genes

MDPI

Article

HIV Protease and Integrase Empirical Substitution Models of Evolution: Protein-Specific Models Outperform Generalist Models

Roberto Del Amparo [1,2] and Miguel Arenas [1,2,3,*]

[1] Centro de Investigacións Biomédicas (CINBIO), University of Vigo, 36310 Vigo, Spain; rdelamparo@uvigo.es
[2] Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain
[3] Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain
* Correspondence: marenas@uvigo.es; Tel.: +34-986-130047

Abstract: Diverse phylogenetic methods require a substitution model of evolution that should mimic, as accurately as possible, the real substitution process. At the protein level, empirical substitution models have traditionally been based on a large number of different proteins from particular taxonomic levels. However, these models assume that all of the proteins of a taxonomic level evolve under the same substitution patterns. We believe that this assumption is highly unrealistic and should be relaxed by considering protein-specific substitution models that account for protein-specific selection processes. In order to test this hypothesis, we inferred and evaluated four new empirical substitution models for the protease and integrase of HIV and other viruses. We found that these models more accurately fit, compared with any of the currently available empirical substitution models, the evolutionary process of these proteins. We conclude that evolutionary inferences from protein sequences are more accurate if they are based on protein-specific substitution models rather than taxonomic-specific (generalist) substitution models. We also present four new empirical substitution models of protein evolution that could be useful for phylogenetic inferences of viral protease and integrase.

Keywords: substitution model of protein evolution; protein evolution; phylogenetic reconstruction; viral protease; viral integrase; HIV

check for updates

Citation: Del Amparo, R.; Arenas, M. HIV Protease and Integrase Empirical Substitution Models of Evolution: Protein-Specific Models Outperform Generalist Models. *Genes* **2022**, *13*, 61. https://doi.org/10.3390/genes13010061

Academic Editor: Domenico Lio

Received: 9 December 2021
Accepted: 22 December 2021
Published: 27 December 2021

1. Introduction

Substitution models of molecular evolution are well established in a variety of phylogenetic methods to obtain accurate inferences of past evolutionary processes [1]. At the protein level, substitution models are frequently applied in evolutionary biology to infer phylogenetic trees [2,3], ancestral sequences [4,5] and selection [6,7], among other applications [1,8]. The current substitution models of protein evolution can be classified roughly into two categories. First, there are parametric or structure-based substitution models that consider structural constraints to model selection on the protein folding stability and function [9–13]. These models provided accurate inferences of protein evolution [10,12]; however, although some of them have been implemented in useful evolutionary frameworks [14,15], their mathematical complexity (i.e., most of them account for site-dependent evolution) and large computational requirements prevented (for the moment) their establishment in phylogenetics. The other category includes empirical substitution models of protein evolution [1,8,16]. These substitution models consist of a 20 × 20 matrix of relative rates of change among amino acids (hereafter, an exchangeability matrix) and 20 amino acid frequencies, which are estimated from large protein databases. These models assume that all of the protein sites evolve under the same substitution process, despite the fact that this is often unrealistic (i.e., it is likely that sites located in the catalytic region of an enzyme evolve under different evolutionary patterns compared to sites located at the protein surface due to selection on the protein function [13,15], and these models also ignore the protein folding stability, leading to unrealistically unstable proteins [17]). However, the mathematical simplicity and rapid computation of empirical substitution models favored their establishment in protein phylogenetics, including their implementation in most of the frameworks for phylogenetic tree reconstructions, e.g., [18], and ancestral sequence reconstructions, e.g., [19,20].

