VIRUS EVOLUTION

Virus Evolution, 2024, **10(1)**, 1–10
DOI: https://doi.org/10.1093/ve/veae005
Advance Access Publication 27 January 2024
Resources

Bases-dependent Rapid Phylogenetic Clustering (Bd-RPC) enables precise and efficient phylogenetic estimation in viruses

Bin Ma,[1,2] Huimin Gong,[1,2] Qianshuai Xu,[1,2] Yuan Gao,[1,2] Aohan Guan,[1,2] Haoyu Wang,[1,2] Kexin Hua,[1,2] Rui Luo,[1,2,*] and Hui Jin[1,2,*]

[1]State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, No.1 Shizishan Street, Wuhan, Hubei 430070, China and [2]College of Veterinary Medicine, Huazhong Agricultural University, No.1 Shizishan Street, Wuhan, Hubei 430070, China
https://orcid.org/0000-0001-8004-2662
*Corresponding authors: E-mail: luorui@mail.hzau.edu.cn; jinhui@mail.hzau.edu.cn

Abstract

Understanding phylogenetic relationships among species is essential for many biological studies, which call for an accurate phylogenetic tree to understand major evolutionary transitions. The phylogenetic analyses present a major challenge in estimation accuracy and computational efficiency, especially recently facing a wave of severe emerging infectious disease outbreaks. Here, we introduced a novel, efficient framework called Bases-dependent Rapid Phylogenetic Clustering (Bd-RPC) for new sample placement for viruses. In this study, a brand-new recoding method called Frequency Vector Recoding was implemented to approximate the phylogenetic distance, and the Phylogenetic Simulated Annealing Search algorithm was developed to match the recoded distance matrix with the phylogenetic tree. Meanwhile, the indel (insertion/deletion) was heuristically introduced to foreign sequence recognition for the first time. Here, we compared the Bd-RPC with the recent placement software (PAGAN2, EPA-ng, TreeBeST) and evaluated it in *Alphacoronavirus, Alphaherpesvirinae*, and *Betacoronavirus* by using Split and Robinson-Foulds distances. The comparisons showed that Bd-RPC maintained the highest precision with great efficiency, demonstrating good performance in new sample placement on all three virus genera. Finally, a user-friendly website (http://www.bd-rpc.xyz) is available for users to classify new samples instantly and facilitate exploration of the phylogenetic research in viruses, and the Bd-RPC is available on GitHub (http://github.com/Bin-Ma/bd-rpc).

Keywords: phylogenetic tree; recoding; insertion/deletion; new sample placement; simulated annealing.

Phylogenetic relationships among species have a pivotal role in almost every branch of biology (Kapli, Yang, and Telford 2020; Yang and Rannala 2012). Furthermore, when humans come across new virus species, they must analyze their phylogenetic relationships to classify and monitor the viruses for public health purposes (of 2020). To clarify the relationships among species, constructing a phylogenetic tree is crucial in evolutionary biology research (Cheon, Zhang, and Park 2020; Vakirlis et al. 2016). Over the past few decades, evolutionary theory and computational phylogenetics have advanced rapidly (Kobert et al. 2014; Aberer, Pattengale, and Stamatakis 2010; Felsenstein 1981; Huelsenbeck and Ronquist 2001; Kobert, Stamatakis, and Flouri 2017; Yang 1993; Yang 1994). Among these, the phylogenetic trees based on Maximum Likelihood (ML) and Bayesian inference methods are often regarded as the 'gold standard' for tree construction nowadays, such as IQ-TREE2, MrBayes, and MEGA (Kapli, Yang, and Telford 2020; Felsenstein 1981; Huelsenbeck and Ronquist 2001; Minh et al. 2020; Tamura, Stecher, and Kumar 2021). This approach provided a more accurate estimation with a realistic substitution model, and most phylogenetic developments were achieved in this framework (Yang and Rannala 2012; Chen, Lewis, and O 2014; Yang 2006).

Nevertheless, the heavy computational demand for these methods is still a serious drawback, which limits the applicability on large datasets (Kapli, Yang, and Telford 2020). Nowadays, the unprecedented accumulation of viral genome sequences calls for the development of speeding algorithms. More and more placement methods with fundamental advances, including UShER, MAPLE, PAGAN2, EPA-ng, and TreeBeST, have been developed to enhance efficiency (Loytynoja, Vilella, and Goldman 2012; Barbera et al. 2019; De Maio et al. 2023; Ruan et al. 2008; Turakhia et al. 2021). Among these methods, researchers spent plenty of effort on inspired innovation, including the mutation-annotated tree (UShER), a novel Felsenstein pruning algorithm (MAPLE), the phylogeny-aware graph algorithm (PAGAN2), the evolutionary placement algorithm (EPA-ng), and the constructing algorithm guided by species tree (TreeBeST). By placing the new samples into the reference tree or optimizing existing algorithms to construct the phylogenetic tree, the UShER and MAPLE achieved better computational efficiency in some specific situations (De Maio et al. 2023; Turakhia et al. 2021). However, the UShER and the MAPLE prefer to perform well on highly similar sequence datasets (De Maio et al. 2023; Turakhia et al. 2021). Meanwhile,

in this study, we found that the other methods (PAGAN2, EPA-ng, and TreeBeST) do not show enough accuracy and robustness in multiple datasets compared with the Maximum-likelihood tree constructed by IQ-TREE2 (Minh et al. 2020; Loytynoja, Vilella, and Goldman 2012; Barbera et al. 2019; Ruan et al. 2008). Therefore, efficient phylogenetic estimation with high accuracy is an important unsolved problem and particularly relevant during the epidemic.

In this work, we describe an efficient method that facilitates rapid and accurate placement, called Bases-dependent Rapid Phylogenetic Clustering (Bd-RPC). Here, we integrate various recoding methods and present a creative way of using the recoding method called Frequency Vector Recoding, which combines nucleotide base frequency into the recoding methods and takes the influence of transition–transversion bias into account. Furthermore, in this study, we shed new light on using indel characters and develop a novel algorithm for the problem of rooted variants in phylogenetic trees called the Phylogenetic Simulated Annealing Search algorithm. Here, we examined the effectiveness of each Bd-RPC module, including the Phylogenetic Simulated Annealing Search algorithm, Indel Recognition, and PCA Improvement in this study. In addition, we compared Bd-RPC's running time and accuracy to the other current state-of-the-art methods of phylogenetic placement (PAGAN2, EPA-ng, TreeBeST) in multiple genera (*Alphacoronavirus, Alphaherpesvirinae*, and *Betacoronavirus*) (Loytynoja, Vilella, and Goldman 2012; Barbera et al. 2019; Ruan et al. 2008). Bd-RPC worked well in accuracy and efficiency and maintained good stability across different datasets. Nowadays, the database of *Alphacoronavirus, Alphaherpesvirinae*, and *Betacoronavirus* has been posted to the website for new sample placement (http://www.bd-rpc.xyz).

Methods

Implementation of algorithms in Bd-RPC

In this study, we developed a software/website for new sample placement called Bd-RPC. To perform the placement function, we first classify the existing sequences, and this section is defined as 'Make Database', which is utilized to obtain the relationships among existing sequences (Fig. 1A and B). After that, we place new samples into the database; the second section is called 'Clustering New Sequences' of Bd-RPC (Fig. 1C and D).

In the 'Make Database' section, Bd-RPC approximates the phylogenetic distance using the sequences treated by the Frequency Vector Recoding and matches the background information, including taxonomy information or phylogenetic tree by using hierarchical clustering as well as the Phylogenetic Simulated Annealing Search algorithm (Fig. 1E and F). In this study, the sequences, highly matched to the background information, are defined as a cluster, and many clusters are merged into a database for further analysis.

