74.1 ± 1.8**
Wdbc	89.8 ± 1.5	90.7 ± 0.9	90.5 ± 1.5	**92.3 ± 0.6**

Table 5
The compositions of datasets

Baseline	Training dataset $\mathscr{L}$	Testing dataset
SVM	$\{X_S, Y_S\} \cup \{X_T, Y_T\}$	$\{X_T^U\}$
tar-SVM	$\{X_T, Y_T\}$	$\{X_T^U\}$
XGBoost	$\{X_S, Y_S\} \cup \{X_T, Y_T\}$	$\{X_T^U\}$
tar-XGBoost	$\{X_T, Y_T\}$	$\{X_T^U\}$
HTL	$\{X_S, Y_S\} \cup \{X_T, Y_T\}$	$\{X_T^U\}$

enhance the generalization of the classifier on the testing data.

After demonstrate the effectiveness of HTL, we compare our method with other three state-of-the-art transfer learning algorithms for medical classification problems, i.e., TrAdaBoost [19], MTLF [55] and CODA [56]. TrAdaBoost is an instance-based transfer learning method which conducts a new mechanism to automatically adjust the weights of training samples. The other two algorithms belonging to feature-based transfer learning methods. The MTLF takes advantage of the Mahalanobis distance [57] to evaluate the distribution differences between the source and target domains instead of MMD in domain adaptation. The CODA conduct a feature selection strategy based on the Pearson correlation coefficient [58], then co-training is used to improve the classifier. We can easily see their effectiveness of narrowing the data variances when the instance-based methods loss its power, especially in datasets *Cervical* and *Sick*. Table 4 shows that TrAdaBoost wins on two datasets and the MTLF algorithm wins on two datasets, while our HTL approach achieves the best results on the other 9 datasets. While in datasets *Cervical* and *Sick*, both of the datasets have imbalanced class problem which may cause the high accuracy. In general, our method can reveal the superiority when compared with other transfer learning methods on benchmarks.

4.2. Experiment on virus datasets

In this part, we evaluate the effectiveness of HTL on viral origin tracing problems and compare it with other transfer learning methods and traditional machine learning methods. Since the discrepancy exists among different virus species, when predicting the newly discovered virus (e.g., COVID-19), the results obtained from the same virus species will be more reliable. However, in the case of limited training samples in coronavirus, only using coronavirus samples are insufficient to construct a robust classification model for COVID-19. Thus, the purpose of HTL is to construct a high generalized model for target domain (coronavirus) by leveraging the source domain (other related virus). The detail of the viral genome datasets and the domain partitions are reported in Table 1. By applying the quantification method described before, the collected refseqes are transformed into the numerical datasets with 4229 dimensions.

The best parameters of JDA are set as {kernel = linear, dimension = 200, $\lambda = 0.1$, gamma = 0.01}, the other parameter settings of HTL are same as Section A. The *SVM* and *XGBoost* are used as the base classifiers for comparison, since these two algorithms outperform than others on our datasets. The target training sets are randomly sampling from the target domain, while the rests are used for testing. To demonstrate the purposes of HTL and simulate the performances with other traditional machine learning methods, we project the experiment mainly from three perspectives. First, due to the information redundancy problem under the high-dimensional genetic datasets, we use 4 different feature inputs for comparison. The primitive feature represents the quantification results with 4229 traits and our HTL also uses the primitive feature for training. The other three feature processing methods comes from [14]. The Gradient Boosted Tree is used to make feature selection, which selects 50 most important features for training. The PN model [14] is adopted as a compared feature extraction method which uses the Blastn to analyse each genome sequence, then finds the top 5 similar virus from the training sets and uses their hosts information to generate the features for each sample. In [14], by combining selected features with viral PN, the classifier can get a better performance. Thus, these combined features are also used for comparison. Second, by setting the training sets