Next, most of the empirical substitution models of protein evolution are based on general nuclear (i.e., JTT [21] and WAG [22]) or mitochondrial (i.e., MtMam [16] and mtREV [23]) proteins, and others were developed from proteins of particular taxonomic levels, including viruses like the human immunodeficiency virus (HIV) [24], influenza virus [25], dengue virus [26] and flavivirus [27], among others. Still, the currently available set of empirical substitution models of protein evolution is very limited, with less than 100 substitution models [1]. Next, it is known that the accuracy of phylogenetic inferences depends on the applied substitution model [28–32]; consequently, the selection of the best-fitting substitution model of evolution currently constitutes a fundamental step in phylogenetics [33]. The limited number of currently available empirical substitution models of protein evolution means that, for a particular dataset of protein sequences, one could not find an appropriate substitution model. For example, using a framework for substitution model selection, the authors of [34] found that the best-fitting empirical substitution model for datasets from proteobacteria and archaea was a model inferred from retroviral Pol proteins, which is likely to improperly describe the evolutionary processes of the cited datasets. Therefore, there is a need for more empirical substitution models of protein evolution, at least while realistic structure-based substitution models are not yet established in phylogenetics. Regarding this concern, as noted previously, most of currently available empirical substitution models of protein evolution are largely generalist (i.e., a single substitution model is based on all of the different proteins existing in a taxonomic level). Regarding this concern, we believe that empirical substitution models that are specific for protein families could more accurately mimic the evolution of a protein dataset belonging to the underlying protein family. For example, at present, phylogenetic inferences from a dataset of HIV protease (PR) or integrase (IN) sequences can be performed under the HIVw or HIVb substitution models [24], which are the only currently available empirical substitution models based on HIV proteins. However, these two empirical substitution models are based on all of the different proteins present in this virus, and thus we consider them to be generalist. As a consequence, we believe that these models can be highly unrealistic when modeling a dataset of a specific HIV protein (i.e., PR or IN). This intuitive reasoning motivated us to investigate whether a protein-specific empirical substitution model of evolution could outperform the currently available set of generalist empirical substitution models of evolution that are commonly used in phylogenetics. In order to test this hypothesis, and also to provide new empirical substitution models that can be useful for certain viral phylogenetic inferences, we developed and evaluated four novel protein-specific empirical substitution models of evolution. In particular, we developed two models for viral PR (one for the HIV PR and another one for the PR of multiple viruses; hereafter, HIVpr and VIRpr, respectively) and two models for viral IN (one for the HIV IN and another one for the IN of multiple viruses; hereafter, HIVin and VIRin, respectively). Next, we evaluated the fitting of these models with other models (including HIVb and HIVw) using independent test data.

2. Materials and Methods

In this section, we describe the data collection, the development of the empirical substitution models of PR and IN evolution, and the evaluation of the developed models by likelihood-based comparisons with currently available empirical substitution models. All of these methodological steps are illustrated in Figure 1.

