RESEARCH ARTICLE

A Computational Model for Predicting RNase H Domain of Retrovirus

Sijia Wu, Xinman Zhang*, Jiuqiang Han*

School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China

* zhangxinman@mail.xjtu.edu.cn (XZ); jqhan@mail.xjtu.edu.cn (JH)

Abstract

RNase H (RNH) is a pivotal domain in retrovirus to cleave the DNA-RNA hybrid for continuing retroviral replication. The crucial role indicates that RNH is a promising drug target for therapeutic intervention. However, annotated RNHs in UniProtKB database have still been insufficient for a good understanding of their statistical characteristics so far. In this work, a computational RNH model was proposed to annotate new putative RNHs (np-RNHs) in the retroviruses. It basically predicts RNH domains through recognizing their start and end sites separately with SVM method. The classification accuracy rates are 100%, 99.01% and 97.52% respectively corresponding to jack-knife, 10-fold cross-validation and 5-fold cross-validation test. Subsequently, this model discovered 14,033 np-RNHs after scanning sequences without RNH annotations. All these predicted np-RNHs and annotated RNHs were employed to analyze the length, hydrophobicity and evolutionary relationship of RNH domains. They are all related to retroviral genera, which validates the classification of retroviruses to a certain degree. In the end, a software tool was designed for the application of our prediction model. The software together with datasets involved in this paper can be available for free download at https://sourceforge.net/projects/rhtool/files/?source=navbar.

Citation: Wu S, Zhang X, Han J (2016) A Computational Model for Predicting RNase H Domain of Retrovirus. PLoS ONE 11(8): e0161913. doi:10.1371/journal.pone.0161913

Editor: Jean-Luc EPH Darlix, "INSERM", FRANCE

Received: April 28, 2016

Accepted: August 15, 2016

Published: August 30, 2016

Data Availability Statement: All relevant data are within the paper.

Funding: This work was supported by grants from the Ph.D. Program Foundation of the Ministry of Education of China (No. 20110201110010). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

1 Introduction

The retroviruses encompass a family of enveloped RNA-containing viruses that utilize reverse transcription of their genomes as an obligate step in virus replication [1]. Based on differences in morphological and biochemical features, retroviruses can be classified into seven genera including alpha-retrovirus, beta-retrovirus, gamma-retrovirus, delta-retrovirus, epsilon-retrovirus, lentivirus and spumavirus [2]. The common replication mode of all these retroviruses leads to reverse flow of genetic information from RNA template to intermediate DNA-RNA hybrid and then to complementary DNA [3]. The procedure thus contributes to virus propagation and influences genetic composition of infected cells [4]. To complete this process, reverse transcriptase enzyme is absolutely required which is encoded by polymerase (*pol*) gene [5]. As a key domain in reverse transcriptase, RNase H (RNH) specifically degrades the RNA strand of DNA-RNA replication intermediate to free the newly-made minus strand for use as a template in the synthesis of the plus strand of DNA [6–9].

The indispensable role of RNH in retroviral replication attracts the attention of some researchers. Their works reveal that inhibiting activity of RNH to break reverse transcription can be helpful to therapeutically intervene in some kinds of diseases, such as neoplasia, autoimmunity and immunosuppression [8–12]. Thus, RNH is qualified as an important and promising pharmaceutical target. But annotated RNHs in online database are not sufficient for statistical research. More new putative RNHs (np-RNHs) in the retroviruses are waiting for prediction with bioinformatics methods. However, the special RNH prediction software has been absent so far, and some classical database search tools [13–15] achieved unsatisfied accuracy rates which were lower than 80% in RNH recognition. Therefore, it is of great importance to come up with a computational RNH prediction model. This model will be conducive to reducing the number of amino acid sequences for biochemical experiment corroboration.

On the basis of amino acid sequences only, the proposed model recognizes start and end sites by SVM method to accomplish the aim of RNH prediction. The remainder of this paper is arranged as follows containing all the requirements of a sequence-based predictor [16, 17]. In Section 2, we will briefly describe the datasets used in this study. In Section 3, we will introduce the general scheme of our model and proper validation methods. Then the experimental results will be provided in Section 4, including the optimal parameters, validity analysis, RNH motif and evolutionary relationship. Finally, we will summarize our work and present the conclusions in Section 5.