Table 6
Classification accuracy on virus datasets

Methods		***10 samples***	***20 samples***	***30 samples***
		ACC ± std	ACC ± std	ACC ± std
Primitive Feature	*SVM*	22.6%±5.6	24.4%±4.8	28.9%±5.3
	XGBoost	21.3%±2.4	28.6%±2.6	29.1%±5.5
	tar-SVM	28.9%±4.1	36.7%±4.5	39.3%±6.2
	tar-XGBoost	–	34.4%±3.9	31.5%±4.2
Feature Selection	*SVM*	25.2%±1.9	26.7%±3.3	24.5%±3.6
	XGBoost	24.7%±7.4	28.6%±5.8	32.3%±5.7
	tar-SVM	29.2%±4.5	28.1%±5.4	23.3%±4.7
	tar-XGBoost	–	32.2%±6.5	35.4%±7.6
PN Model	*SVM*	19.0%±2.1	25.5%±4.6	27.2%±5.5
	XGBoost	17.1%±3.9	29.0%±5.9	23.3%±6.6
	tar-SVM	10.3%±5.9	22.7%±5.6	27.8%±4.6
	tar-XGBoost	–	19.5%±5.5	25.1%±7.0
Feature Selection + PN Model	*SVM*	15.4%±4.4	22.0%±4.7	29.6%±6.4
	XGBoost	19.2%±2.9	33.9%±6.3	36.3%±7.7
	tar-SVM	27.1%±5.0	26.9%±6.6	33.9%±5.5
	tar-XGBoost	–	30.6%±5.3	38.4%±9.0
TrAdaBoost		30.1%±2.3	35.4%±3.3	44.7%±3.6
MTLF		34.7%±3.9	36.1%±4.5	48.1%±4.2
CODA		28.3%±5.9	30.4%±5.1	35.1%±4.8
***Our Method:** HTL*		**36.5%±4.3**	**40.7%±5.3**	**51.5%±3.2**

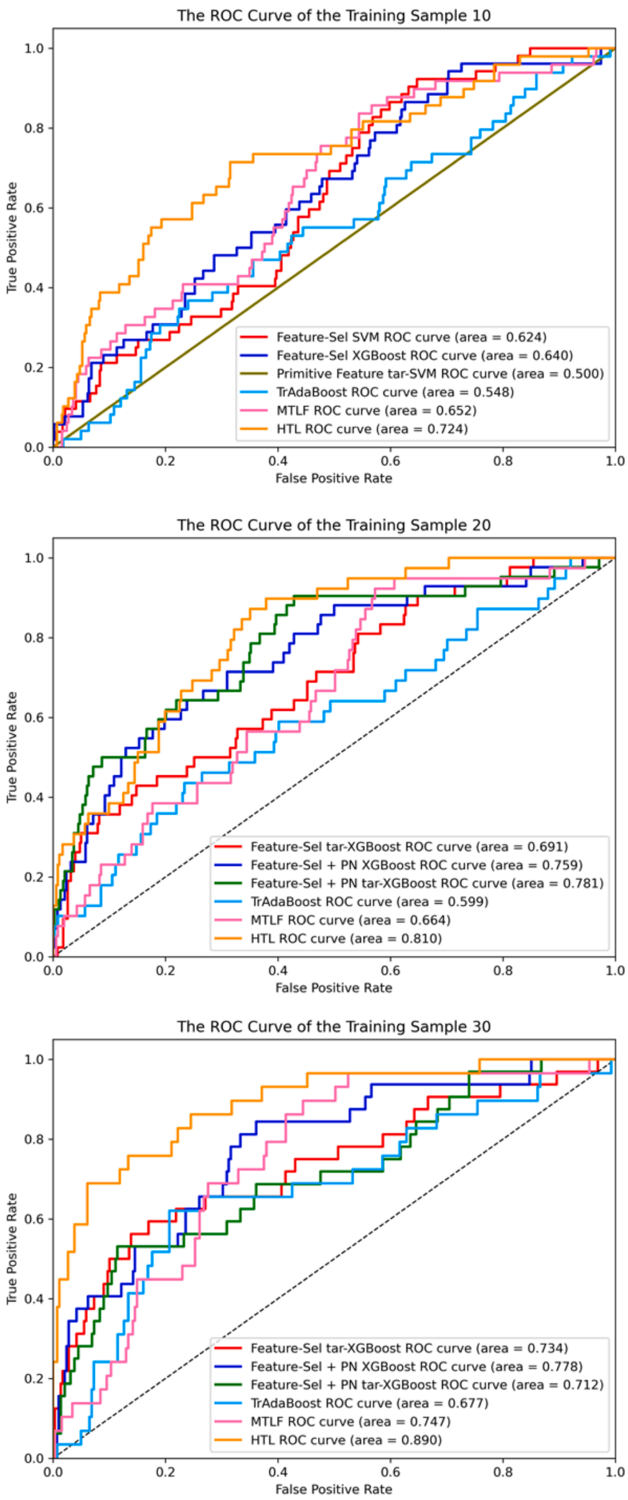


Fig. 4. The ROC curve of the different amounts of target training datasets. In each experiment group, Top-5 best performance methods are selected to compare the AUC with HTL.