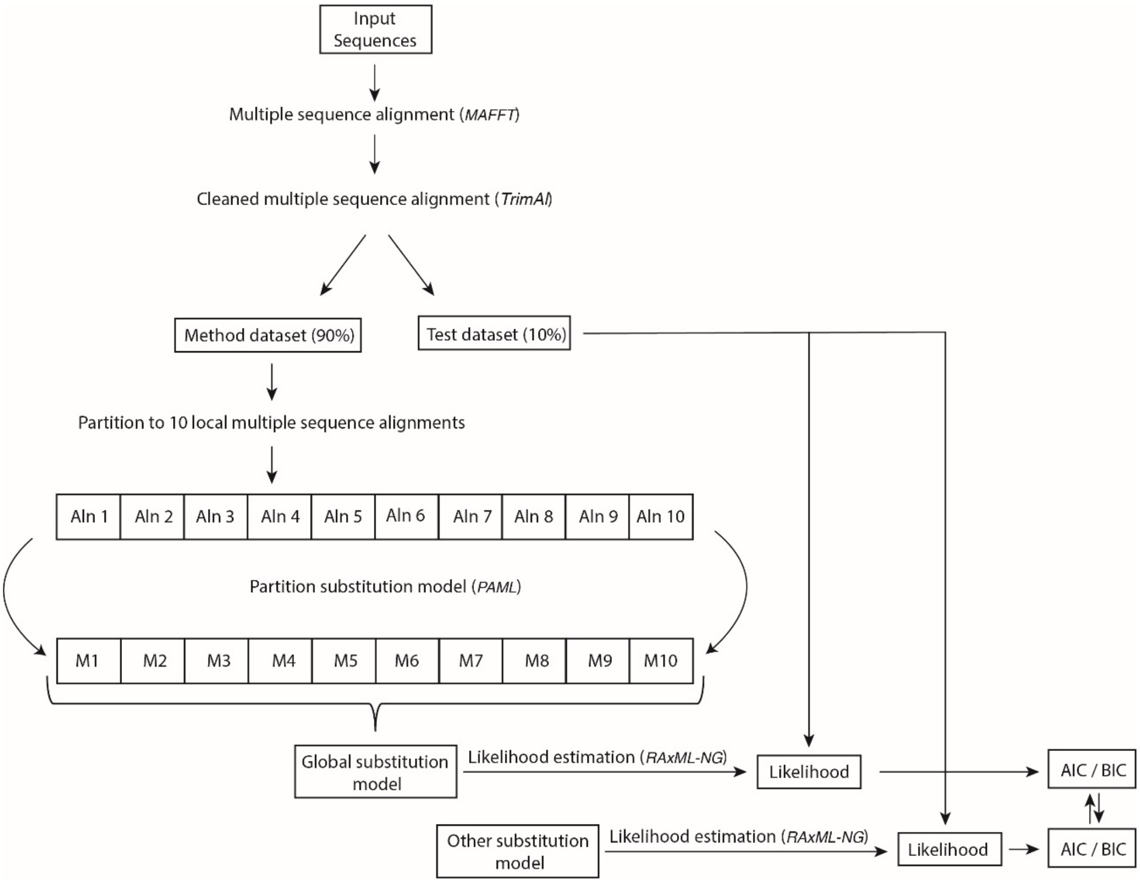
Input Sequences
Multiple sequence alignment (MAFFT)
Cleaned multiple sequence alignment (TrimAl)
Method dataset (90%)
Test dataset (10%)
Partition to 10 local multiple sequence alignments
Aln 1
Aln 2
Aln 3
Aln 4
Aln 5
Aln 6
Aln 7
Aln 8
Aln 9
Aln 10
Partition substitution model (PAML)
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
Global substitution model
Likelihood estimation (RAxML-NG)
Likelihood
AIC / BIC
Other substitution model
Likelihood estimation (RAxML-NG)
Likelihood
AIC / BIC

Figure 1. Pipeline for the inference and evaluation of the empirical substitution models of PR and IN evolution. The input protein sequences were aligned and cleaned (removing duplicate sequences and uninformative sites). Next, the resulting multiple-sequence alignment (MSA) was split into two datasets: a method dataset (for the inference of the substitution model, including most of the sequences) and a test dataset (for the evaluation of the substitution model). Indeed, the method dataset was split into 10 local method datasets (due to computational limitations), and we inferred a local (partition) substitution model for each one. The resulting local substitution models were averaged to obtain a global substitution model. Finally, we calculated the AIC and BIC scores for the global substitution model and other currently available empirical substitution models in order to evaluate them, considering the likelihood of every model with the test dataset.

2.1. Study Data of HIV-1 and General Virus Protease and Integrase

We collected all of the protein sequences of HIV PR and IN available in GenBank to develop the substitution models of HIV PR and IN, respectively. We only considered sequences with a length similar to the natural length of HIV PR (99 amino acids) and IN (163 amino acids) in order to avoid uninformative sequences with multiple indels. We obtained a total of 55,000 and 23,000 sequences for HIV PR and IN, respectively. Next, for each dataset, we obtained a multiple-sequence alignment (MSA) using MAFFT [35]. The resulting MSAs were further refined by removing sequences with multiple gaps (we only allowed sequences with less than 30% gaps) with TrimAl [36], following previous studies, e.g., [25,37]. The final MSAs included 16,900 and 14,764 sequences with lengths of 99 and 163 amino acids for the HIV PR and IN, respectively. Concerning the development of substitution models based on PR and IN from multiple viruses, we collected sequences of viral PR and IN from the PFAM database (codes PF00077 and PF00665 for the PR and IN, respectively). We also applied the previously indicated filtering to obtain the final MSAs, which included a total of 1605 and 34,282 sequences for the viral PR and IN, respectively.