2 Datasets

All the retroviral protein sequences involved in this work were collected from UniProt Knowledgebase (UniProtKB) [18, 19] and divided into two datasets. One named benchmark dataset contains 105 *pol* sequences with RNH annotations so as to complete the establishment of RNH prediction model. These data are all qualified to meet the criteria which stress RNH domains to be non-repetitive, manually annotated or reviewed and consistent in different databases. The other one includes 149,692 *pol* sequences and 320 non-*pol* sequences both without RNH annotations for predicting potential np-RNHs.

3 Methods

3.1 Prediction algorithm

Stage 1: Sample preparation. There are two sets of samples corresponding to start and end sites prediction respectively. The start samples are expressed as:

$$\begin{aligned} ss &= pol(i : i + L1 - 1)\ i = start(\text{RNH}) + d1\ d1 = [-20, 20] \\ &= \begin{cases} \text{positive start sample if } d1 = 0 \\ \\ \text{negative start sample if } d1 \neq 0 \end{cases} \end{aligned}, \tag{1}$$

while the end samples are given by:

$$\begin{aligned} es &= pol(i - L2 + 1 : i)\ i = end(\text{RNH}) + d2\ d2 = [-20, 20] \\ &= \begin{cases} \text{positive end sample if } d2 = 0 \\ \\ \text{negative end sample if } d2 \neq 0 \end{cases} \end{aligned}. \tag{2}$$

Here, $L1$ and $L2$ represent the length of start and end samples severally. *start*(RNH) and *end* (RNH) denote either true start and end site for a training *pol* sequence, or achieved start and end position after Smith-Waterman alignment [20] with training RNH domains for an inquired sequence.

Stage 2: Feature extraction. The widely recognized position weight matrix (PWM) [21–23] was employed to represent the conservative level of amino acid sequences in the benchmark dataset. By aligning positive start samples, negative start samples, positive end samples and negative end samples, PWMs are defined as:

$$PWM_{\alpha ip} = \ln \frac{20^k \times (f_{\alpha ip} + \sqrt{N}/20^k)}{(N + \sqrt{N})} \quad k = 1,\ 1 \le i \le L$$

$$[p, L] = \begin{cases} [1, L1] & \text{for positive start samples} \\ [2, L1] & \text{for negative start samples} \\ [3, L2] & \text{for positive end samples} \\ [4, L2] & \text{for negative end samples} \end{cases} . \tag{3}$$

Here, $f_{\alpha ip}$ refers to the absolute frequency of amino acid α in the i-th position of N aligned sequences for the p-th matrix. With these four matrixes, the scoring functions (SF) are given as follows:

$$SF_p = \sum_{i=1}^{L} PWM_{\alpha ip}. \tag{4}$$

The genuine start or end sample tends to have a larger value of SF_p ($p = 1$ or 3) and a smaller value of SF_p ($p = 2$ or 4).

Stage 3: SVM classifier. Support vector machine (SVM) [24] is a supervised algorithm based on statistical learning theory. Its basic idea is to construct an optimal hyperplane as the discriminative surface with the largest distance to the nearest training data points of any class. Like in some other publications [25–27], software LIBSVM [28] was used to fulfill the classification between positive start (end) samples and negative start (end) samples. LIBSVM can be obtained from the website http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Stage 4: RNH prediction. If the inquired sequence did not contain both of start and end site identified by SVM, it was inferred that no RNH domain exists in this sequence. Otherwise, the np-RNH could be predicted by the recognition of start and end site.

3.2 Performance assessment

The proposed RNH model was tested by jack-knife and n-fold cross-validation (10-fold CV or 5-fold CV) methods for an objective and comprehensive assessment. The quantitative evaluation results were measured by sensitivity (Se), specificity (Sp), overall accuracy (Acc) and Matthew's correlation coefficient (Mcc). They are defined as follows:

$$\begin{cases} Se = \dfrac{Tp}{Tp + Fn} \\ Sp = \dfrac{Tn}{Tn + Fp} \\ Acc = \dfrac{Tp + Tn}{Tp + Tn + Fp + Fn} \\ Mcc = \dfrac{Tp \times Tn - Fp \times Fn}{\sqrt{(Tp + Fn)(Tp + Fp)(Tn + Fn)(Tn + Fp)}} \end{cases} , \tag{5}$$

where Tp and Fn are the numbers of positive samples that are predicted to be positive and negative respectively, analogously, Tn and Fp are the numbers of negative samples that are predicted to be negative and positive respectively.

