Virus Database and Online Inquiry System Based on Natural Vectors

Evolutionary Bioinformatics
Volume 13: 1–7

DOI: 10.1177/1176934317746667

Rui Dong[1], Hui Zheng[2], Kun Tian[1], Shek-Chung Yau[3], Weiguang Mao[1], Wenping Yu[4], Changchuan Yin[2], Chenglong Yu[5,6], Rong Lucy He[7], Jie Yang[2] and Stephen ST Yau[1]

[1]Department of Mathematical Sciences, Tsinghua University, Beijing, China. [2]Department of Mathematics, Statistics, and Computer Science, The University of Illinois at Chicago, Chicago, IL, USA. [3]Information Technology Services Center, The Hong Kong University of Science and Technology, Kowloon, Hong Kong. [4]College of Computer and Control Engineering, Nankai University, Tianjin, China. [5]Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA, Australia. [6]School of Medicine, Flinders University, Adelaide, SA, Australia. [7]Department of Biological Sciences, Chicago State University, Chicago, IL, USA.

ABSTRACT: We construct a virus database called VirusDB (http://yaulab.math.tsinghua.edu.cn/VirusDB/) and an online inquiry system to serve people who are interested in viral classification and prediction. The database stores all viral genomes, their corresponding natural vectors, and the classification information of the single/multiple-segmented viral reference sequences downloaded from National Center for Biotechnology Information. The online inquiry system serves the purpose of computing natural vectors and their distances based on submitted genomes, providing an online interface for accessing and using the database for viral classification and prediction, and back-end processes for automatic and manual updating of database content to synchronize with GenBank. Submitted genomes data in FASTA format will be carried out and the prediction results with 5 closest neighbors and their classifications will be returned by email. Considering the one-to-one correspondence between sequence and natural vector, time efficiency, and high accuracy, natural vector is a significant advance compared with alignment methods, which makes VirusDB a useful database in further research.

KEYWORDS: virus, database, natural vector, genome sequences, classification

RECEIVED: July 30, 2017. **ACCEPTED:** October 5, 2017.

TYPE: Original Research

FUNDING: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Tsinghua University start-up fund.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Stephen ST Yau, Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China. Email: yau@uic.edu

Rong Lucy He, Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA. Email: rhe@csu.edu

Jie Yang, Department of Mathematics, Statistics, and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607, USA. Email: jyang06@uic.edu

Introduction

Viruses are tiny organisms that may cause a number of diseases in eukaryotes. They are quite diverse and different from other biological entities. Some viruses have RNA genomes and some have DNA genomes. Some viruses have single-stranded genomes, whereas others have multiple-stranded genomes. The structures and replication strategies are also quite different.[1,2] Due to the high diversity of viruses, the corresponding classification problem is complicated. The International Committee on Taxonomy of Viruses (ICTV) defined a universal taxonomic scheme for all the viruses. According to the ICTV scheme, viral classification starts at the highest level of *order* and continues as follows: *order, family, subfamily, genus*, and *species*. Another main scheme used for the classification of viruses is the Baltimore classification system.[3] Based on their nucleic acid, strandedness (single-stranded or double-stranded), and methods of replication, viruses are divided into the following 7 groups: dsDNA, ssDNA, dsRNA, ssRNA-RT, (+)ssRNA, (−) ssRNA, and dsDNA-RT.

More and more new deadly viruses have been discovered, such as human immunodeficiency virus, severe acute respiratory syndrome, Ebola virus, West Nile virus, H7N9, and their variants. They continually threaten human's life. Nowadays, scientists are able to study the nucleotide sequences of viruses and develop algorithms to classify and predict the characteristics of new viruses. One major approach is to analyze the similarity between the new viruses and known ones. Multiple sequence alignment (MSA) has been widely accepted and used by biologists and virologists now. However, its heavy computational cost causes serious inefficiency, which makes it impossible to analyze the phylogeny of whole genomes. Besides, MSA methods may fail in diverse systems of different families of RNA viruses. Meantime, different evolutionary models can lead to different results, which implies that the models and assumptions are unnatural and unpersuasive. Another popular category of alignment-free methods is based on the statistics of oligomers frequency and associated with a fixed-length segment, known as *k*-mers. However, it ignores the positional information of nucleotides.