with different compositions, we want to prove the constrains of the traditional machine learning methods in the conditions of limited training sets and unbalanced class problems. Furthermore, directly using the source domain may cause a negative effect on the classification model because of the discrepancies exist among different virus species. The compared baseline methods are implemented by *SVM* and *XGBoost*, Table 5 shows the different compositions of training datasets. Third, we conduct different amounts of target training datasets to prove the superiorities of our model in few sample problems. What's more, transfer learning methods are also adopted to make a further comparation. For each set of experiments, we randomly select a corresponding number (1, 2, 3) of target domain from each category to form 3 comparative experiments. We perform 10 epochs for all experiments and calculate the mean accuracy and standard deviation for them.

The final experiment results are collectively illustrated in Table 6. We can see from the Table 6 that HTL achieves significantly better performance than both the traditional machine learning methods and transfer learning methods. While analysing from the vertical direction of the Table 6, none of the adopted traditional machine learning methods can obtain a satisfactory prediction accurate on our datasets. For the four adopted feature inputs, the combination of feature selection with PN have a better performance than others especially when the target training samples reach to 30.

The genome feature quantification method calculates the intricate information for the sequence, which will lead to the information redundancy when training the classifiers. In the cases of small amounts of target training datasets, traditional methods don't have the abilities to build a highly generalized models and show a poor performance on target testing datasets. Note that, when there is only one training sample for each class, XGBoost fails to build a model for classification. For each leaf node contains different classes of samples except the last classification node, makes it hard to select a low entropy subset. Furthermore, the feature redundancy further aggravates the emergence of this phenomenon. The results of directly training the classifiers by source domain cannot achieve a better performance than only using the target training sets, even worse than the latter. This evidence means the source domain will cause a negative impact on learning the classifier for target testing sets. However, by conducting the transfer learning methods to decrease the variances among virus species, the accuracy can have a distinct improvement. MTLF achieves a closer result with HTL, which means the domain adaptation methods have a strong ability to decrease the variances and alleviate the feature redundancy. Compared from the horizontal of the Table 6, with the increase of target domain training data set, the accuracy of the algorithm increases gradually, especially for MTLF, which means the target training sets can bring more information to predict the hosts for testing sets.

To further demonstrate the performance of HTL, we select the Top-5 methods in each experiment group and compare their ROC curve with HTL. The experiment results are reported in Fig. 4. When the training samples in the target domain is lower than 30, only using the target training sets is insufficient to build a robust model. By leveraging the transfer learning methods, the source domain can be regarded as a better supplementary and improve the generalization on target domain. In our experiment, both TrAdaBoost and MTLF perform better than most of the traditional machine learning methods. Although, transfer learning can effectively decrease the discrepancy between source and target domain, class imbalance problem still constrains the conventional methods to get a better performance. We can discover from the Fig. 4 that the AUC of HTL always appears the best performance than other compared methods. Furthermore, the AUC value of our method is between 0.7 and 0.9, which indicates the prediction results of HTL has a high confidence.

In general, by decreasing the variance between target virus species and related virus species, our method outperforms than other compared traditional machine learning methods and transfer learning methods. The ensemble learning model guarantees the generalization and the robustness for the target testing set. What's more, the final ensemble strategy decreases the influence by the resample method which can be demonstrated in the standard deviation.