4 Results and Discussion

4.1 Parameter analysis

Two sets of parameters will be discussed in detail in this section. They are the range of candidate sites (*d*1, *d*2) and length of samples (*L*1, *L*2). The appropriate values of these parameters were trained by jack-knife and n-fold CV methods based on the benchmark dataset.

The absolute distances between positive sites and corresponding predicted sites by SW-alignment were counted for *d*1 and *d*2 parameters. They are all smaller than 20 amino acids (AA). Thus, samples prepared as formula (1) and formula (2) definitely contain all positive ones and a certain number of negative ones for a better classification by SVM method.

The curves of *Mcc* versus different sample lengths were plotted in Fig 1 for *L*1 and *L*2 parameters. In this figure, *Mcc* values with *L*1 = 12 AA and *L*2 = 14 AA reach the peaks with fewer calculations under jack-knife test, as well as show very small differences from others under 10-fold CV test. Thus, the optimal *L*1 and *L*2 were set as aforesaid when taken into account both of the performance and computational amounts.

4.2 Validity analysis

The proposed model has been tested from different aspects to demonstrate its validity in np-RNH prediction (Table 1, Fig 2). The first one is a test by increasingly used and widely recognized validation methods [16]. The classification accuracy rates are 100%, 99.01% and 97.52% respectively corresponding to jack-knife, 10-fold CV and 5-fold CV test. Another one is a comparison with three other classical database search tools. The differences are 20.95, 24.76, and 80.95 points after prediction accuracy of RNH model minus that of PSI-BLAST, CS-BLAST, and HMMER3 severally. The last one is a contrast of np-RNHs predicted in *pol* sequences and non-*pol* sequences. RNH model discovered 14,033 and zero np-RNHs after separately scanning 149,692 *pol* sequences and 320 non-*pol* sequences, which is in consistency with the existence of RNH as a domain in reverse transcriptase of *pol* [4, 8]. These results all confirm the validity of RNH model to a certain degree.

4.3 Sequence analysis

Weblogo software [29, 30] was employed to generate RNH motif in Fig 3. In this figure, the underlined D10, E58, D81 and D153 are four active-site residues based on 71 annotations of

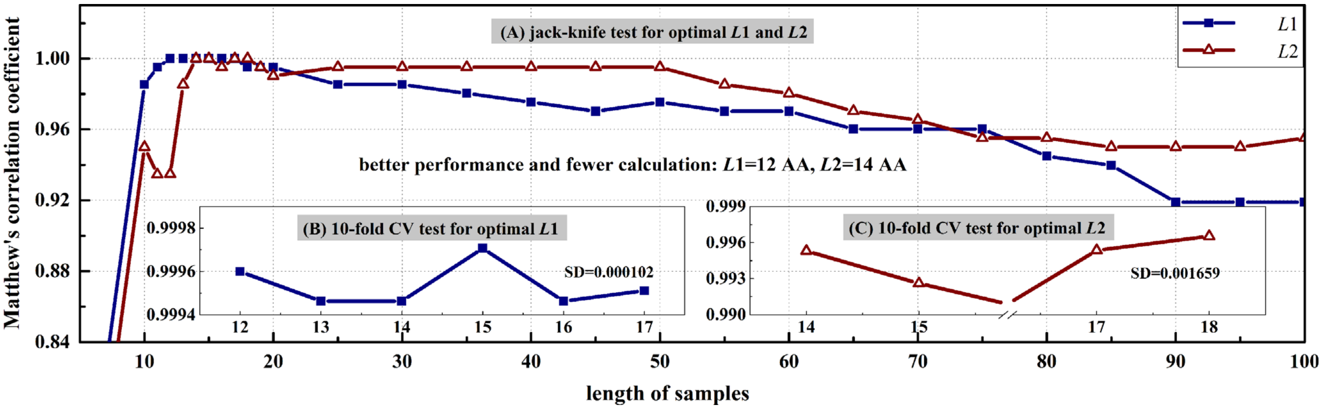

Fig 1. Change of average *Mcc* values versus different sample lengths. (A) Performance changes along with sample lengths from 10 AA to 100 AA under jack-knife test. (B)-(C) Performance varies along with optimal sample lengths selected in Fig 1(A) under 10-fold CV test. SD denotes the standard deviation of *Mcc* values.

doi:10.1371/journal.pone.0161913.g001

Table 1. Comparison results between RNH tool and three other classical database search tools.