In the 'Clustering New Sequences' section, Bd-RPC first realigns the existing sequences with the new samples using the multiple sequence alignment program (Multiple Alignment using Fast Fourier Transform, MAFFT) and analyses the indel changes to recognize the foreign sequences (Katoh and Standley 2013). One of the functions of the 'Clustering New Sequences' section is pathogen identification based on taxonomy with high speed. Bd-PRC calculates the distance among total sequences (containing new samples) using the same recoding methods as in the 'Make Database' section and classifies the new samples into the database's clusters through the minimum distance for microorganisms identification. The other function of the 'Clustering New Sequences' section is constructing the phylogenetic tree. Bd-RPC will extract the clusters for tree construction using IQ-TREE2 and combine them into the phylogenetic tree users provided as the output (Minh et al. 2020).

The detailed workflow of Bd-RPC is available in Supplementary Methods.

Data collection and phylogenetic analysis

As of 25 February 2021, 30,142 *Betacoronavirus* sequences with complete clinical data (Host, Collection Date, and Isolated Country) were obtained from the National Center for Biotechnology Information (NCBI) and GISAID (http://gisaid.org/). In this study, a total of 29,220 SARS-CoV-2 sequences were de-replicated using the mash algorithm of the dRep software, and the spike (S) gene was selected by python scripts (Olm et al. 2017). Here, the sequences with an ANI of 0.9992 were chosen, which contained a total of eight clades (S, L, V, G, GH, GR, GV, and O) of GISAID (Supplementary Figure S1A). Furthermore, the reference sequences and the first reported sequences (alpha, beta, gamma) of SARS-CoV-2 have been put into the total sequences. Meanwhile, the high-quality genome sequences with all three types of clinical data were selected in the other sequences. In contrast, the 127 sequences with any two types of clinical data were defined as the unclassified sequences for evaluation (Supplementary Figure S1B). As a result, 1,482 sequences of *Betacoronavirus* (Spike gene) were identified as high-quality genome sequences for database creation, and 127 sequences were chosen to evaluate the software in the clustering region. Here, high-quality sequences have been aligned using the MAFFT program, and the phylogenetic tree was generated by IQ-TREE2 for the dataset preparation of the 'Make Database' section in Bd-RPC (Supplementary Figure S1C) (Minh et al. 2020), (Katoh and Standley 2013). The clinical data and taxonomy information were collected by python scripts, and manual adjustments were made afterward (Supplementary Table S1 and Supplementary Table S2).

We introduced two other datasets, including DNA and RNA viruses (*Alphacoronavirus* and *Alphaherpesvirinae*) to the new sample placement for a comprehensive evaluation. Among these viruses, the ORF1ab gene of *Alphacoronavirus* and the genes associated with the DNA replication machinery (UL5, UL8, UL9, UL29, UL30, UL42, UL52) of *Alphaherpesvirinae* were selected separately by python scripts for the tests. Here, the sequences that contained full clinical data (Host, Collection Date, and Isolated Country) were defined as the high-quality sequences for database creation, and the other sequences were used to place into the database. The high-quality sequences were aligned using the MAFFT program, and the phylogenetic trees were generated using IQ-TREE2 (Minh et al. 2020), (Katoh and Standley 2013). Finally, 1,139 sequences of *Alphacoronavirus* were defined as high-quality sequences for database creation, and 171 sequences were used to place into the database (Supplementary Table S3 and Supplementary Table S4). Meanwhile, 197 sequences of *Alphaherpesvirinae* were used in database creation, and the other 217 sequences were selected for placement (Supplementary Table S5 and Supplementary Table S6).

Optional parameters evaluation of Bd-RPC

In this study, we compared the performance of Bd-RPC with different parameter settings. To evaluate the robustness of the database creation, simulated sequences with different lengths, phylogenetic distances, and nucleotide base preferences were generated based on the *Betacoronavirus* phylogenetic tree by Seq-gen for database evaluation (Rambaut and Grassly 1997). Here, ten replications of the experiment were performed for each simulation