Many virus databases have been constructed but on a limited subset. Examples include the influenza virus database,[4] the RNA virus database,[5] the LANL hemorrhagic fever virus database,[6] and the Hepatitis B virus database.[7] The GenBank provides a

comprehensive database for viral genomes but with significantly high-missing ICTV labels, partly due to the different definitions of "species". For example, the missing rates of family and genus labels at GenBank are as high as 12% and 30%, respectively.[8] For virus research, it is crucial to take the whole virus database as an entirety to reveal the relationships for classification.

Our VirusDB introduced here is based on a newly developed method, natural vector representation,[8,9–11] which is highly accurate and efficient. In our latest version (April 14, 2017), we extract all the viruses with complete genomes information directly downloaded from National Center for Biotechnology Information (NCBI). All the genomes are obtained by biological experiments rather than the prediction from other models, which reduces the errors by human involvement to the greatest extent. With the rapid development of sequencing technology such as next-generation sequencing, the sequence resources would be more and more abundant and accurate, thus the prediction results would be more satisfying as well.

Materials and Methods

There are 2 kinds of methods for alignment-free genetic sequence comparison in the literature.[12] One is based on the statistics of oligomer frequencies and associated with a fixed-length segments, and the other one is based on information theory and chaos theory. In the past 10 years, alignment-free comparison has attracted much attention from researchers, and more and more methods and models have been proposed. The natural vector method is an alignment-free approach to characterize the genetic sequences and could be used to classify and to predict the DNA/RNA and protein sequences with high efficiency and accuracy.[8,10–14]

Let $S = (s_1, s_2, ..., s_n)$ be a nucleotide sequence with length n, $s_i \in \{A, C, G, T\}$. For $\alpha = A, C, G, T$, we define $w_\alpha(s) = 1$, if $s = \alpha$; and $w_\alpha(s) = 0$ otherwise.

The 12-dimemsional natural vector of S is defined as $(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_A^2, D_C^2, D_G^2, D_T^2)$,[8,10] where $n_\alpha = \sum_{i=1}^{n} w_\alpha(s_i)$ denotes the number of occurrences of letter α in S, $\mu_\alpha = \sum_{i=1}^{n} i \cdot w_\alpha(s_i) / n_\alpha$ is the mean position of letter α, $D_\alpha^2 = \sum_{i=1}^{n} (i - \mu_\alpha)^2 w_\alpha(s_i) / n_\alpha n$ is a scaled variance of positions of letter α.

For example, the nucleotide sequence "ACGTA" will be represented as "(2, 1, 1, 1, 3, 2, 3, 4, 0.8, 0, 0, 0)."

Using the natural vector representation, if a viral genome consists of a single-nucleotide sequence, known as single-segmented, then the virus will be represented by a 12-dimensional numerical vector in the database. If a viral genome contains more than one nucleotide sequence, known as multiple-segmented, then each nucleotide sequence is represented by a natural vector and this virus will be represented by multiple 12-dimensional vectors in the database. According to the distance matrix, we find the nearest neighbor of each virus, i.e., the virus that has the smallest distance to it. We check the Baltimore, family, genus, and species labels of the virus and count it as a case of misclassification if they are not consistent.

We first build a distance matrix among all viruses in the database. If 2 viruses are both single-segmented, then the distance between them is defined as the Euclidean distance between their natural vectors of the 2 viruses. If at least 1 of the 2 viruses is multiple-segmented, the distance is defined as the Hausdorff distance of 2 sets of natural vectors.[10]

For example, if set $A = \{a_1, \ldots, a_p\}$ and set $B = \{b_1, \ldots, b_q\}$, the Hausdorff distance between them is defined as follows:

$$H(A,B) = \max(h(A,B), h(B,A)), \text{ where } h(A,B) = \max_{a \in A} \min_{b \in B} \| a - b \|$$