Method	Train/test		Scan		
			Protein [d]	number	np-RNHs
RNH tool	**5-fold CV**	*Acc* = 97.52% [a]	*pol*	149,692	14,033
		Se1 = 99.62%, *Sp1* = 100%, *Acc1* = 99.99%, *Mcc1* = 99.80% [b]			
		Se2 = 97.90%, *Sp2* = 100%, *Acc2* = 99.95%, *Mcc2* = 98.92% [c]			
	10-fold CV	*Acc* = 99.01%			
		Se1 = 99.91%, *Sp1* = 100%, *Acc1* = 99.998%, *Mcc1* = 99.96%	non-*pol*	320	0
		Se2 = 99.08%, *Sp2* = 100%, *Acc2* = 99.98%, *Mcc2* = 99.53%			
	jack-knife	*Acc* = 100%			
	self-consistency	*Acc* = 105/105 = 100%			
PSI-BLAST	*Acc* = 83/105 = 79.05% (100%-79.05% = 20.95%)		-	-	-
CS-BLAST	*Acc* = 79/105 = 75.24% (100%-75.24% = 24.76%)		-	-	-
HMMER3	*Acc* = 20/105 = 19.05% (100%-19.05% = 80.95%)		-	-	-

[a]: the performance of RNH domain prediction.

[b]: the performance of start site prediction.

[c]: the performance of end site prediction.

[d]: the proteins without RNH annotations for scanning.

doi:10.1371/journal.pone.0161913.t001

the benchmark dataset. These residues can be grouped into the DEDD motif, which coordinates with divalent metal ions to facilitate RNA hydrolysis during the catalytic process [31]. It is noteworthy that regions around DEDD motif are more conserved than others. The result supports the literatures [32, 33] that DEDD motif is essential for metal binding and catalytic activity. These conservativeness properties around D10 and D153 make it convenient to recognize start and end sites of RNH domains.

Two properties were analyzed on RNH domains in Fig 4. One is length property and the other is hydrophobicity calculated by grand average of hydropathy (GRAVY) [34, 35]. This

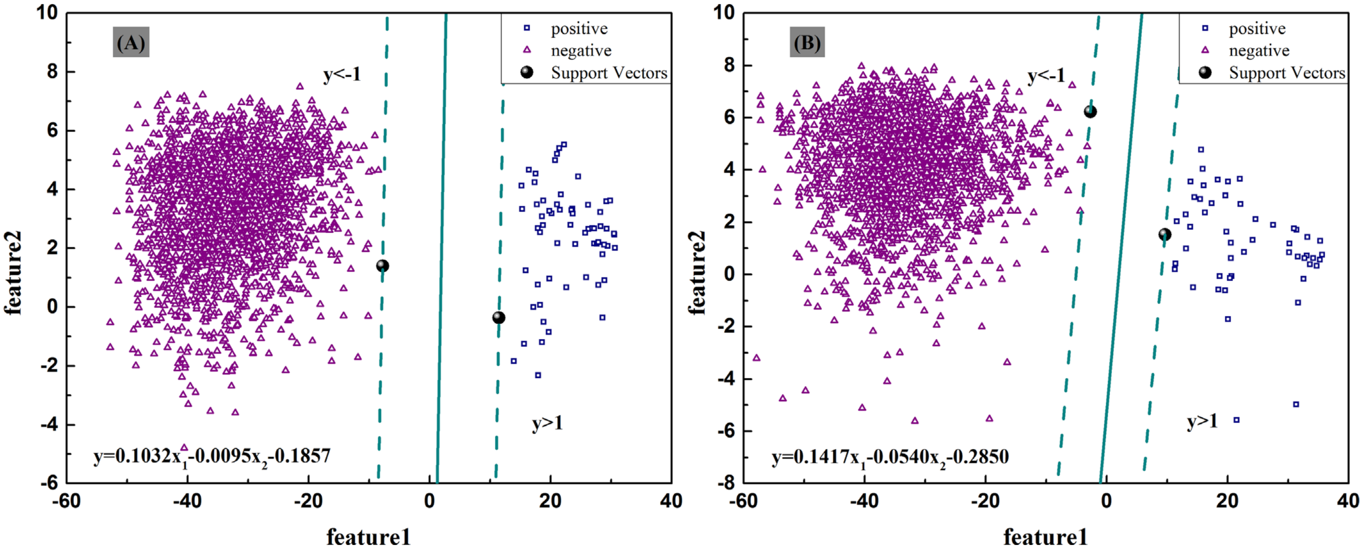

Fig 2. Classification results between positive and negative samples by SVM method. The solid lines represent the optimal classification hyperplanes constructed for (A) start samples and (B) end samples.