4.3. Prediction results for COVID-19

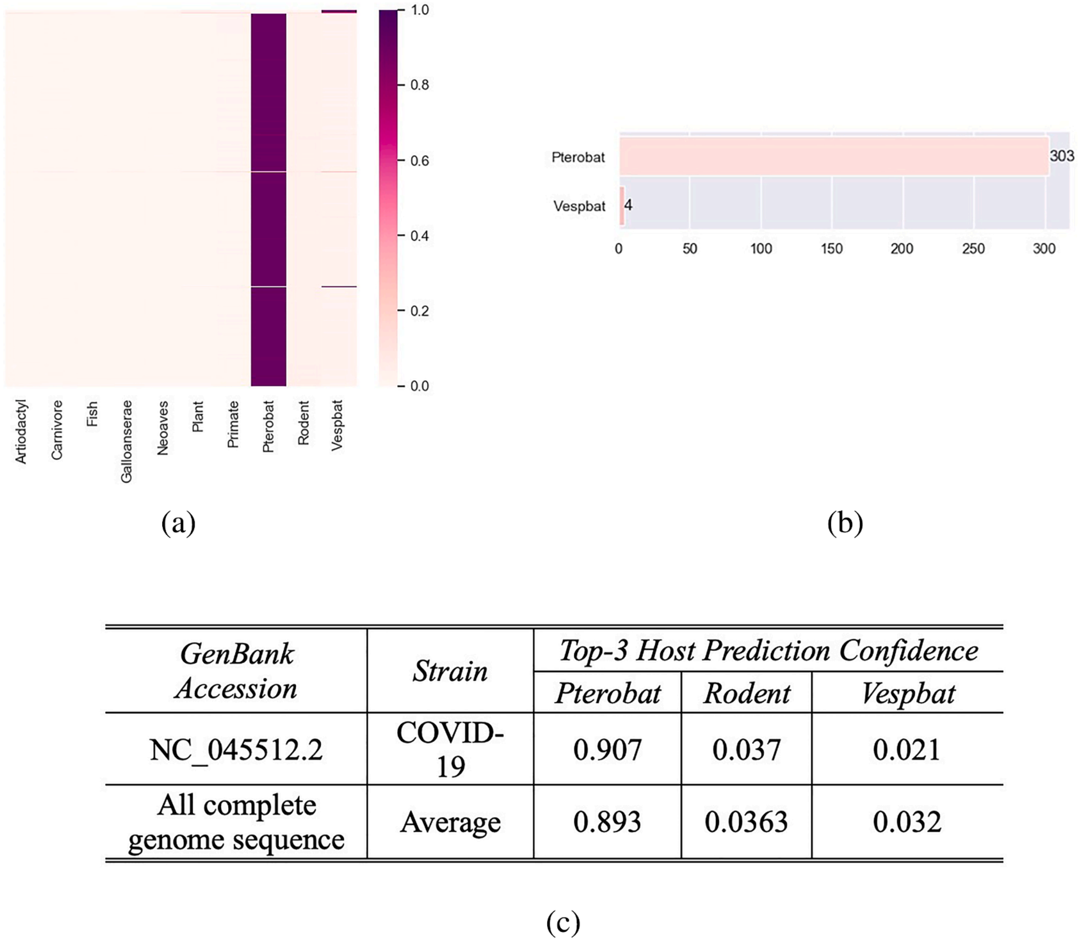
After evaluating the effectiveness and performance of the HTL on virus datasets, we conduct our method to provide the prediction results for COVID-19. All 307 complete genome sequences and 1 refseq of COVID-19 are pre-processed with the same methods described before. We keep the same experiment settings in Section B, but change the target training set into all coronavirus datasets. The experiment results of HTL get the similar prediction results like most mainstreams for COVID-19 that the reservoir hosts of COVID-19 may come from bat. More specifically, the reservoir host of the COVID-19 is Petrobat. While for the complete genome sequences, there are 303 prediction results for Petrobat and 4 results for Vespbat. As shown in Fig. 5 (a), we transform the output probabilities of HTL into a heat map to visualize the prediction information for the complete genome sequences of COVID-19. All the prediction results for Petrobat get high probabilities, which reveals the prediction results for Petrobat have high confidence. We can obviously find in Fig. 5 (c) that the Petrobat gets the highest mean confidence nearly 89.3%, the Rodent and the Vespbat get 3.6% and 3.2% prediction confidence for average. Looking back to the prediction results of refseq, the results are similar with the average probabilities of complete genome sequences, which the Petrobat gets 90.1%, the Rodent and the Vespbat get 3.7% and 2.1% respectively. The complete genome sequences contain many similar samples of COVID-19, which means Petrobat may be the most probable hosts.

To make our prediction results convincible, we use Blastsn to search the Top-5 closely related viruses from our datasets and show the results in Table 7. Compared with the refseq of COVID-19 (NC_045512.2), the SARS coronavirus has the highest sequence identity of 80.3% and other Bat coronavirus with 78% and 67.4% sequence identities. Beyond our expectations, there are two sequences of coronavirus with the hosts of Neoaves have the high identities with COVID-19. These two sequences are submitted in [59], which demonstrates the avian and mammalian coronaviruses have similar genome characteristics and structures. In their work, the bats and birds, the warm-blooded flying vertebrates, are ideal hosts for the coronavirus gene source. This conclusion presents another conjecture that whether avian could be the intermediate host or have some relationships with the COVID-19, it needs further studies. However, judging from the prediction results of our method, the Neoaves only has 0.3% probability to be the hosts for COVID-19.

Table 7
The sequence identities with the refseq of COVID-19 (NC_045512.2).