2.2. Inference of Novel Empirical Substitution Models for HIV and General Virus Protease and Integrase

We split every dataset in two datasets: (i) the method dataset, which includes 90% of the sequences, and was used to infer the substitution model, and (ii) the test dataset, which includes 10% of the sequences, and was used to evaluate the developed substitution model. The inference of an empirical substitution model requires a large number of sequences [38], and therefore we incorporated most of them into this group. However, we note that 10% of the sequences provided to the test datasets include a large number of sequences (1700 and 1500 for HIV PR and IN, respectively; 144 and 3400 for general viral PR and IN, respectively). Actually, we benefited from the large number of sequences that are available for these proteins. Note that they have been frequently sequenced due to their relevant role as common antiretroviral drug targets [39–42]. Next, we found that the large number of sequences present in some method datasets (in particular, those with more than 10,000 sequences, which are 3 of the 4 method datasets) caused computational limitations that forced us to split them into 10 partitions with the same size (Figure 1). For each partitioned method dataset, we inferred an empirical substitution model (in particular, the exchangeability matrix and amino acid frequencies) under the maximum-likelihood (ML) method implemented in *PAML* [19], using an ML phylogenetic tree previously reconstructed with *RAxML-NG* [18] under the best-fitting substitution model selected by *ProtTest* [43]. We allowed *PAML* to internally optimize the model and phylogenetic tree according to the input data [19]. Using the ML method implemented in *PAML*, we obtained 10 local exchangeability matrices and sets of amino acid frequencies that we applied to calculate (by their average) the global exchangeability matrix and amino acid frequencies (Figure 1).

2.3. Evaluation of the Novel Empirical Substitution Models of HIV-1 and General Virus Protease and Integrase

We evaluated the inferred empirical substitution models using the test datasets. First, we applied *ProtTest* to every test dataset in order to identify the best-fitting empirical substitution model among the currently available empirical substitution models. Next, we obtained the likelihood and the Akaike Information Criterion (AIC) [44] and Bayesian Information Criterion (BIC) [45] scores of the fitting of every substitution model (including the corresponding empirical substitution model developed in this study and the top 5 of the currently available empirical substitution models that best fit with the studied test dataset and were selected by *ProtTest*) with the test dataset using *RAxML-NG* (Figure 1). Additionally, we applied a statistical *t*-test to compare the AIC and BIC scores obtained from the empirical substitution models developed in this study with the scores obtained from other currently available empirical substitution models (the top 5 best-fitting empirical substitution models among the set of currently available empirical substitution models) with the test datasets.

3. Results and Discussion

3.1. Novel Empirical Substitution Models for HIV and General Virus Protease and Integrase

We developed protein-specific empirical substitution models of evolution for the viral PR (one for the HIV PR and another one for the PR of multiple viruses; HIVpr and VIRpr, respectively) and for the viral IN (one for the HIV IN and another one for the IN of multiple viruses; HIVin and VIRin, respectively). These four new empirical substitution models of protein evolution are presented in Tables S1–S4 (Supplementary Material). The developed substitution models are based on symmetric exchangeability matrices (Q), which despite potentially being more unrealistic than asymmetric exchangeability matrices, are well-established in phylogenetics due to the more simple calculation of the probability matrix (P) for a given period of time (t), $P(t) = exp(Qt)$. Therefore, the development of symmetric exchangeability matrices allows a wider implementation of the models in phylogenetic frameworks.