doi:10.1371/journal.pone.0161913.g002

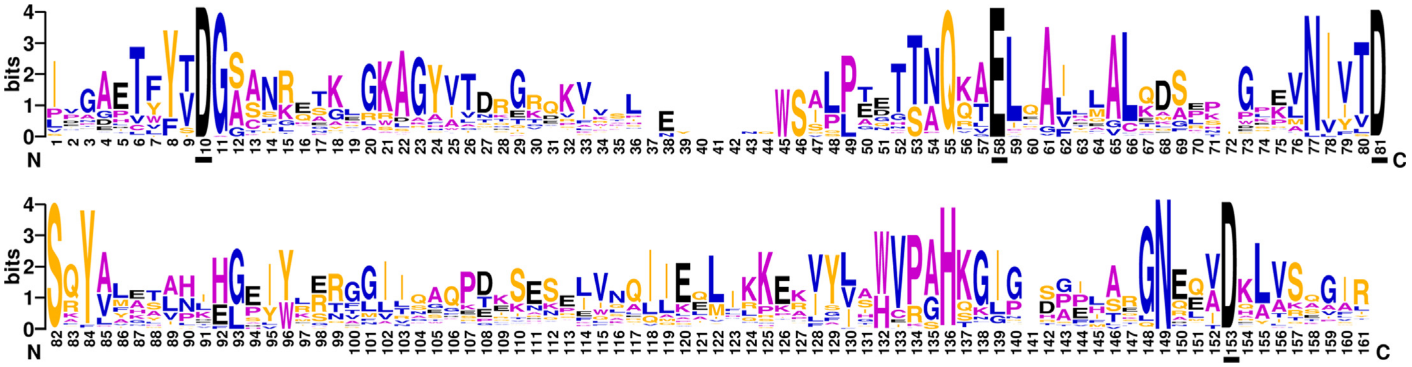

Fig 3. RNH motif plotted by weblogo software. The overall height of the stack indicates the sequence conservation at each position, the height of symbols within the stack represents the relative frequency of amino acid at that position, while the four underlined residues (D10, E58, D81 and D153) display the DEDD motif in RNH domain.

doi:10.1371/journal.pone.0161913.g003

figure shows that RNHs in the same genus share particular length and GRAVY values. In other words, it can be concluded that length and hydrophobicity of RNH domains may have some relationships with retroviral genera. Based on the conclusion, Intracisternal A-particle (unclassified-retrovirus) may be a part of alpha-retrovirus, beta-retrovirus or delta-retrovirus. On the other hand, these two characteristics of RNH domains can be regarded as a corroboration of retroviral genera defined by *pol* and other elements [1, 2].

4.4 Evolutionary relationship analysis

MEGA software [36] was used to create the phylogenic tree by maximum likelihood method in Fig 5. This figure gives a comparison result between homology of RNHs within genera and that of inter-genera. It is expected and clear that RNHs in the same genus show higher homology than different genera. Thus, Intracisternal A-particle (unclassified-retrovirus) should be classified as beta-retrovirus, coinciding with analysis result in section 4.3. From another perspective, the evolutionary relationship in RNH of retroviruses can be seen as a technical validation of their classification in the literature [1], which uses *pol* sequence as one of the elements to define retroviral genera.

4.5 RNH prediction software

A software tool has been developed to further facilitate the application of proposed prediction model. It was implemented with C programing language for windows environment. This tool receives and sends files both with the format of FASTA. The input files contain inquired amino acid sequences, while the output files include the predicted np-RNHs. Each np-RNH annotation in export files indicates clearly its start and end site in the inquired sequence.

5 Conclusion

A computational RNH model was put forward in this paper based on amino acid sequences only with high classification accuracy. This model discovered 14,033 np-RNHs after scanning sequences without RNH annotations. Based on these predicted np-RNHs and annotated RNHs, a preliminary experiment was performed to analyze the length, hydrophobicity and evolutionary relationship of RNH domains. The results indicate a correlation between these