GenBank accession	*Virus species*	*Strain*	*Identity (%)*	*Hosts*
NC_004718.3	Coronavirus	SARS coronavirus	80.3%	Pterobat
NC_014470.1	Coronavirus	Bat coronavirus	78.0%	Pterobat
NC_016991.1	Coronavirus	White-eye coronavirus HKU16	68.9%	Neoaves
NC_016992.1	Coronavirus	Sparrow coronavirus HKU17	67.8%	Neoaves
NC_030886.1	Coronavirus	Rousettus bat coronavirus	67.4%	Pterobat

Phylogenetic analysis of COVID-19 revealed that the hosts of the closely related genomes can also support our prediction results. For a better visualization, we only adopt 20 refseqes of the closely related sequence in Fig. 6. All the bootstrap values are 100, which means the credibility when building the branches of the tree. We can see form the



(a)

(b)

GenBank Accession	*Strain*	*Top-3 Host Prediction Confidence*		
		Pterobat	*Rodent*	*Vespbat*
NC_045512.2	COVID-19	0.907	0.037	0.021
All complete genome sequence	Average	0.893	0.0363	0.032

(c)

Fig. 5. The statistics results of the prediction results and the heat map of the hosts probabilities for COVID-19. In (a), the depth of the color represents the hosts probability of each virus; (b) illustrates the prediction results of complete genome sequences of COVID-19; (c) records the Top-3 prediction confidence of refseq and average of complete genome sequences

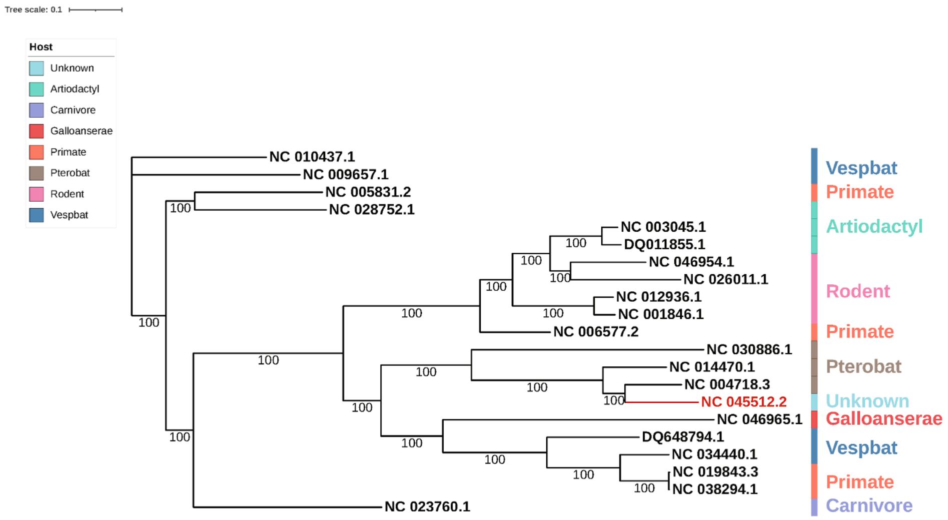


Fig. 6. The phylogenetic tree constructed by the refseq of COVID-19 (NC_045512.2) and Top-20 related sequences.

Fig. 6., the SARS coronavirus (NC_004718.3) [60] and the Bat coronavirus (NC_014470.1) [61] have the closest clades with COVID-19, the Rousettus bat coronavirus (NC_030886.1)[62] forms one clade close to the above three viruses. The hosts in the branch which constructed by these three sequences all comes from Pterobat. Thus, according to the phylogenetic tree, COVID-19 has the highest homology with the SARS coronavirus (NC_004718.3) which reveals the host of COVID-19 may come from Pterobat. The evidence from phylogenetic analyses and the sequence identities analyses have the high consistencies with the prediction result from our method, which indicates the Petrobat is the most probable hosts for COVID-19.

5. Discussion

As demonstrated in the reported experiments, our approach obtains better classification results not only for UCI benchmarks, but also the viral datasets. It provides a promising yet easy-to-use method for addressing the origin tracing problems for the newly discovered virus. Based on both the overall experimental results and the virology analyses, we summarize our approach as follows.

Traditional origin tracing methods consume a large number of times and labours, which delay the isolation for the hosts animals and the pathology research of the virus. Machine learning provides a precise way to discover the latent information from annotated genome sequences and make the prediction for specific tasks. According to our experiments, when predicting the hosts for newly discovered virus, the classification model trained from the same viral species will get more reliable results. However, under the conditions of limited annotated virus and class-imbalance problem, directly train the classifiers from other viral datasets will lose their efficiencies by the discrepancies among different viral species. By conducting a combination of transfer learning and ensemble learning, the fundamental weakness of traditional methods can be well solved. From Section 4, the simulation results indicate that our method provides a promising way to improve the accuracy and generalization on UCI medical benchmarks, especially the virus datasets.