We found that, among the currently available empirical substitution models (excluding the models developed in the present study), the HIVb substitution model produced the best fitting with all of the test datasets, except for the viral IN test dataset, for which the selected model was WAG. Next, we present qualitative comparisons (quantitative comparisons are shown in the next section) between the substitution models developed in this study and the currently available best-fitting substitution models (Figure 2 and Figures S1–S3, Supplementary Material). In general, we found that the HIVpr and VIRpr substitution models decreased (compared with the HIVb model) most of relative rates of change among the amino acids considering the PR function, which suggests a higher specificity. For example, the PR presents a catalytic aspartic acid (Asp-25) that is usually conserved due to selection to maintain the protein function [46,47]. Consequently, if it is not conserved, it could only change to glutamic acid (also a potent acidic nucleophile) in order to conserve the physicochemical properties and functionality of the catalytic region. In agreement with the consequences of this selection pressure, the HIVpr and VIRpr substitution models presented a lower substitution rate from aspartic acid to any other amino acid (compared with HIVb), except for its substitution to glutamic acid that presents physicochemical properties similar to aspartic acid (Figure 2 and Figure S1). We only found some amino acid changes where the HIVpr or VIRpr substitution models displayed a higher relative substitution rate than the HIVb substitution model, such as the substitutions serine/threonine and serine/asparagine (in HIVpr), which involve amino acids with similar physicochemical properties. Less-intuitive cases involved the substitutions serine/histidine (in HIVpr) or lysine/tyrosine (in VIRpr), which imply a physicochemical change between polar and basic amino acids, and can occur at the protein surface involved in the protein solubility [46,47]. Concerning comparisons between the HIVin and HIVb substitution models, and between the VIRin and WAG substitution models (note that HIVb and WAG were selected as the best-fitting substitution models among the currently available set of empirical substitution models, excluding the models developed in this study), we found again that the HIVin and VIRin models present more restrictive relative rates of change for the aspartic and glutamic acids than the selected models (HIVb and WAG) (Figures S3 and S4). Again, note that the main catalytic sites of the integrase are aspartic and glutamic acids (Asp-64, Asp-116 and Glu-152) [48,49]. Comparing HIVin and HIVb, or VIRin and WAG, we also observed a few amino acid changes with an increase in their relative rate of change, such as isoleucine/methionine, isoleucine/leucine, leucine/methionine and serine/threonine (in HIVin) or cysteine/tyrosine (in VIRin), where the amino acids involved presented similar physicochemical properties, agreeing with selection pressure for the maintenance of the protein function [40,50,51].