The natural vector method has quite a few successful applications recently. In 2011, we applied natural vector to cluster H1N1 genomes and reported as significant results in dendrogram, which coincided with biologists' analyses in the work by Deng et al.[9] We also predict that the A(H1N1) genomes are originally from swine flu virus genome lineage, which shows the direction toward how to resist the threat of the new influenza; in 2013, we also proposed 12-dimensional natural vectors for classifying all single-segmented viruses (DNA and RNA genomes) which broadened the range of natural vector.[8] Among the 7 Baltimore classes, the error rates of classifying Baltimore labels were below 0.01% for Baltimore I, II, IV, V, and VII, and the error rates of classifying family labels were 0 for Baltimore II, III, V, VI, and VII. After validating with the published references, we successfully predicted 21 missing labels of viruses. In the work by Huang et al,[10] we extended the natural vector approach to include multiple-segmented viruses by introducing Hausdorff metric in the GenBank at that time. The error rates of the predictions of 2384 viruses were 3.5% for Baltimore labels, 4.6% for family labels, 0.3% for subfamily labels, and 4.4% for genus labels. We also analyzed the influenza A(H7N9) virus whose genome consists of 8 segments and drew the conclusion that the analysis based on whole genomes through Hausdorff distance is more reliable than the classical one based only on 2 segments, which proves that our method performs well in multiple-segmented viruses. In 2015, we applied natural vectors on Ebola viruses of the 2014 outbreak.[11] The accuracy rates on classifying family and genus labels were 100%. Our phylogenetic analysis showed that a protein named VP24 is the most consistent one to the variation of virulence among the 7 proteins related to Ebola viruses, which suggests that VP24 would be a pharmaceutical target for preventing and treating Ebola virus. As natural vector can reflect core information stored in sequences and genomes, we use it to construct the virus classification system introduced in this article.

Results

Database

Our VirusDB contains all the single-segmented and multiple-segmented reference viral genomes in GenBank. The reference nucleotide sequence and the label information of Baltimore

class, family, subfamily, genus, and species of each virus are included. Every viral nucleotide sequence is analyzed and converted into a unique 12-dimensional natural vector. Because the correspondence between the natural vectors and the existing viral genomes is one to one, the distance between 2 viral genomes based on natural vectors successfully reflects their biological distance.[8,10] Relative distances ensure the influence of different scopes of families or species, whereas absolute distances show little significance to the final result as the length and the scope of categories vary from one to another.

As natural vector method is nonassumption and does not rely on any model, it can be applied well on the whole virus dataset in GenBank. Thus, it can help fill in the high-missing classification labels in GenBank fast and accurately. Meantime, it provides insight into the sequences or genomes of viruses and their variants and shows the direction for biological experiments and medical research, especially for viral diseases.

Currently, we have 5 database versions available which are labeled by the downloading time from GenBank: v1, July 27, 2012; v2, November 21, 2012; v3, May 22, 2013; v4, February 20, 2014; and v5, April 14, 2017. The first 2 versions only contain single-segmented viruses. We keep these 5 versions available to users to track the changes of viral information due to corrections, updates, etc. The last 3 versions include both single-segmented and multiple-segmented viruses. The latest version contains 4390 viruses with complete genomes information directly downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/). Those genomes are obtained from biological experiments rather than predicted by other models, which may introduce many assumptions and errors from the models. We have tested the fifth version to make sure that it is stable with high prediction accuracy.

Our database stores the following information of each virus:

- Number of segments: 1 if the virus is single segmented, and some integer larger than 1 if the virus is multiple segmented;
- Accession: the access number of the virus in GenBank;
- GI: the GenBank GI number of the viral nucleotide sequence;
- GenPart1: first class label of the nucleotide sequence extracted from NCBI;
- GenPart2: if the sequence is from ssRNA, the property of positive/negative-strand information;
- Balt: Baltimore classification labels I, II, III, IV, V, VI, VII of the virus;
- Shape: the shape of the virus, circular or linear;
- Order: order label of the virus;
- Family: family label of the virus;
- Subfamily: subfamily label of the virus;
- Genus: genus label of the virus;
- Virus: the name of the virus;
- Length: the length of the nucleotide sequence;
- 12-dimensional natural vectors.