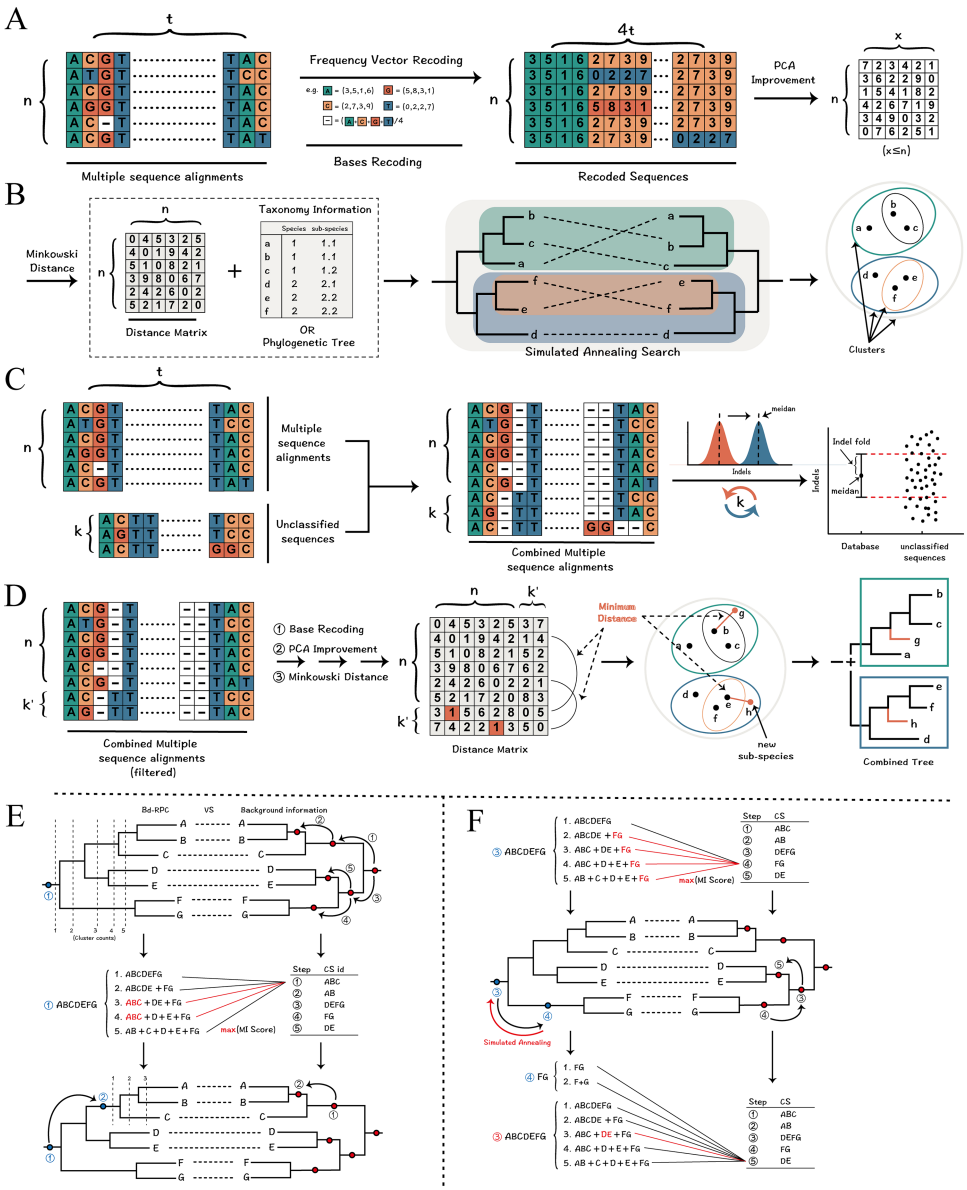

Figure 1. Overview of the Bd-RPC workflow. The Bd-RPC includes two section: 'Make Database' and 'Clustering New Sequences'. The Make Database section consists of two main components: (A) calculating the distance matrix and (B) matching the distance matrix with the background information. (A) In the first component, each base of multiple sequence alignments is initially recoded as a numeric vector based on user-defined methods, causing the width of the sequence matrix to increase substantially. Bd-RPC extracts the principle components to reduce the width of the sequence matrix by PCA Improvement, and the output matrix is used to calculate the distance matrix using Minkowski Distance. (B) Two types of background information are accepted (taxonomy information or phylogenetic tree), and the background information is matched with the hierarchical tree generated by the distance matrix using the Simulated Annealing Search algorithm (described in E and F). The cluster whose MI score is larger than the cutoff value is accepted for database creation, and the density of the cluster is defined as the max sequences' distance of this cluster. The Clustering New Sequences section consists of two main components: (C) Indel Recognition and (D) new sample placement. (C) Unclassified sequences are inserted into multiple sequence alignments using the MAFFT to generate a combined multiple sequence alignment for downstream analysis. Bd-RPC counts the indel change resulting from each new sequence insertion and compares the median of the change to the Indel Fold to assess whether the sequence belongs to the database. (D) The distance matrix is calculated using the combined multiple sequence alignments (filtered) based on the database creation settings. The minimum distance (orange numbers) of each new sample is selected and compared with the density of the corresponding cluster to place the new samples in the existing clusters. IQ-TREE2 will construct the subtree tree using the new sequences and the sequences from the corresponding cluster, if the phylogenetic tree is provided by users to the Bd-RPC. The subtrees will then be merged into a single tree as the output. For the Simulated Annealing Search algorithm, (E) the tree on the left is the hierarchical clustering result (dotted line) of the Bd-RPC distance matrix, and the one on the right is the tree transformed by background information including taxonomy information and the phylogenetic tree. The point on the branch (Bd-RPC hierarchical result and Background information) is the searching point for matching trees. For each searching point of the background information, the corresponding sequence (CS) ids are collected and compared with the sequence id of hierarchical results performed by Bd-RPC. The MI score of each clustering result is calculated independently, and the Bd-RPC searching point jumps forward to the maximum value for further searching. (F) After each jump of the Bd-RPC searching point, the Bd-RPC will perform a deeper search at the previous searching point (red arrow), thus reducing the impact of local maximum. By comparing the MI scores of the two searching points of Bd-RPC, the branch with the maximums value is chosen for the next jump.

parameter. For sequence length, we simulated different sequence datasets ranging from 1 Kbp to 1 Mbp. For phylogenetic distance, the sequence length was set to 5 Kbp and the average phylogenetic distances of the simulated sequences ranged from 0.1 to 3 times the average phylogenetic distance of *Betacoronavirus*. Meanwhile, for nucleotide base preferences, the sequence length was also set to 5 Kbp and the frequency of each base nucleotide varied from 0.1 to 0.7. The ratio of high-quality clusters (Matching Identity [MI] score >0.8) was used to determine the performance of Bd-RPC.

To assess the robustness of Indel Recognition, we first inserted other genera sequences into the dataset and collected the number of indels for performance evaluation (Supplementary Table 7). Additionally, we removed one species from each subgenus (HKU1, MERS-CoV, GCCDC1, SARS-CoV-2) and reinserted them into the residual database (337 sequences) for further assessment.

To evaluate the performance of recoded and uncoded genetic distance, we first selected the distance (Recoding Method 1, P, and K80 distance) and compared it against the patristic distance derived from the ML tree in three datasets: *Betacoronavirus, Alphacoronavirus*, and *Alphaherpesvirinae*. Furthermore, we used all seventeen uncoded genetic distance metrics available in the R package 'ape' for more comprehensive comparison with the six recoded distance metrics mentioned in this study (Paradis, Claude, and Strimmer 2004). In this comparison, the high-quality cluster ratio and new sample placement ratio were employed to evaluate the performance of these distance metrics in all three datasets.

Performance evaluation of Bd-RPC

We generated different ways to evaluate the performance of the Bd-RPC with different types of databases. For the database created by taxonomy information, we offered two different datasets to evaluate Bd-RPC. Firstly, we selected data up to 2015 for database creation (409 sequences) and placed the remaining data (1,073 sequences) for software performance evaluation (Supplementary Figure S2). Secondly, we chose total high-quality sequences for the database creation (1,482 sequences), and 127 sequences, which only carry two clinical data, were treated as new samples for testing (Supplementary Figure S1B).

In terms of Bd-RPC using the database created by the phylogenetic tree, we compared Bd-RPC to four other recent placement software: EPA-ng v.0.3.8; PAGAN2 v.1.53; TreeBeST v.1.92; and MAPLE v.0 18 (Minh et al. 2020; Loytynoja, Vilella, and Goldman 2012; Barbera et al. 2019; De Maio et al. 2023; Ruan et al. 2008). To assess the adaptability of Bd-RPC, we picked DNA as well as RNA viruses (*Alphacoronavirus, Betacoronavirus*, and *Alphaherpesvirinae*) in this comparison and employed various genes (ORF1ab, Spike, and Concatenation of UL5, UL8, UL9, UL29, UL30, UL42, UL52) to evaluate the software's robustness. Here, the MAPLE v.0.18 is not suitable for the datasets whose branches are typically more than 0 01 (De Maio et al. 2023), and a detailed list of commands used to perform each algorithm can be found in Supplementary Table S8. Furthermore, the running time of phylogenetic placement was also collected by the python scripts. The accuracy of these placements was evaluated using the Robinson-Foulds distance and Split distance, which was assessed using the R package 'TreeDist' and the software TOPD/FMTS (Robinson and Foulds 1981; Smith and Schwartz 2020; Bogdanowicz and Giaro 2011; Puigbo, Garcia-Vallve, and McInerney 2007). Meanwhile, the statistical analysis among each result of the software was performed using the one-sided paired Student's *t*-test.

Program implementation

Bd-RPC was developed using Python 3 and tested in macOS and Linux operating systems. In this study, all analyses were run on a 40-core Ubuntu 18.04.5 system with 187 GB of RAM, and the code is available on GitHub (http://github.com/Bin-Ma/bd-rpc). The Bd-RPC website was designed and implemented using the Django framework (https://www.djangoproject.com), which improves maintainability, extendibility, and portability. The website has been tested in several web browsers, such as Firefox, Google Chrome, and Internet Explorer.

Results

Evaluation of the 'Bases Recoding' module in the 'Make Database' section

In this study, six recoding methods were designed for distance estimation, and the distance displayed a high correlation (Pearson correlation coefficient >0.96) to the patristic distance measured from the phylogenetic tree (Table 1 and Supplementary Figure S3). Moreover, Method 1 as the default method, which showed the highest Pearson correlation coefficients with 0.965, was selected for further analysis. At the same time, it can be shown that the recoding method can be used to approximate phylogenetic distance with high accuracy.

Evaluation of the 'PCA Improvement' module in the 'Make Database' section

After the Base Recoding, the principal component analysis (PCA) was applied to speed up the distance estimation (Hotelling 1936). The running time showed that the PCA Improvement using 1,482 components enhanced computational efficiency significantly with a *P*-value of $1^{*}10^{-47}$ for the one-sided Student's *t*-test (Supplementary Figure S4A). As the length of the sequence rose from 1 kbp to 1 Mbp, the reduction time due to PCA Improvement increased from −1.3 s to 4450.6 s (Supplementary Figure S4B). In terms of the accuracy assessment, we calculated the Pearson Correlation Coefficient between the distance calculated by each number of components and the raw recoded distance using the Euclidean Distance (Supplementary Figure S4C). It can be found that the Pearson Correlation Coefficient by using the PCA Improvement with 1,482 components was equal to 1, representing a strong positive correlation. This shows that the 'PCA Improvement' can significantly improve computational speed while maintaining accuracy. Meanwhile, the recoded sequences were simplified using the PCA Improvement with 1,482 components for further analysis.

Evaluation of the 'Phylogenetic Simulated Annealing Search' algorithm in the 'Make Database' section

The Phylogenetic Simulated Annealing Search algorithm was used to maximize the MI score in terms of the database created by the phylogenetic tree, as shown in 'Supplementary Methods'. Comparing the number of high-quality clusters, the database using the Phylogenetic Simulated Annealing Search algorithm showed a significant increase in the proportion of high-quality clusters from 87.5 per cent to 93.4 per cent (Supplementary Figure S4D). It can be inferred that the Phylogenetic Simulated Annealing Search algorithm can effectively improve the matching between the hierarchical distance and the phylogenetic tree.