Although HTL achieves a better performance than all the compared methods, some problems still need to be further studied: 1) For the variances exist among the different virus species, confusing the related virus as one domain group is worthy of discussion. Evaluating the relationships among different virus species and separating the related domain into multiple domains seems more convinced. 2) Though the JDA can significantly reduce the variances among different virus species, the huge calculated quantity consumes enormous computing resources. The large-scale genome datasets with high dimensions become hard to obtain the best transform matrix. As PN model does, the evaluating for the variances among different virus species may dig out from the perspective the sequence genomes. 3) In this paper, we adopt the CDS of the complete sequences for quantification. While for each CDS, some coding regions may have the low relationship with the hosts information, but some regions have more discriminative information which will benefit for hosts prediction. So, finding the CDS regions related to the host is of great significance for improving the accuracy of the algorithm. And other sequence quantification methods can be used for pre-processing in the future work.

To predict the hosts for the newly discovered virus, we provide a complete work flow from collecting the virus sequences, genome sequence processing, feature quantification and construction the classification model. The final output of HTL gets a high prediction convince that the Petrobat is the hosts for COVID-19. The sequence identity and phylogenetic analysis of the refseq of COVID-19 get the same results with our method, which convince the prediction results from HTL. However, despite the prediction results of the Petrobat, several facts suggest that another animal is acting as an intermediate host between bats and humans. No traces of bats were found in the Huanan seafood market where the COVID-19 first broke out. The latest research [63] reveal that Malayan pangolins could be the intermediate host for COVID-19, for they detected and found the associated coronavirus in pangolin samples, belonging to two subgenus of the COVID-19, in which a receptor binding domain has a closely relationship with COVID-19. In [64], the researchers found mink viruses have a closer infectivity pattern to COVID-19 by deep learning method. These consequences of infectivity pattern analysis illustrate that bat and mink could be the candidate reservoirs of COVID-19. From the above, on the basis of the prediction results from HTL, it seems likely that the Petrobat may be the initial hosts of COVID-19, and might have another intermediate host, like pangolins, mink or other wild animal, to transmit the COVID-19 to humans.

6. Conclusion

In this paper, we present a workflow of predicting the hosts for the newly discovered virus, including data collection, data processing, feature quantification and construct the classification model. We construct a virus benchmarks which can be used to proceeding further studies for virus. Considering the limited training samples and class-imbalance problems, we extend transfer learning and ensemble learning to construct a hybrid transfer learning model for virus hosts prediction. We achieve a promising simulation results with other compared methods for UCI benchmarks and virus datasets. Finally, the prediction results reveal that the Petrobat might be the reservoir host of COVID-19, which provides a research interests for virologists to do more investigates.

CRediT authorship contribution statement

Yun Yang: Conceptualization, Funding acquisition, Project administration. **Jing Guo:** Software, Methodology, Writing - original draft. **Pei Wang:** Supervision, Writing - review & editing. **Yaowei Wang:** Formal analysis. **Minghao Yu:** Data curation, Visualization. **Xiang Wang:** Investigation. **Po Yang:** Validation. **Liang Sun:** Funding acquisition, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors are grateful to GenBank, for all the adopted gene sequences and hosts information are downloaded from the website. We would like to acknowledge the financial support provided by the Natural Science Foundation of China (NSFC) under Grant No. 61876166, No.61663046 and No.91849132, CAMS Innovation Fund for Medical Sciences (2018-I2M-1-002), Beijing Hospital Nova Project (BJ-2018-139), Yunnan provincial major science and technology special plan projects: digitization research and application demonstration of Yunnan characteristic industry, under Grant: 202002AD080001.

References

[1] M. Woolhouse, E. Gaunt, Ecological origins of novel human pathogens, Crit. Rev. Microbiol. 33 (4) (2007) 231–242.

[2] A.M. Saéz, et al., Investigating the zoonotic origin of the West African Ebola epidemic, EMBO Mol. Med. 7 (1) (2015) 17–23.

[3] H.-D. Klenk, W. Garten, Host cell proteases controlling virus pathogenicity, Trends Microbiol. 2 (2) (1994) 39–43.

[4] B. Hu, et al., Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus, PLoS Pathog. 13 (11) (2017), e1006698.