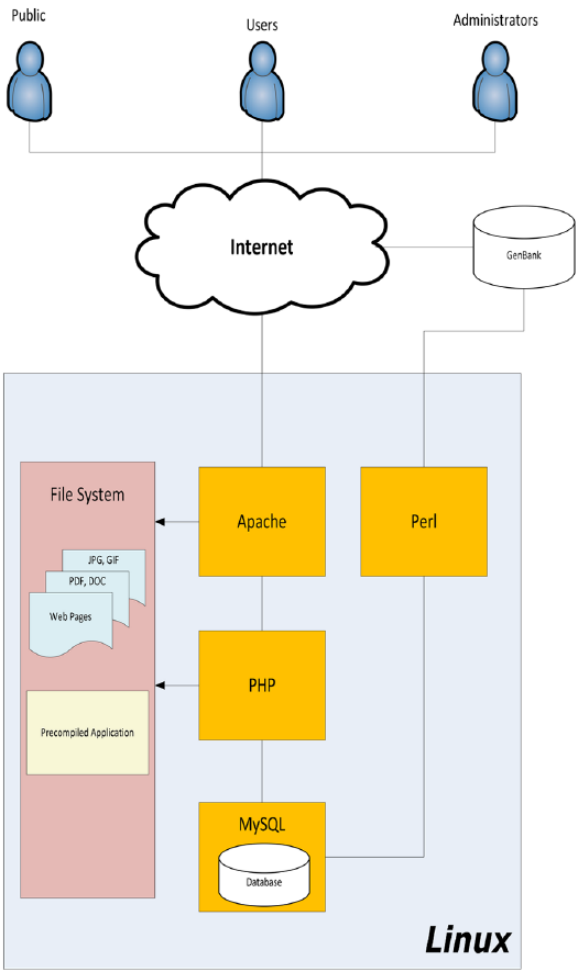

Figure 1. The high-level design of the inquiry system.

Online inquiry system

The online inquiry system at VirusDB facilitates the broad scientific user community with the natural vector method, including virologists, computational biologists, genomics and proteomics scientists, molecular biologists, and clinicians. To cater the balance between development and cost as well as system maintenance and performance, a Linux-Apache-MySQL-PHP (LAMP) solution is applied. Figure 1 outlines the high-level design of the inquiry system.

A LINUX-based host running on CentOS 6 is used as a server that contains all the required program and application modules for implementing the system. It also serves as a file repository for the public and administrative document storage system.

The core of the whole system is a database management system, implemented using MySQL databases. The core database is used to store the genome sequences, their associated natural vectors, and classification information. Perl programming language is adopted as the scripting language to bridge the internal database and online GenBank database. Apache http server, the most widely adopted web server, is used to serve both general public and internal administrators. WordPress has been adopted as the presentation layer. PHP is used as the scripting language for the Web application development.

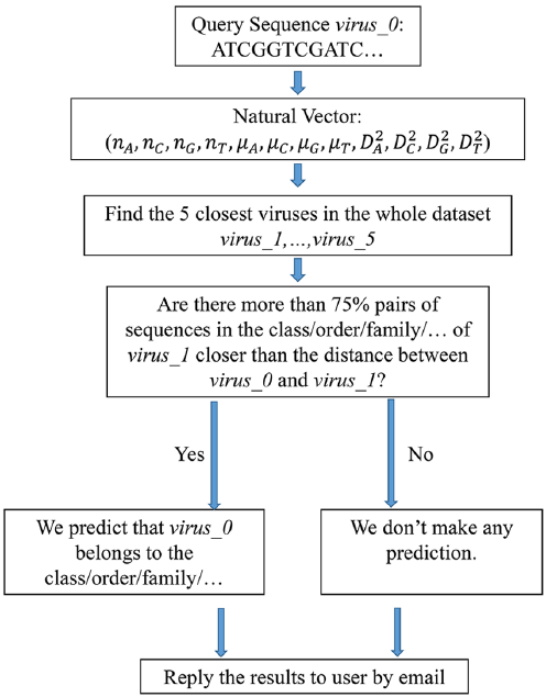

Figure 2. The basic workflow of VirusDB.