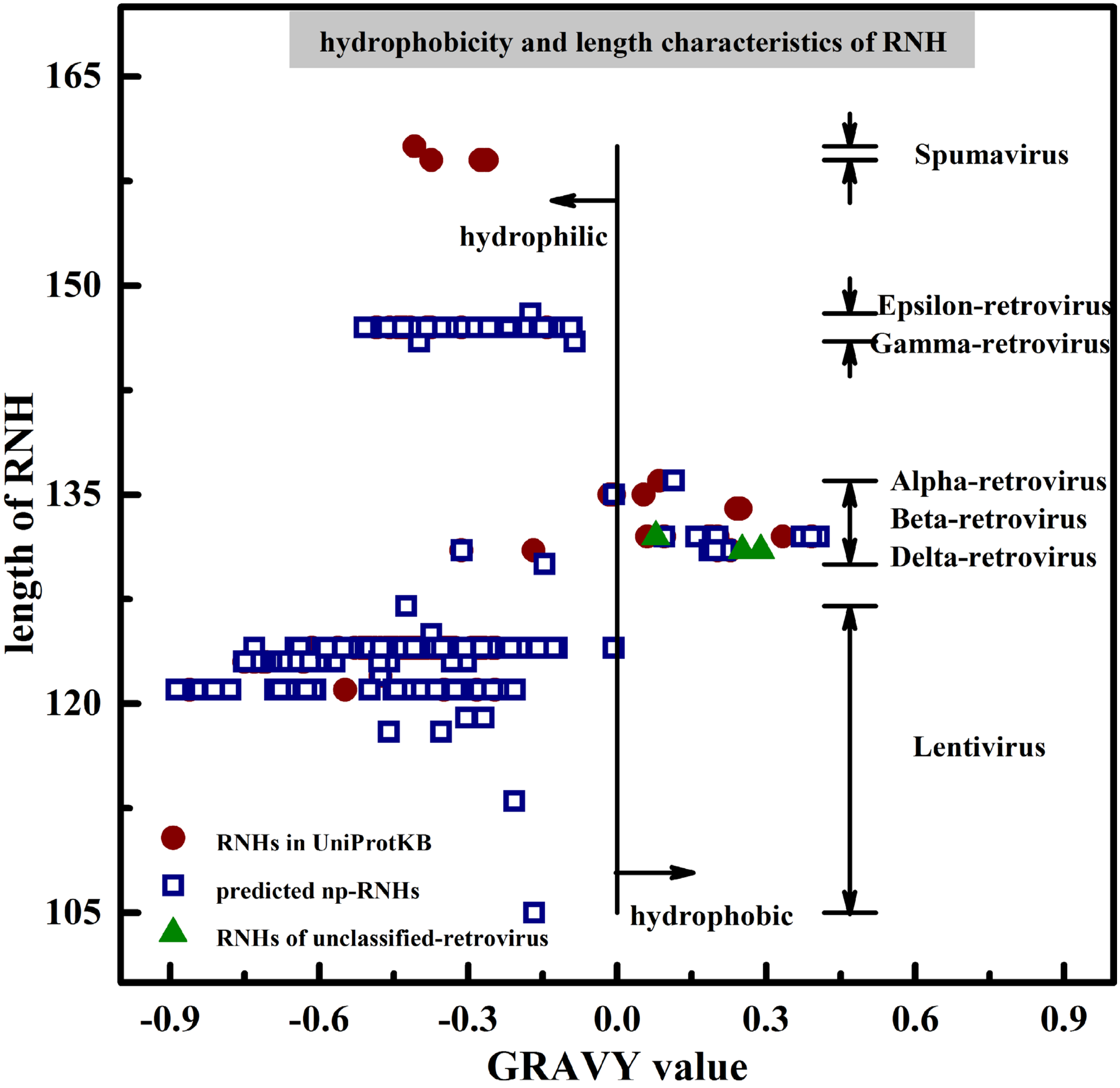

Fig 4. Hydrophobicity and length characteristics analyzed on RNH domains. RNHs in lentivirus are hydrophilic whose lengths range from 105 AA to 127 AA, RNHs in alpha-retrovirus, beta-retrovirus and delta-retrovirus are hydrophilic or hydrophobic whose lengths range from 130 AA to 136 AA, RNHs in epsilon-retrovirus and gamma-retrovirus are hydrophilic whose lengths range from 146 AA to 148 AA, while RNHs in spumavirus are hydrophilic whose lengths range from 159 AA to 160 AA.

doi:10.1371/journal.pone.0161913.g004

three characteristics and retroviral genera, which confirms the classification of retroviruses to some extent. To further facilitate the application of our proposed model, software named RNHtool has been developed.