Evaluation of the 'Make Database' section using the simulated datasets

To evaluate the stability of Bd-RPC's database creation, we generated 1,482 simulated sequences based on the *Betacoronavirus* phylogenetic tree with various lengths, phylogenetic distances,

Table 1. Recoding methods and the correlation against patristic distance.

	Recoding methods	Recoded distance		
	A	Transitions	Transversions	Pearson correlation
Method 1	$(1-\pi_A, 0, 0, 0, 1-\pi_A, 0)$	$\sqrt{(1-\pi_i)^2+(1-\pi_j)^2+(\pi_i-\pi_j)^2}$	$\sqrt{2\cdot(1-\pi_i)^2+2\cdot(1-\pi_j)^2}$	0.9650
Method 2	$(1-\pi_A, 0, 0, 0)$	$\sqrt{(1-\pi_i)^2+(1-\pi_j)^2}$	$\sqrt{(1-\pi_i)^2+(1-\pi_j)}$	0.9624
Method 3	$(\pi_A, 0, 0, 0, \pi_A, 0)$	$\sqrt{\pi_i{}^2+\pi_j{}^2+(\pi_i-\pi_j)^2}$	$\sqrt{2\cdot(\pi_i{}^2+\pi_j{}^2)}$	0.9627
Method 4	$(\pi_A, 0, 0, 0)$		$\sqrt{\pi_i{}^2+\pi_j{}^2}$	0.9606
Method 5	(1, 0, 0, 0, 1, 0)	$\sqrt{2}$	2	0.9645
Method 6	(1, 0, 0, 0)	$\sqrt{2}$	$\sqrt{2}$	0.9620

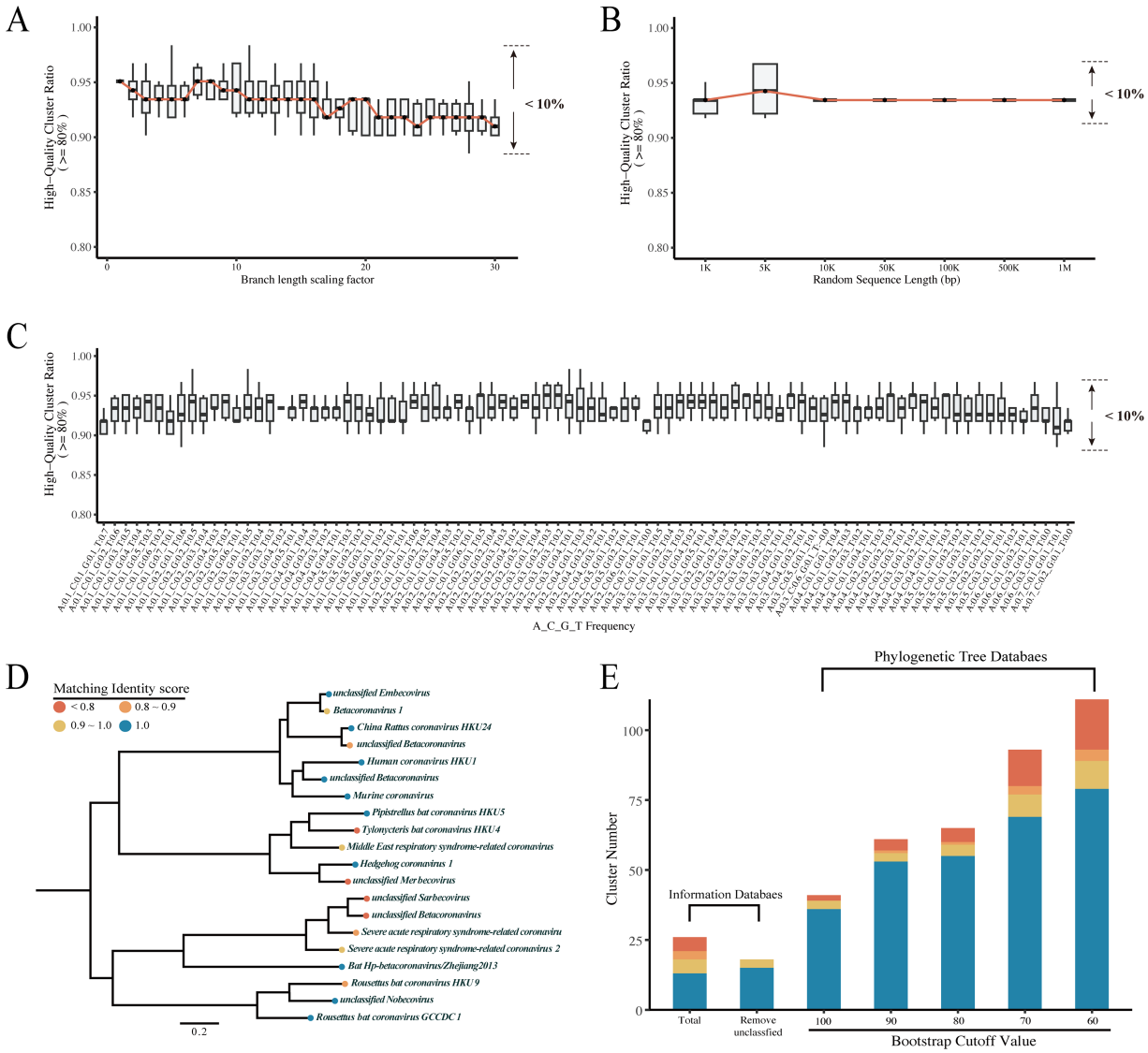

Figure 2. Performance evaluation of database creation. (A–C) The Simulated sequences and the high-quality cluster ratio (MI score ≥0.8) of the phylogenetic database were used to evaluate the robustness of Bd-RPC. (A) The x-axis represents the branch length scaling factor to simulate the sequences with various phylogenetic distances, and the ratio of high-quality clusters is represented on the y-axis. (B) Each box on the polyline corresponds to the simulation sequence's length, shown on the x-axis, and the high-quality cluster ratio is shown on the y-axis. (C) The frequency of the four bases is changed from 0.1 to 0.7, and detailed information of base frequency is available on the x-axis. Each box represents ten simulation sequences with specific base frequency, and the high-quality cluster ratio is shown on the y-axis. (D) The Betacoronavirus ML tree was generated using a random sequence in each species, and the colorful dots reflect the Matching Identity score of each cluster. (E) Statistical results of the Matching Identity score in the databases, established using two types of background information, were displayed as a stacked bar chart. The first and second bar represent the results with or without unclassified species, and the others show the distribution of the Matching Identity score under different boot-strap cutoff values.

and nucleotide base preferences to evaluate the robustness of Bd-RPC (Fig. 2A, B, and C).

For each value of average phylogenetic distances, the ratio fluctuations caused by phylogenetic distance changes were within a small range size (<10 per cent) (Fig. 2A). Similarly, the results of the ratio calculated in different sequence lengths also showed a small fluctuation (<10 per cent) in Bd-RPC, emphasizing the stability of Bd-RPC (Fig. 2B). Meanwhile, to evaluate the impact caused by nucleotide base frequency on database creation, the percentage of high-quality clusters in each database was collected. The results showed that as the frequency of the four bases changed from 0.1 to 0.7, the ratio of high quality clusters maintained a small variation (<10 per cent), highlighting the stability of Bd-RPC (Fig. 2C). Altogether, it was suggested that Bd-RPC maintained the robustness in the section 'Make Database'.