Web interface design overview

The structure of Web interface of whole system consists of 8 parts: Home, About, Content, Submit Sequence, Get Result, Current Collection, FAQ, and Contact Us.

The whole Web system contains 2 components, namely, as follows:

- Virus Classification Project using natural vector method;
- Virus Prediction.

Virus Classification Project is the main homepage for the project. All 8 categories except the Submit Sequence and Get Result are under this homepage. The Home section displays our research progress on natural vector method, published papers, and funding support. The About section contains the overview of the whole project. The Content section contains detailed information of the whole project. The Submit Sequences web application aims at providing online interface to users. The Current Collection provides online interfaces for the public to read the current virus classification. FAQ displays several commonly asked questions and the corresponding answers, especially about the inputs and outputs of the system, and will be updated regularly according to feedback from users.

Virus Prediction System includes Submit Sequences and Get Result. Our basic workflow is shown in Figure 2. The users are inquired to choose either single segment or multiple segments of the viruses. With a little instruction beginning with a symbol ">" on the first line (which is similar to FASTA file), they can submit a nucleotide sequence with or without the Baltimore and family labels. If no label is chosen, the calculation would be done in the whole database and the prediction of which Baltimore class and family it belongs to would be made. If the sequence and Baltimore class label are submitted, but without a family label, the calculation will be done within the Baltimore class, and then the prediction of whether the query virus is in the Baltimore class or not would be made. If the family label is also submitted, then the system would predict whether it is in the family or not. As soon as the system receives a request with the verification code to ensure human requiring rather than attack from a malicious hacker, it will calculate the corresponding natural vector as well as the distances between the submitted sequence and the viruses in VirusDB. Besides, when we make a prediction, we set up a cutoff rate of 75%, which means that only when the distance between the reference virus and its nearest neighbor is smaller than 75% of the distances in the family, we will make the prediction that this virus belongs to the same class as its nearest neighbor. In this way, the accuracy rate is dramatically improved up to 90% for the predicted ones.[8]

Figure 3 displays the screenshots of the procedure of submission and prediction of a single-segmented virus. It shows that the system successfully predicted a Parvoviridae virus. For a multiple-segmented virus, users can choose "Multiple Segments" and upload the corresponding FASTA files to get the results.

We design Get Result section to allow users to retrieve their previous requests. That is also why we send the results back to users via email so that when they want to retrieve the results, they can easily get the Request ID from their previous email. They can get results by submitting the Request ID which is generated by the system.

The Current Collection allows users to find the viruses from a given family group. Figure 4 shows the screenshots of finding all the Filoviridae viruses in our database through Current Collection. By clicking the neighbors and segments buttons, one can receive more information about the virus. We also add the links to GenBank in the current version, which contain the link of virus itself and of each segment.

Discussion

We plan to build the System Administration section to serve the administrative tasks for the system in the future. It will include Approve Submission, List Submission, Upload New Database, and Manual Database Update. Once a user submitted a sequence, a notification will be sent to the moderator. Should the administrator accept the submission, further processing, as well as calculation will be done.

Moreover, we have some back-end process. Computation Service is designed because sequence processing using natural vectors involves a huge volume of computer processing. It provides program modules written in Perl (or further enhanced using C/C++) to compute natural vectors, build distance matrices, and find neighbors of a virus. Automatic Database Updating refers to retrieving latest sequences from GenBank and recalculating the natural vectors and its associated distances with other