[5] C.B. Reusken, et al., Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study, Lancet. Infect. Dis 13 (10) (2013) 859–866.

[6] P. Zhou, et al., Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin, BioRxiv, 2020. [Online]. Available: https://doi.org/10.1101/2020.01.22.914952.

[7] D. Benvenuto, M. Giovanetti, A. Ciccozzi, S. Spoto, S. Angeletti, M. Ciccozzi, The 2019-new coronavirus epidemic: evidence for virus evolution, J. Med. Virol. 92 (4) (2020) 455–459.

[8] J.-F.-W. Chan, et al., A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, The Lancet 395 (10223) (2020) 514–523.

[9] R. Lu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet 395 (10224) (2020) 565–574.

[10] M. Viana, et al., Assembling evidence for identifying reservoirs of infection, Trends Ecol. Evol. 29 (5) (2014) 270–279.

[11] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (6) (2015) 321–332.

[12] M.K. Leung, A. Delong, B. Alipanahi, B.J. Frey, Machine learning in genomic medicine: a review of computational problems and data sets, Proc. IEEE 104 (1) (2015) 176–197.

[13] C.L. Eng, J.C. Tong, T.W. Tan, Predicting host tropism of influenza A virus proteins using random forest, BMC Med. Genomics 7 (S3) (2014) S1.

[14] S.A. Babayan, R.J. Orton, D.G. Streicker, Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes, Science 362 (6414) (2018) 577–580.

[15] M. Zhang, L. Yang, J. Ren, N.A. Ahlgren, J.A. Fuhrman, F. Sun, Prediction of virus-host infectious association by supervised learning methods, BMC Bioinf. 18 (3) (2017) 60.

[16] F. Mock, A. Viehweger, E. Barth, M. Marz, “Viral host prediction with Deep Learning,” bioRxiv, 2019. [Online]. Available: https://doi.org/10.1101/575571.

[17] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.

[18] K. Weiss, T.M. Khoshgoftaar, D.D. Wang, A survey of transfer learning, J. Big Data 3 (1) (2016) 1–40.

[19] W. Dai, Q. Yang, G.R. Xue, Y. Yu, Boosting for transfer learning, in: International Conference on Machine Learning, 2007, pp. 193-200.

[20] B. Tan, Y. Song, E. Zhong, Q. Yang, Transitive transfer learning, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1155–1164.

[21] M.N.A. Khan, D.R. Heisterkamp, Adapting instance weights for unsupervised domain adaptation using quadratic mutual information and subspace learning, in: International Conference on Pattern Recognition, 2017, pp. 1560-1565.

[22] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Networks 22 (2) (2011) 199.

[23] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer feature learning with joint distribution adaptation, IEEE International Conference on Computer Vision (2014) 2200–2207.

[24] M. Long, J. Wang, G. Ding, J. Sun, P.S. Yu, Transfer joint matching for unsupervised domain adaptation, IEEE Conference on Computer Vision and Pattern Recognition (2014) 1410–1417.

[25] J. Wang, Y. Chen, S. Hao, W. Feng, Z. Shen, Balanced distribution adaptation for transfer learning, IEEE International Conference on Data Mining (2017) 1129–1134.

[26] Z. Zhao, Y. Chen, J. Liu, Z. Shen, M. Liu, Cross-people mobile-phone based activity recognition, in: International Joint Conference on Artificial Intelligence, 2011, pp. 2545-2550.

[27] S.J. Pan, J.T. Kwok, Q. Yang, Transfer learning via dimensionality reduction, AAAI 8 (2008) 677–682.

[28] J. Davis, P. Domingos, Deep transfer via second-order Markov logic, in: International Conference on Machine Learning, 2009, pp. 217-224.

[29] L. Mihalkova, R. J. Mooney, Transfer learning from minimal target data by mapping across relational domains, in: International Jont Conference on Artifical Intelligence, 2009, pp. 1163-1168.

[30] Z.H. Zhou, Ensemble methods: foundations and algorithms, Taylor & Francis (2012) 77–79.

[31] Z.-H. Zhou, M. Li, Tri-training: Exploiting unlabeled data using three classifiers, IEEE Trans. Knowl. Data Eng. 11 (2005) 1529–1541.

[32] Y. Yang, J. Jiang, Hybrid sampling-based clustering ensemble with global and local constitutions, IEEE Trans. Neural Netw. Learn. Syst. 27 (5) (2017) 952–965.