Evaluation of the 'Make Database' section using the *Betacoronavirus* dataset

Specifically, for the database created by taxonomy information, we found that the majority of species had great MI scores (>0.8), and some species with low MI scores (<0.8) were most defined as 'unclassified' by NCBI (Fig. 2D). To explore the influence of unclassified species, we removed sequences defined as 'unclassified' by NCBI from the input data and noticed that the MI score had a great improvement (MI of all clusters >0.9), as shown in Fig. 2E. It can be inferred that the MI score can be used to evaluate the consistency between the clustering result of recoded distance matrix and taxonomy information.

Besides, for the database created by phylogenetic tree, the bootstrap cutoff value is optional for users, and the program will stop searching if the branch's bootstrap value is less than the bootstrap cutoff value. As shown in Fig. 2E, the results indicated that most clusters' MI score was greater than 0.8, and the cluster with a high MI score (>0.8) was defined as a high-quality cluster in this study. As the bootstrap cutoff value increased from 60 to 100, the number of high-quality clusters was distributed from 39 to 93, while the number of high-quality clusters identified by taxonomy information was 18 or 21. Aside from these findings, as the bootstrap cutoff value decreased, the percentage of clusters with low MI scores increased from 4.8 per cent to 16.2 per cent, which may result in misclassifications. As a result, the database created using the phylogenetic tree contained more numbers of high-quality clusters, which performed better than the database created using taxonomy information. Furthermore, the results showed that decreasing the bootstrap cutoff value would cause misclassifications that may affect the accuracy of Bd-RPC.

Evaluation of the 'Indel Recognition' module in the 'Clustering New Sequence' section

In the 'Clustering New Sequence' section, Bd-RPC first removed the foreign sequences by Indel Recognition. To identify the foreign sequence among the new samples, the indel number was re-counted after the alignment with the addition of the new sequences. In *Betacoronavirus*, foreign sequence recognition was determined by the default Indel Fold, defined as the median fold-change (±10 per cent) of existing sequences' indels (Supplementary Figure S5). Here, Supplementary Figure S5A showed the movement of the indel quantity distribution after inserting foreign sequence.

By analyzing the median indel number of each sample, it can be found that the sequences have achieved a correct classification by comparing to the taxonomy information of NCBI (Supplementary Figure S5B). To assess the validity of Indel Recognition, four species of *Betacoronavirus* were removed from the database. Supplementary Figure S5C indicated that the four species treated as the new samples in this assessment were correctly classified into the *Betacoronavirus* by performing Indel Recognition. In summary, these results demonstrated the feasibility and high classification accuracy of the Indel Recognition module.

The accuracy and efficiency evaluation of Bd-RPC with database constructed by taxonomy information

After removing the foreign samples through Indel Recognition, the remaining sequences were classified into clusters of the existing database. It was found that the number of *Betacoronavirus* sequences rose sharply after 2015 (Supplementary Figure S2). Here, 1,066 new sequences were correctly classified within about 3 minutes. Only seven samples, which belong to the species with few sequences uploaded in NCBI before 2015, were mostly classified into the correct sub-genus (6/7) but not species (Supplementary Figure S6). Among these, SARS-CoV-2 was an outbreak in November 2019 and was first deemed as SARS-like Coronaviruses in 2020 (of 2020). Based on the existing sequences before 2015, Bd-RPC correctly classified the SARS-CoV-2 sequences into the SARS-CoV. Meanwhile, the BtRs-BetaCoV and Pangolin coronavirus were deemed 'unclassified' in NCBI, which were identified as SARS-like CoV after the worldwide outbreak of SARS-CoV-2 (Lam et al. 2020; Zhou et al. 2020). This study successfully classified these sequences into the SARS-CoV by Bd-RPC using the existing sequences before 2015.

Moreover, the 127 unclassified sequences with any two types of clinical data were employed as the new samples for evaluating the robustness of Bd-RPC performance. Here, all 127 samples were classified into the clusters of the existing database with 1,482 sequences using about 1 minute, and the Bd-RPC correctly identifies new samples into the corresponding species compared to the taxonomy information from NCBI (Fig. 3). The Bd-RPC using the database created by taxonomy information (user provided) enables rapid and accurate identification of new samples.

The comparation of frequency vector recoding to the other distance metrics

To evaluate the performance of recoded versus uncoded genetic distance, the recoded distance (Method 1) and uncoded genetic distance (P and K80 distance) were selected to compare against the patristic distance derived from the ML tree. As shown in Supplementary Figure S7, these distance metrics were evaluated in three datasets (*Betacoronavirus*, *Alphacoronavirus*, and *Alphaherpesvirinae*). When the patristic distances between sequences are close to zero, the recoded distance using Frequency Vector Recoding shows a steeper slope compared to the other two distances (P and K80 distance). This indicates a higher discriminative power of recoded distance for sequences with close patristic distances.

Moreover, we evaluated the performance of six recoded and seventeen uncoded genetic distance metrics in Bd-RPC using high-quality cluster ratio and new sample placement ratio. In *Betacoronavirus*, while all distance metrics enabled completely new sample placement, six recoded distance metrics outperformed the uncoded genetic distance metrics in terms of high-quality cluster ratio (Supplementary Figure S8A). Statistical analysis confirmed that these six recoded distance metrics significantly surpassed the uncoded genetic distance metrics in high-quality cluster ratio (Supplementary Figure S8B). Similar trends were observed in *Alphacoronavirus*, where the six recoded distance metrics significantly outperformed the uncoded genetic distance metrics in

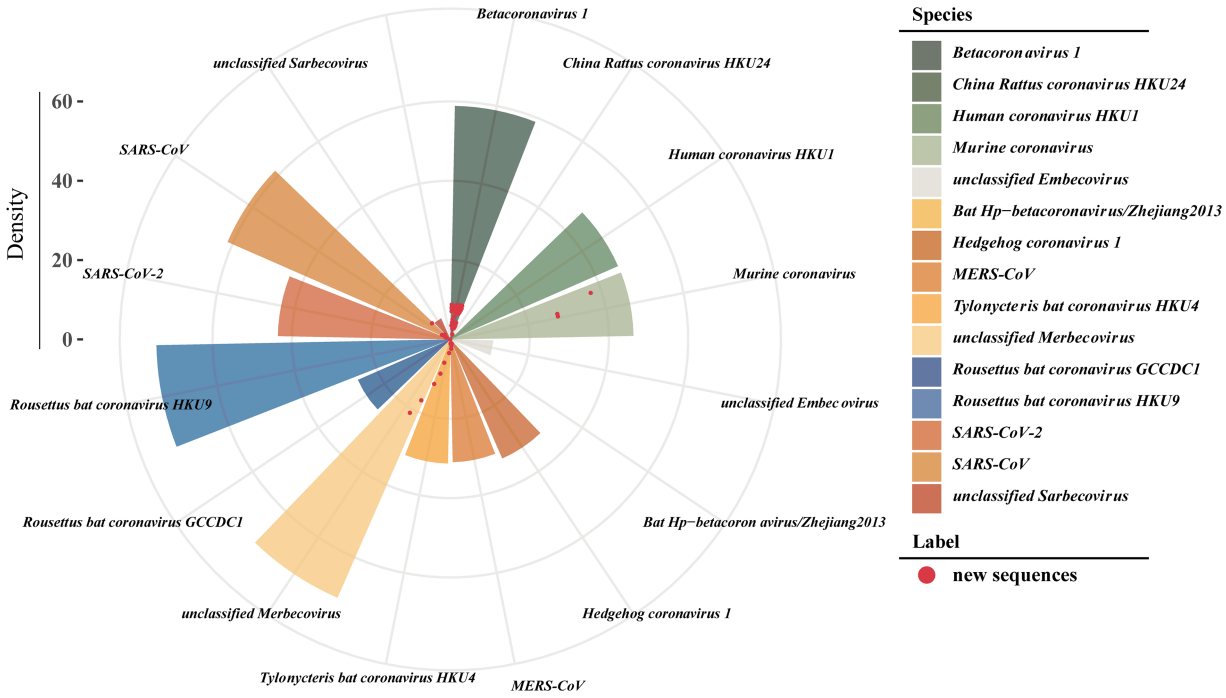

Figure 3. Accuracy evaluation of new sample placement in taxonomy database. The radar chart displays the 1.5-fold density of clusters (MI score ≥0.8, default value) and the shortest distance of new samples (red dot) to the cluster in the database. The colorful sectors show the clusters in the graphic, and the radius (distance-to-center) denotes the density of the clusters (new samples).

terms of high-quality cluster ratio (Supplementary Figure S8C and D). Additionally, it was noted that most distance metrics failed to place all new sequences in *Alphaherpesvirinae*. In this dataset, recoded distance metrics outperformed uncoded genetic distances in terms of new sample placement ratio but were inferior to some uncoded genetic distance metrics in high-quality cluster ratio (Supplementary Figure S8E and F).