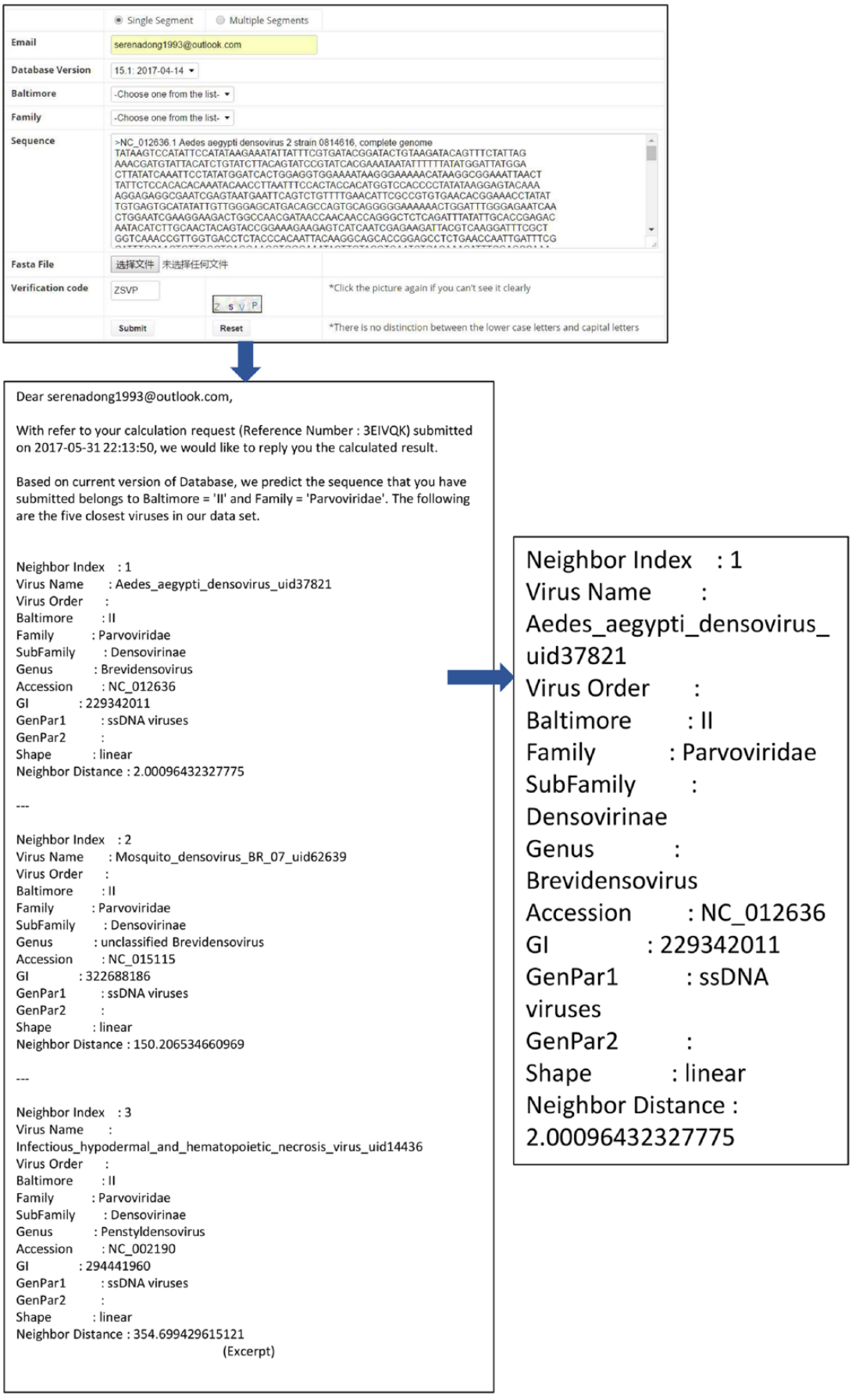

Figure 3. Screenshots of procedure of submission and prediction of a single-segmented virus.