[33] Y. Yang, Z. Li, W. Wang, D. Tao, An adaptive semi-supervised clustering approach via multiple density-based information, Neurocomputing (2017).

[34] Y. Yang, J. Jiang, HMM-based hybrid meta-clustering ensemble for temporal data, Knowl.-Based Syst. vol. 56, no. C (2014) 299–310.

[35] Y. Yang, K. Chen, Time series clustering via RPCL network ensemble with different representations, IEEE Trans. Syst. Man & Cybernetics Part C 41 (2) (2011) 190–199.

[36] Y. Yang, K. Chen, Temporal data clustering via weighted clustering ensemble with different representations, IEEE Trans. Knowl. Data Eng. 23 (2) (2010) 307–320.

[37] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. vol. 33, no. suppl_1 (2005) D501–D504.

[38] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[39] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statistics, pp. 1189-1232, 2001.

[40] J.L. Geoghegan, S. Duchêne, E.C. Holmes, Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families, PLoS Pathog. 13 (2) (2017), e1006215.

[41] S. Roux, S.J. Hallam, T. Woyke, M.B. Sullivan, Viral dark matter and virus–host interactions resolved from publicly available microbial genomes, elife, vol. 4, p. e08490, 2015.

[42] J. Ren, K. Song, M. Deng, G. Reinert, C.H. Cannon, F. Sun, Inference of Markovian properties of molecular sequences from NGS data and applications to comparative genomics, Bioinformatics 32 (7) (2016) 993–1000.

[43] L. Wan, G. Reinert, F. Sun, M.S. Waterman, Alignment-free sequence comparison (II): theoretical power of comparison statistics, J. Comput. Biol. 17 (11) (2010) 1467–1490.

[44] G. Reinert, D. Chew, F. Sun, M.S. Waterman, Alignment-free sequence comparison (I): statistics and power, J. Comput. Biol. 16 (12) (2009) 1615–1634.

[45] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[47] J. Huang, A.J. Smola, A. Gretton, K.M. Borgwardt, B. Scholkopf, Correcting sample selection bias by unlabeled data, in: International Conference on Neural Information Processing Systems, 2006, pp. 601–608.

[48] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2) (2008) 539–550.

[49] G. Kudla, A.W. Murray, D. Tollervey, J.B. Plotkin, Coding-sequence determinants of gene expression in Escherichia coli, Science 324 (5924) (2009) 255–258.

[50] K. Katoh, J. Rozewicki, K.D. Yamada, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization, Briefings Bioinf. 20 (4) (2019) 1160–1166.

[51] I. Letunic, P. Bork, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation, Bioinformatics 23 (1) (2007) 127–128.

[52] D. Kunec, N. Osterrieder, Codon pair bias is a direct consequence of dinucleotide bias, Cell Reports 14 (1) (2016) 55–67.

[53] K.M. Borgwardt, A. Gretton, M.J. Rasch, H. Kriegel, B. Schölkopf, A.J. Smola, Integrating structured biological data by Kernel maximum mean discrepancy, Bioinformatics 22 (14) (2006) 49–57.

[54] C. Blake. UCI repository of machine learning databases [Online] Available: www.ics.uci.edu/~mlearn/MLRepository.html.

[55] Y. Xu, et al., A unified framework for metric transfer learning, IEEE Trans. Knowl. Data Eng. 29 (6) (2017) 1158–1171.

[56] M. Chen, K.Q. Weinberger, J. Blitzer, Co-training for domain adaptation, in: Advances in neural information processing systems, 2011, pp. 2456-2464.

[57] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, Chemometrics and Intelligent Laboratory Systems 50 (1) (2000) 1–18.

[58] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise reduction in speech processing: Springer, 2009, pp. 1-4.

[59] P.C. Woo, et al., Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus, J. Virol. 86 (7) (2012) 3995–4008.

[60] R. He, et al., Analysis of multimerization of the SARS coronavirus nucleocapsid protein, Biochem. Biophys. Res. Commun. 316 (2) (2004) 476–483.

[61] J.F. Drexler, et al., Genomic characterization of SARS-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences, J. Virol. (2010).

[62] J.O. Obameso, et al., The persistent prevalence and evolution of cross-family recombinant coronavirus GCCDC1 among a bat population: a two-year follow-up, Sci. China Life Sci. 60 (12) (2017) 1357–1363.

[63] K. Xiao, et al., Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins, Nature (2020) 1–7.

[64] H. Zhu, et al., Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm, BioRxiv, 2020. [Online]. Available: https://doi.org/10.1101/2020.01.21.914044.