The comparation of Bd-RPC to the other placement software

This study compared the recent developments of phylogenetic placement software (PAGAN2, EPA-ng, and TreeBeST) with Bd-RPC in running time and accuracy. Multiple datasets, including *Alphacoronavirus*, *Alphaherpesvirinae*, and *Betacoronavirus*, were selected to evaluate the performance and robustness of Bd-RPC in new sample placement with five repetitions. The Split and Robinson-Foulds distances, frequently employed to quantify variation between phylogenetic trees, were used to assess the accuracy of software (Robinson and Foulds 1981; Smith and Schwartz 2020; Bogdanowicz and Giaro 2011). In this study, the recoded distance metric (Method 1) was used in *Alphacoronavirus* and *Betacoronavirus*, while the 'logdet' distance metric was employed for *Alphaherpesvirinae* (Paradis, Claude, and Strimmer 2004).

In the *Betacoronavirus* datasets, the running time of Bd-RPC was significantly faster than PAGAN2 and TreeBest (Fig. 4A). Meanwhile, Bd-RPC achieved the highest similarity to the ML tree compared with other placement software (PAGAN2, EPA-ng, and TreeBeST) through the Split and Robinson-Foulds distances (Fig. 4B). Similar results were observed in *Alphacoronavirus*, where the Bd-RPC achieved the highest accuracy and maintained high efficiency (faster than PAGAN2 and TreeBest) (Fig. 4C and D). Furthermore, in the *Alphaherpesvirinae* dataset, Bd-RPC showed the highest accuracy of all placement software and achieved better efficiency than PAGAN2 (Fig. 4E and F). Overall, these findings showed that Bd-RPC maintained a good efficiency compared with the other placement software in viruses. Notably, Bd-RPC produced the most stable and accurate results among all methods in viruses.

Discussion

In this study, we provided a highly flexible, efficient computational software/website for new sample placement called Bd-RPC. Here, we heuristically introduced Frequency Vector Recoding, Indel Recognition, and the Phylogenetic Simulated Annealing Search algorithm into the new sample placement. Testing against multiple genera (*Alphacoronavirus, Alphaherpesvirinae*, and *Betacoronavirus*) revealed that the new software provided efficient performance and maintained stable accuracy.

For the phylogenetic analysis using recoding methods, plenty of effort has been made on bases and amino acids in recent years (Sridhar et al. 2007; Konishi et al. 2019; Phillips and Penny 2003; Hernandez, Ryan, and Uyeda 2021; Vera-Ruiz et al. 2014). However, it is well known that the estimation accuracy of recoding methods carries serious drawbacks (Vera-Ruiz et al. 2014). For example, the six-stats recoding in the protein phylogenetic analysis was ineffective in the face of high saturation (Hernandez, Ryan, and Uyeda 2021). Our study presents a creative way of using the recoding method called Frequency Vector Recoding, which combines nucleotide base frequency into the recoding methods and considers the influence of transition–transversion bias. In practice, it was shown that Method 1 revealed the highest similarity to the patristic distance calculated by the ML tree, with the Pearson correlation coefficient equal to 0.965 in *Betacoronavirus* (Table 1). Compared with the uncoded genetic distance, recoded distance provides higher discriminative power for sequences with close patristic distances (Supplementay Figure S7). Furthermore,

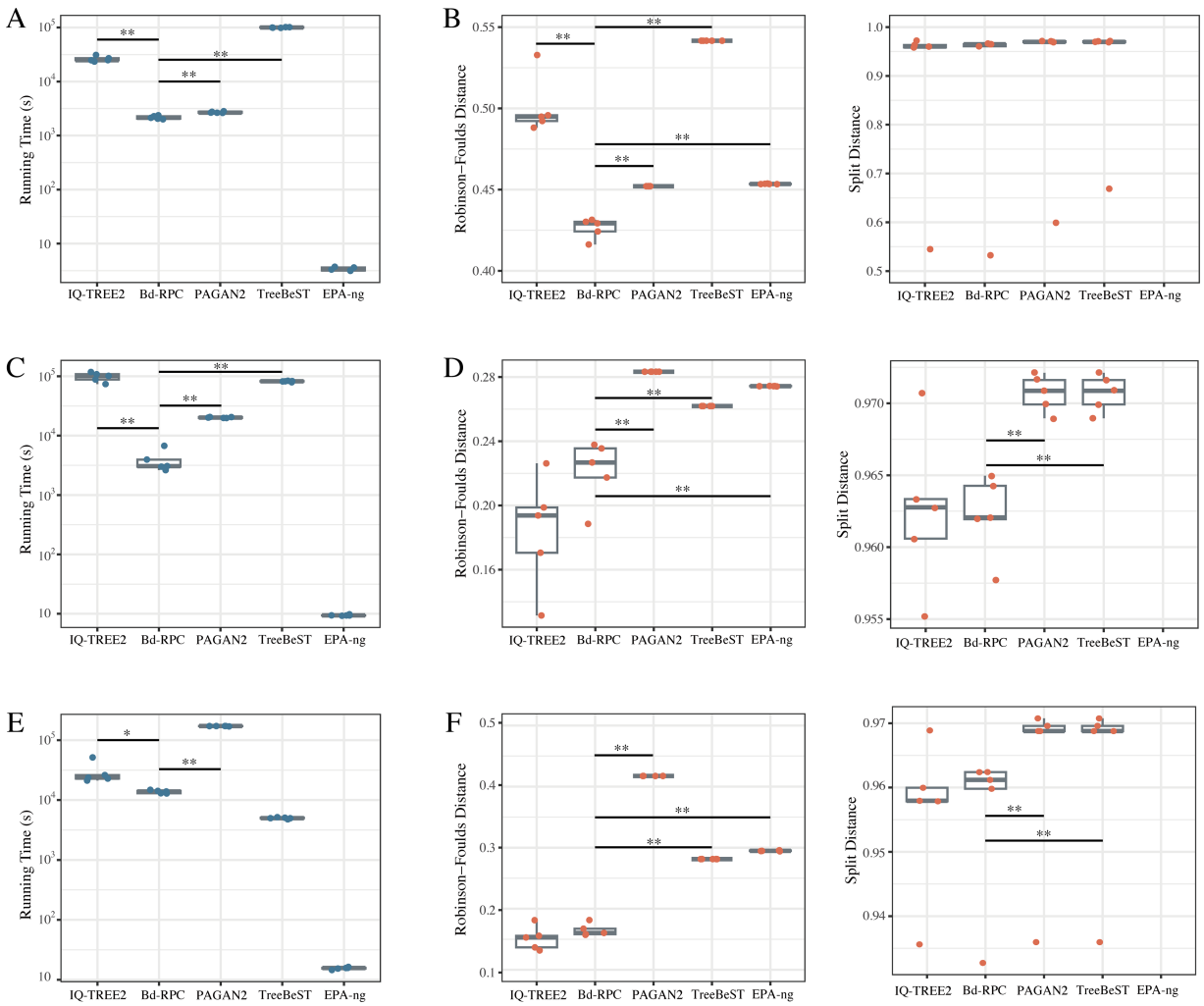

Figure 4. Running time and the distance to ML tree of each software in *Betacoronavirus, Alphacoronavirus, and Alphaherpesvirinae*. (A, C, E) The running time of each software in Betacoronavirus, Alphacoronavirus, and Alpha- herpesvirinae. The blue scatter represents the running time of each examination of the software. The running time of Bd-RPC was compared with that of other software using the one-sided paired Student's *t*-test. '*' represents *P*-value <0.05, and '**' represents *P*-value <0.01. (B, D, F) The Split and Robinson-Foulds distances were used to quantify the variation between the placement results and the ML tree. The orange scatter represents the results of each examination of the software. The statistical analysis among each result of the software was performed using the one-sided paired Student's *t*-test. '*' represents *P*-value <0.05, and '**' represents *P*-value <0.01. Here, the result of EPA-ng v.0.3.8 is not suitable for TOPD/FMTS to calculate Split Distance. (A, B) Statistics in Betacoronavirus. (C, D) Statistics in Alphacoronavirus. (E, F) Statistics in Alphaherpesvirinae.