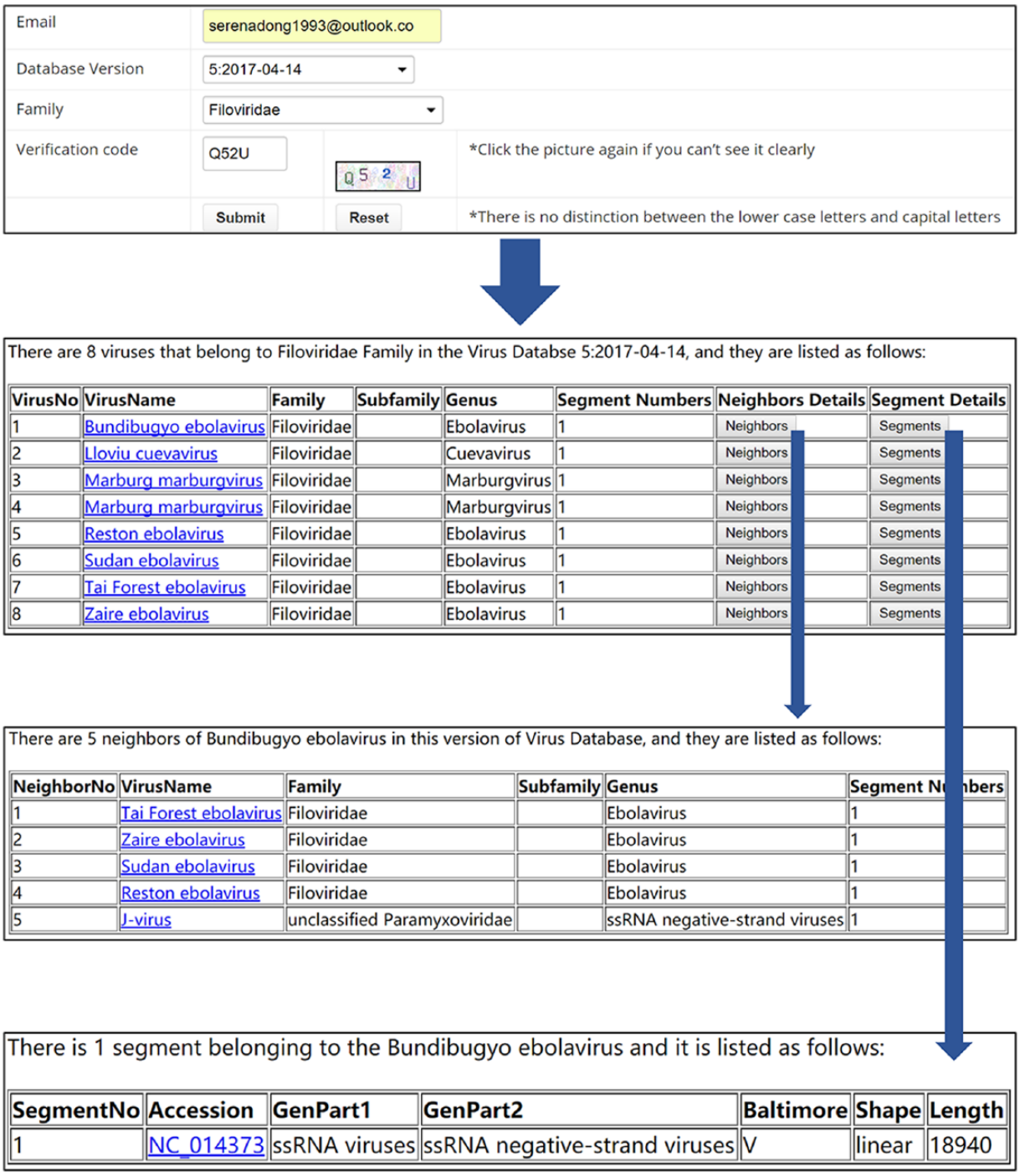

Figure 4. The Filoviridae viruses in our database through Current Collection.

viruses. The following system administration and maintenance tasks will run on a predetermined schedule: system backup, database housekeeping, usage report generation, routine integrity check, and performance reporting and tuning.

We plan to build the Approve Submission interface to allow moderator to view the submission and approve/reject submission possibly with reasons. The Upload New DB Dataset interface will allow project leader to operate the system to download a new version of the dataset. The new dataset would then be verified against GenBank and the FASTA files would then be retrieved. Natural vectors and distance matrices would then be recalculated and updated. This database is planned to update every year to ensure its advance with times and new technology.