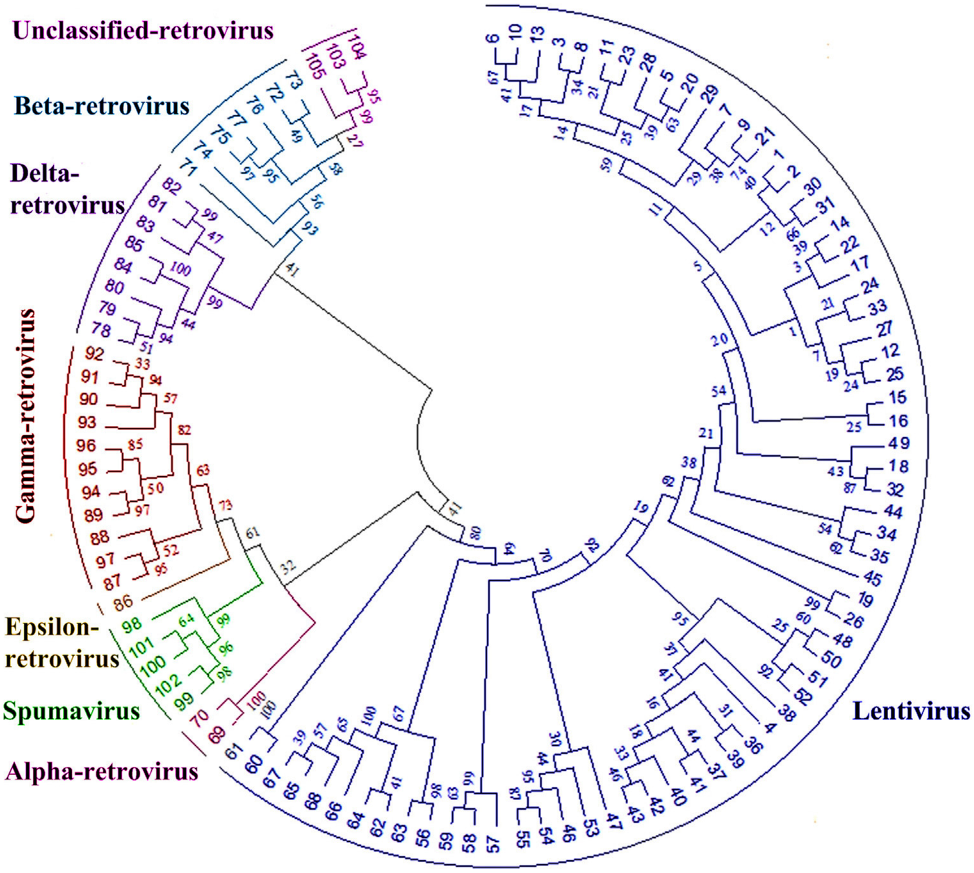

Fig 5. The evolutionary relationship in RNH of retroviruses. The leaf nodes denote annotated RNHs in the benchmark dataset, and the edge lengths describe the phylogenic relationship between these nodes.

doi:10.1371/journal.pone.0161913.g005

Acknowledgments

We are grateful to our colleagues in the School of Electronic and Information Engineering, Xi'an Jiaotong University for their help during the course of this work.

Author Contributions

Conceptualization: SJW JQH.

Data curation: SJW.

Formal analysis: SJW JQH.

Funding acquisition: JQH.

Investigation: SJW.

Methodology: SJW XMZ JQH.

Project administration: JQH XMZ.

Resources: JQH.

Software: SJW.

Supervision: JQH.

Validation: SJW.

Visualization: SJW.

Writing – original draft: SJW XMZ.

Writing – review & editing: SJW XMZ.