the recoded distance metrics demonstrated superior performance over the uncoded genetic distance metrics in the *Alphacoronavirus* and *Betacoronavirus* datasets, as evidenced by the higher-quality clustering ratio and new sample placement ratio (Supplementary Figure S8). In practice, the distance calculation using the recoding methods was quite rapid, making it ideal for the initial location (not for estimating the final distances). Here, Bd-RPC provided a flexible framework allowing users to choose from six recoding methods used in this study or provided new recoding methods. In total, different from the previous research (Sridhar et al. 2007; Konishi et al. 2019; Phillips and Penny 2003; Hernandez, Ryan, and Uyeda 2021; Vera-Ruiz et al. 2014), the recoding methods were selected for approximating the phylogenetic distance rapidly, and the distance calculated by recoded sequences was used to estimate the analysis precision by matching the background information (taxonomy information or phylogenetic tree).

The Phylogenetic Simulated Annealing Search algorithm was employed to enhance the number of high-quality clusters in the database constructed using the phylogenetic tree. The Simulated Annealing algorithm is a probabilistic technique frequently employed in prior research to approximate the global optimum of a given function in Machine Learning (Zhan et al. 2016), traveling salesman problem (Zhang et al. 2020), (Aarts, Korst, and van Laarhoven 1988), and protein structure prediction (Rere, Fanany, and Arymurthy 2015; Chou and Carlacci 1991), but was never used in phylogenetic analysis. By accepting suboptimal solutions, this algorithm allows for a more thorough search for the global optimal solution (Kirkpatrick, Gelatt, and Vecchi 1983). Similarly, in this study, the Phylogenetic Simulated Annealing Search algorithm was developed to find the optimal global match of the list of sequence id between recoded sequence distance matrix and the phylogenetic tree of existing sequences. Through the evaluation, it can be found that the database utilizing the Phylogenetic Simulated Annealing Search algorithm exhibited a significant increase in the proportion of high-quality clusters from 87.5 per cent to 93.4 per cent (Supplementary Figure S4D). It can

be inferred that the Phylogenetic Simulated Annealing Search algorithm effectively reduced incorrect convergence and obtained more high-quality clusters.

The previous study showed that the indel variation models yielded biologically unrealistic estimations in constructing the phylogenetic tree (Saurabh et al. 2012). The indel was employed to recognize foreign sequences by counting the median fold-change of sequences' indel after adding new samples, but not estimate the phylogenetic tree in this study. As a result of the assessments, Bd-RPC can clearly distinguish foreign sequences through the Indel Recognition while maintaining high robustness with the default cutoff value (±10 per cent) (Supplementary Figure S5B and C). For the current method in foreign sequence recognition, PhyClip identifies outlier sequences using the patristic distance, and the sequences scaled away from the cluster's median patristic distance are recognized as foreign sequences (Han et al. 2019). Therefore, the phylogenetic tree construction is necessary for running PhyClip, and implanting PhyClip is unsuitable for Bd-RPC to place new sequences since the phylogenetic tree is output for Bd-RPC. In this study, Bd-RPC employs the Indel Recognition to distinguish the foreign sequences, and it need not construct the phylogenetic tree beforehand.

For the database created using taxonomy information, Bd-RPC classified the new samples into the cluster through the minimum distance, which can identify new samples at the species level. Here, the performance of Bd-RPC was evaluated on two datasets. It can be found that the Bd-RPC accurately identified unclassified sequences in each dataset, and the placement accuracy was improved from 99.3 per cent to 100 per cent as the sizes of the database grew (Fig. 3A and Supplementary Figure S6). Notably, the unclassified sequences (before 2020) such as SARS-CoV-2, BtRs-BetaCoV, and Pangolin coronavirus, which were classified into the SARS-CoV nowadays, have been correctly classified by Bd-RPC with the sequences obtained before 2015 (of 2020; Lam et al. 2020; Zhou et al. 2020).

To assess the performance of Bd-RPC in generating the phylogenetic tree, the running time and accuracy of the current state-of-the-art placement software (PAGAN2, EPA-ng, TreeBeST) were collected and compared with those of Bd-RPC in multiple genera (*Alphacoronavirus, Alphaherpesvirinae*, and *Betacoronavirus*) (Loytynoja, Vilella, and Goldman 2012; Barbera et al. 2019; Ruan et al. 2008; De Maio et al. 2023). Here, Bd-RPC maintained the highest precision with great efficiency and showed stable performance on all three virus genera (Fig. 4). It can be concluded that Bd-RPC provides a flexible framework that can be used in a variety of viruses, and its performance has been demonstrated in *Alphacoronavirus, Alphaherpesvirinae*, and *Betacoronavirus*. In addition, among these methods, only UShER and Bd-RPC of these phylogenetic placement tools built user-friendly websites (https://genome.ucsc.edu/cgi-bin/hgPhyloPlace and https://www.bd-rpc.xyz), which is convenient for researchers. Nowadays, the databases of *Alphacoronavirus* (ORF1ab), *Alphaherpesvirinae* (Concatenation of UL5, UL8, UL9, UL29, UL30, UL42, UL52), and *Betacoronavirus* (Spike) have been posted to the website for phylogenetic placement. In the future, the efficiency of Bd-RPC could be further improved by combining other tree construction software, such as fastTree and MAPLE, to replace IQ-TREE2 used in version 1.0 Bd-RPC for subtree construction (Minh et al. 2020; De Maio et al. 2023; Price, Dehal, and Arkin 2010). Meanwhile, more and more databases of other species would be updated on the website.

Bd-RPC provides a brand new way to place new samples, and a user-friendly website was built to offer convenient and real-time service for users. For the first time, this study integrates the recoding methods to save the computing time of tree construction and sheds new light on using indels in foreign sequence recognition. Furthermore, the Phylogenetic Simulated Annealing Search algorithm serve to find the optimal global match of the list of sequence id between recoded sequence distance matrix and the phylogenetic tree of existing sequences, which achieved good performance in the proportion of high-quality clusters. Bd-RPC provides a novel, automated, flexible, and efficient framework that can be generalized to place new samples and monitor pathogen dynamics.

Data availability

All data used in this work are available at https://github.com/Bin-Ma/bd-rpc/tree/master/example, collected from NCBI (http://ncbi.nlm.nih.gov) and GISAID (https://www.gisaid.org/). Bd-RPC online toolkit is available to users at http://www.bd-rpc.xyz. The source code and user manual are available at https://github.com/Bin-Ma/bd-rpc.

Supplementary data

Supplementary data is available at *VEVOLU Journal* online.

Acknowledgements

This work was supported by grants from the National Key Research and Development Program of China (2022YFD1801005), the National Natural Science Foundation of China (32272989), Natural Science Foundation of Hubei Province (2021CFA016), the earmarked fund for CARS-41, Hubei Province Natural Science Foundation for Distinguished Young Scholars (2020CFA060), and Applied Basic Research Project of Wuhan (Grant No. 2020020601012254).

Conflict of interest: The authors declare no competing interests.