Viral classification may provide clues for understanding the origins and mechanisms of transmission of viruses, and thus we are also working on a graphical representation of our prediction results. For better user's experience, a complete and detailed analysis of the query virus is essential. A new phylogenetic graphical representation has been proposed in 2013[8]; we plan

to implement it to provide the users with visual results as well. Traditional phylogenetic trees are also in our plan for a more direct and clear representation.

Currently, VirusDB can calculate the genome sequences of a virus. However, proteins have more direct functions during the procedure of virus infection and transmission. The protein sequences are also an important data source to understand the functions and relationships of viruses, and previous works have made some progress.[15] In 2016, we constructed a 60-dimensional protein space to analyze the evolutionary relationships of 4021 viruses by whole proteomes in the NCBI Reference Sequence Database.[13] The accuracy for randomly chosen 351 viruses data set can reach 95.4%, whereas the k-mer can only get 71.2% accuracy ($k=6$ as the optimal). This inspires us to add the proteome information into our system, thus in subsequent version, users can input the proteome and get a prediction result based on genomes and proteomes. Predictions based on the protein sequences would be more related to the practical impact of viruses and would put more insight into the functions and mechanisms of viruses.

Conclusions

This database VirusDB and online inquiry system serve the purpose of storing viral genome reference sequences and related information, providing online tools for computing natural vectors of genome sequences submitted by the users and returning 5 nearest known viral genomes. Based on the high efficiency and accuracy of natural vector method, it serves as a valuable tool for virus analysis.

Acknowledgements

SSTY is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Author Contributions

SSTY conceived and designed the experiments. RD, HZ, S-CY, WM, and WY analyzed the data.

RD and HZ wrote the first draft of the manuscript. All the authors contributed to the writing of the manuscript, agree with manuscript results and conclusions, jointly developed the structure and arguments for the paper, made critical revisions and approved final version, and reviewed and approved the final manuscript.

REFERENCES

1. Emiliani C. Extinction and viruses. *Biosystems*. 1993;31:155–159.
2. Wessner DR. The origin of viruses. *Nat Educ*. 2010;3:37.
3. Baltimore D. Expression of animal virus genomes. *Bacteriol Rev*. 1971;35:235–241.
4. Chang S, Zhang J, Liao X, et al. Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research. *Nucleic Acids Res*. 2006;35:376–380.
5. Belshaw R, Oliverira T, Markowitz S, Bambaut A. The RNA virus database. *Nucleic Acids Res*. 2008;37:431–435.
6. Kuiken C, Thurmond J, Dimitrijevic M, Yoon H. The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. *Nucleic Acids Res*. 2011;40:587–592.
7. Hayer J, Jadeau F, Deleage G, Kay A, Zoulim F, Combet C. HBVdb: a knowledge database for Hepatitis B Virus. *Nucleic Acids Res*. 2012;41:566–570.
8. Yu C, Hernandez T, Zheng H, et al. Real time classification of viruses in 12 dimensions. *PLoS ONE*. 2013;8:e64328.
9. Deng M, Yu C, Liang Q, He RL, Yau SST. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS ONE*. 2011;6:e17293.
10. Huang HH, Yu C, Zheng H, et al. Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Molec Phylogenet Evol*. 2014;81:29–36.
11. Zheng H, Yin C, Hoang T, He RL, Yang J, Yau SS. Ebolavirus classification based on natural vectors. *DNA Cell Biol*. 2015;34:418–428.
12. Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics*. 2003;19:513–523.
13. Li Y, Tian K, Yin C, He RL, Yau SS. Virus classification in 60-dimensional protein space. *Mol Phylogenet Evol*. 2016;99:53–62.
14. Li Y, He L, He RL, Yau SS. Zika and Flaviviruses phylogeny based on the alignment-free natural vector method. *DNA Cell Biol*. 2017;36:109–116.
15. Yu C, Deng M, Cheng SY, Yau SC, He RL, Yau SS. Protein space: a natural method for realizing the nature of protein universe. *J Theor Biol*. 2012;318:197–204.