References

1. Coffin JM. Structure and Classification of Retroviruses. Springer US; 1992. pp. 19–49.
2. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006; 3(3):1–11.
3. Telesnitsky A, Goff SP. Reverse Transcriptase and the Generation of Retroviral DNA. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997.
4. Rosenberg N. Overview of Retrovirology. Springer New York; 2010. pp. 1–30.
5. Temin HM. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proceedings of the National Academy of Sciences. 1993; 90(15):6900–6903.
6. Miller JT, Rausch JW, Grice SFJL. Evaluation of Retroviral Ribonuclease H Activity. Humana Press; 2001. pp. 335–354.
7. Champoux JJ, Schultz SJ. Ribonuclease H: properties, substrate specificity and roles in retroviral reverse transcription. Febs Journal. 2009; 276(6):1506–1516. doi: 10.1111/j.1742-4658.2009.06909.x PMID: 19228195
8. Beilhartz GL, Götte M. HIV-1 Ribonuclease H: Structure, Catalytic Mechanism and Inhibitors. Viruses. 2010; 2(4):900–926. doi: 10.3390/v2040900 PMID: 21994660
9. Klarmann GJ, Hawkins ME, Le GS. Uncovering the complexities of retroviral ribonuclease H reveals its potential as a therapeutic target. Aids Reviews. 2002; 4(4):183–194. PMID: 12555693
10. O'Neill HC. The diversity of retroviral diseases of the immune system. Immunology and Cell Biology. 1992; 70:193–199. PMID: 1333443
11. Tomozumi I. Action of anti-HIV drugs and resistance: reverse transcriptase inhibitors and protease inhibitors. Current Pharmaceutical Design. 2004; 10(32):4039–4053. PMID: 15579086
12. Johnson WE. Assisted suicide for retroviruses. Nature Biotechnology. 2007; 25(6):643–644. PMID: 17557098
13. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25 (8):3389–3402.
14. Biegert A, Söding J. Sequence context-specific profiles for homology searching. Proceedings of the National Academy of Sciences. 2009; 106(10):3770–3775.
15. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. 2011; 39(8):W29–W37.
16. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology. 2011; 273(1):236–247. doi: 10.1016/j.jtbi.2010.12.024 PMID: 21168420
17. Wu S, Han J, Liu R, Liu J, Lv H. A computational model for predicting fusion peptide of retroviruses. Computational Biology and Chemistry. 2016; 61:245–250. doi: 10.1016/j.compbiolchem.2016.02.013 PMID: 26963379
18. Consortium UP. UniProt: a hub for protein information. Nucleic Acids Research. 2015; 43(Database issue):D204–D212. doi: 10.1093/nar/gku989 PMID: 25348405
19. Magrane M, Consortium UP. UniProt Knowledgebase: a hub of integrated protein data. Database-the Journal of Biological Databases and Curation. 2011; 2011(1):bar009–bar009.
20. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981; 147(1):195–197. PMID: 7265238
21. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000; 16(16):16–23.
22. Sinha S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics. 2006; 22(14):e454–e454. PMID: 16873507

23. Xia X. Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction. Scientifica. 2012; 2012:917540–917540. doi: 10.6064/2012/917540 PMID: 24278755

24. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20(3):273–297.

25. Byvatov E, Schneider G. Support vector machine applications in bioinformatics. Applied Bioinformatics. 2003; 2(2):67–77. PMID: 15130823

26. Cui Y, Han J, Zhong D, Liu R. A novel computational method for the identification of plant alternative splice sites. Biochemical and Biophysical Research Communications. 2013; 431(2):221–224. doi: 10.1016/j.bbrc.2012.12.131 PMID: 23313482

27. Lv H, Han J, Liu J, Zheng J, Liu R, Zhong D. CarSPred: A Computational Tool for Predicting Carbonylation Sites of Human Proteins. Plos One. 2014; 9(10):e111478–e111478. doi: 10.1371/journal.pone.0111478 PMID: 25347395

28. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. Acm Transactions on Intelligent Systems and Technology. 2011; 2(3):389–396.

29. Schneider TD, Stephens RM. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Research. 1990; 18(20):6097–6100. PMID: 2172928

30. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Research. 2004; 14(6):1188–1190. PMID: 15173120

31. Nowotny M, Yang W. Stepwise analyses of metal ions in RNase H catalysis from substrate destabilization to product release. Embo Journal. 2006; 25(9):1924–1933. PMID: 16601679

32. Katayanagi K, Okumura M, Morikawa K. Crystal structure of Escherichia coli RNase HI in complex with Mg^{2+} at 2.8 Å resolution: proof for a single Mg^{2+}-binding site. Proteins: Structure Function and Bioinformatics. 1993; 17(4):337–346.

33. Ho M-H, De VM, Dal PM, Klein ML. Understanding the effect of magnesium ion concentration on the catalytic activity of ribonuclease H through computation: does a third metal binding site modulate endonuclease catalysis? Journal of the American Chemical Society. 2010; 132(39):13702–13712. doi: 10.1021/ja102933y PMID: 20731347

34. Gasteiger E, Hoogland C, Gattiker A, Duvaud SE, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. Proteomics Protocols Handbook. 2005; 112(112):571–607.

35. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology. 1982; 157(1):105–132. PMID: 7108955

36. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. 2011; 28(10):2731–2739. doi: 10.1093/molbev/msr121 PMID: 21546353