References

Aarts, E. H. L., Korst, J. H. M., and van Laarhoven, P. J. M. (1988) 'A Quantitative Analysis of the Simulated Annealing Algorithm: A Case Study for the Traveling Salesman Problem', *Journal of Statistical Physics*, 50: 187–206.

Aberer, A. J., Pattengale, N. D., and Stamatakis, A. (2010) 'Parallel Computation of Phylogenetic Consensus Trees', *Procedia Computer Science*, 1: 1065–73.

Barbera, P. et al. (2019) 'EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences', *Systematic Biology*, 68: 365–9.

Bogdanowicz, D., and Giaro, K. (2011) 'Matching Split Distance for Unrooted Binary Phylogenetic Trees', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9: 150–60.

Chen, M. K., Lewis, L., and O, P. (2014) *Bayesian Phylogenetics: Methods, Algorithms, and Applications* (CRC Press).

Cheon, S., Zhang, J., and Park, C. (2020) 'Is Phylotranscriptomics as Reliable as Phylogenomics?', *Molecular Biology and Evolution*, 37: 3672–83.

Chou, K. C., and Carlacci, L. (1991) 'Simulated Annealing Approach to the Study of Protein Structures', "*Protein Engineering, Design and Selection*", 4: 661–7.

De Maio, N. et al. (2023) 'Maximum Likelihood Pandemic-scale Phylogenetics', *Nature Genetics*, 55: 746–52.

——— (1994) 'Estimating the Pattern of Nucleotide Substitution', *Journal of Molecular Evolution*, 39: 105–11.

Felsenstein, J. (1981) 'Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach', *Journal of Molecular Evolution*, 17: 368–76.

Han, A. X. et al. (2019) 'Phylogenetic Clustering by Linear Integer Programming (Phyclip)', *Molecular Biology and Evolution*, 36: 1580–95.

Hernandez, A. M., Ryan, J. F., and Uyeda, J. (2021) 'Six-State Amino Acid Recoding Is Not an Effective Strategy to Offset Compositional Heterogeneity and Saturation in Phylogenetic Analyses', *Systematic Biology*, 70: 1200–12.

Hotelling, H. (1936) 'Relations between Two Sets of Variates', *Biometrika*, 28: 321–77.

Huelsenbeck, J. P., and Ronquist, F. (2001) 'MRBAYES: Bayesian Inference of Phylogenetic Trees', *Bioinformatics*, 17: 754–5.

Kapli, P., Yang, Z., and Telford, M. J. (2020) 'Phylogenetic Tree Building in the Genomic Age', *Nature Reviews Genetics*, 21: 428–44.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Kirkpatrick, S., Gelatt, C. D., Jr, and Vecchi, M. P. (1983) 'Optimization by Simulated Annealing', *Science*, 220: 671–80.

Kobert, K. et al. (2014) 'The Divisible Load Balance Problem and Its Application to Phylogenetic Inference', *Lecture Notes in Computer Science*, 8701: 204–16.

Kobert, K., Stamatakis, A., and Flouri, T. (2017) 'Efficient Detection of Repeating Sites to Accelerate Phylogenetic Likelihood Calculations', *Systematic Biology*, 66: 205–17.

Konishi, T. et al. (2019) 'Principal Component Analysis Applied Directly to Sequence Matrix', *Scientific Reports*, 9: 19297.

Lam, T. T. et al. (2020) 'Identifying SARS-CoV-2-related Coronaviruses in Malayan Pangolins', *Nature*, 583: 282–5.

Loytynoja, A., Vilella, A. J., and Goldman, N. (2012) 'Accurate Extension of Multiple Sequence Alignments Using a Phylogeny-aware Graph Algorithm', *Bioinformatics*, 28: 1684–91.

Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–4.

Coronaviridae Study Group of the International Committee on Taxonomy, of, V. (2020) 'The Species Severe Acute Respiratory Syndrome-related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2', *Nature Microbiology*, 5: 536–44.

Olm, M. R. et al. (2017) 'dRep: A Tool for Fast and Accurate Genomic Comparisons that Enables Improved Genome Recovery from Metagenomes through De-replication', *ISME Journal*, 11: 2864–8.

Paradis, E., Claude, J., and Strimmer, K. A. P. E. (2004) 'Analyses of Phylogenetics and Evolution in R Language', *Bioinformatics*, 20: 289–90.

Phillips, M. J., and Penny, D. (2003) 'The Root of the Mammalian Tree Inferred from Whole Mitochondrial Genomes', *Molecular Phylogenetics & Evolution*, 28: 171–85.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) 'FastTree 2—Approximately Maximum-likelihood Trees for Large Alignments', *PLoS One*, 5: e9490.

Puigbo, P., Garcia-Vallve, S., and McInerney, J. O. (2007) 'TOPD/FMTS: A New Software to Compare Phylogenetic Trees', *Bioinformatics*, 23: 1556–8.

Rambaut, A., and Grassly, N. C. (1997) 'Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution along Phylogenetic Trees', *Computer Applications in the Biosciences : CABIOS*, 13: 235–8.

Rere, L. M. R., Fanany, M. I., and Arymurthy, A. M. (2015) 'Simulated Annealing Algorithm for Deep Learning', *Procedia Computer Science*, 72: 137–44.

Robinson, D. F., and Foulds, L. R. (1981) 'Comparison of Phylogenetic Trees', *Mathematical Biosciences*, 53: 131–47.

Ruan, J. et al. (2008) 'TreeFam: 2008 Update', *Nucleic Acids Research*, 36: D735–740.

Saurabh, K. et al. (2012) 'Gaps: An Elusive Source of Phylogenetic Information', *Systematic Biology*, 61: 1075–82.

Smith, M. R., and Schwartz, R. (2020) 'Information Theoretic Generalized Robinson-Foulds Metrics for Comparing Phylogenetic Trees', *Bioinformatics*, 36: 5007–13.

Sridhar, S. et al. (2007) 'Algorithms for Efficient Near-perfect Phylogenetic Tree Reconstruction in Theory and Practice', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4: 561–71.

Tamura, K., Stecher, G., and Kumar, S. (2021) 'MEGA11: Molecular Evolutionary Genetics Analysis Version 11', *Molecular Biology and Evolution*, 38: 3022–7.

Turakhia, Y. et al. (2021) 'Ultrafast Sample Placement on Existing tRees (Usher) Enables Real-time Phylogenetics for the SARS-CoV-2 Pandemic', *Nature Genetics*, 53: 809–16.

Vakirlis, N. et al. (2016) 'Reconstruction of Ancestral Chromosome Architecture and Gene Repertoire Reveals Principles of Genome Evolution in a Model Yeast Genus', *Genome Research*, 26: 918–32.

Vera-Ruiz, V. A. et al. (2014) 'Statistical Tests to Identify Appropriate Types of Nucleotide Sequence Recoding in Molecular Phylogenetics', *BMC Bioinformatics*, 15: 1–11.

Yang, Z. (1993) 'Maximum-likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites', *Molecular Biology and Evolution*, 10: 1396–401.

Yang, Z. (2006) *Molecular Evolution: A Statistical Approach* (Oxford University Press).

Yang, Z., and Rannala, B. (2012) 'Molecular Phylogenetics: Principles and Practice', *Nature Reviews Genetics*, 13: 303–14.

Zhan, S. H. et al. (2016) 'List-Based Simulated Annealing Algorithm for Traveling Salesman Problem', *Computational Intelligence and Neuroscience*, 2016: 1–12.

Zhang, L. et al. (2020) 'Protein Structure Optimization Using Improved Simulated Annealing Algorithm on a Three-dimensional AB Off-lattice Model', *Computational Biology and Chemistry*, 85: 107237.

Zhou, H. et al. (2020) 'A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein', *Current Biology*, 30: 2196–2203 e2193.

Virus Evolution, 2024, **10(1)**, 1–10
DOI: https://doi.org/10.1093/ve/veae005
Advance Access Publication 27 January 2024
Resources