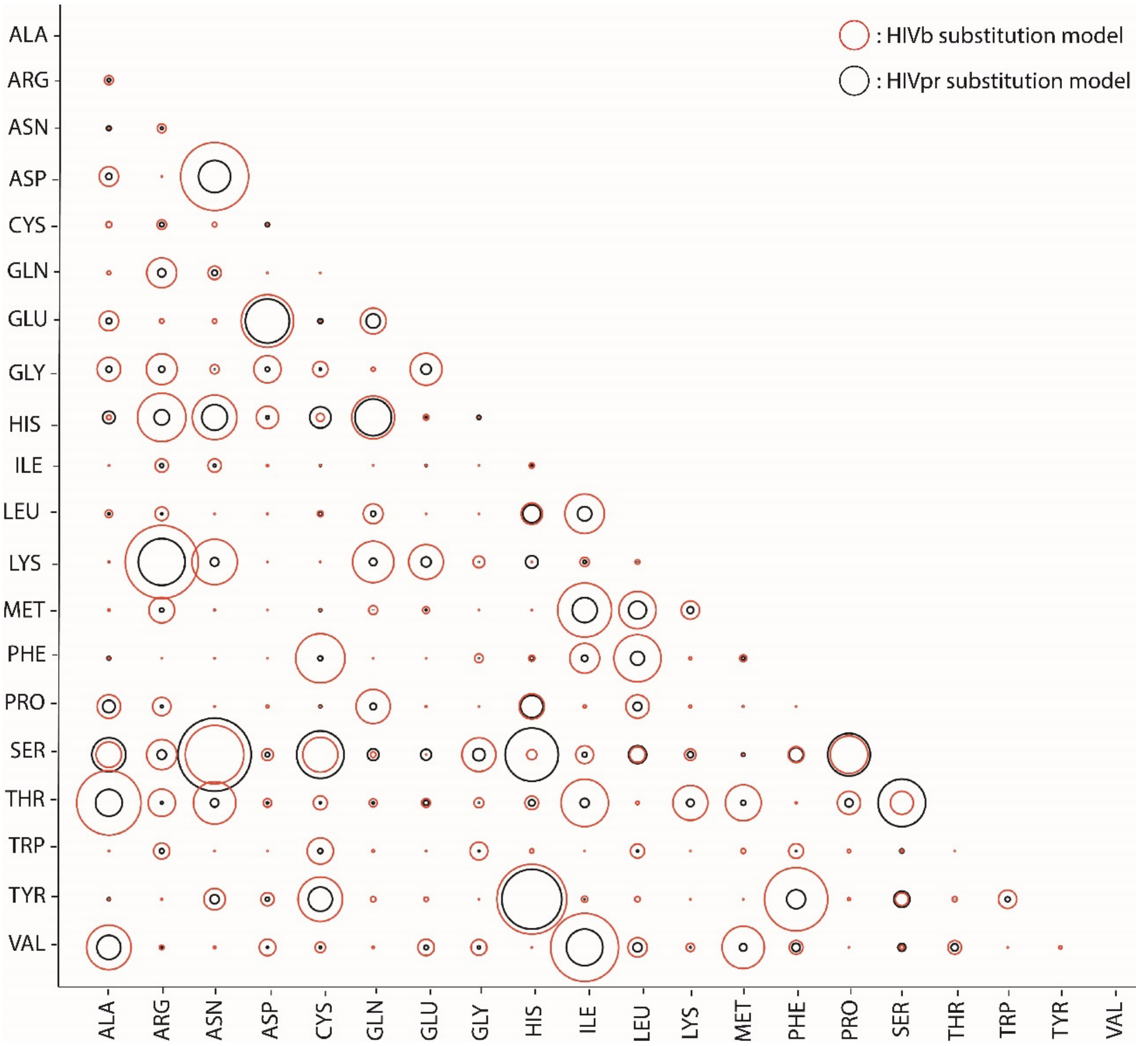

Figure 2. Comparison of HIVpr and HIVb empirical substitution models concerning their relative substitution rates. The plot displays the exchangeability matrix of the relative substitution rates among amino acids for the HIVpr (developed in this study, black circles) and HIVb (the best-fitting substitution model in the set of currently available substitution models, red circles) empirical substitution models of evolution. This plot provides an illustrative comparison between the cited models; the specific parameter values of the HIVpr substitution model are presented in Table S1.

3.2. Likelihood-Based Comparisons Indicate That the Novel Empirical Substitution Models Outperform the Currently Available Empirical Substitution Models

For every test dataset, we found that the novel substitution model that we inferred for every corresponding protein family outperforms the currently available best-fitting substitution models in terms of likelihood. In particular, we found that the currently available best-fitting substitution models (excluding the substitution models developed in this study) for the test datasets of HIV PR, viral PR, HIV IN and viral IN were HIVb, HIVb, HIVb and WAG, respectively. Next, all of the models developed in this study (HIVpr, VIRpr, HIVin and VIRin) provided lower AIC and BIC scores than the cited currently available best-fitting substitution models, and also than the top five currently available best-fitting substitution models (p-values < 0.05; Figure 3 and Figure S4; Supplementary Material). These results indicate that phylogenetic analyses of viral PR and IN are more accurate if they are based on a substitution model of evolution developed from the corresponding studied protein family (such as the substitution models developed in the present study) than if they are based on a generalist substitution model such as those which are currently available.

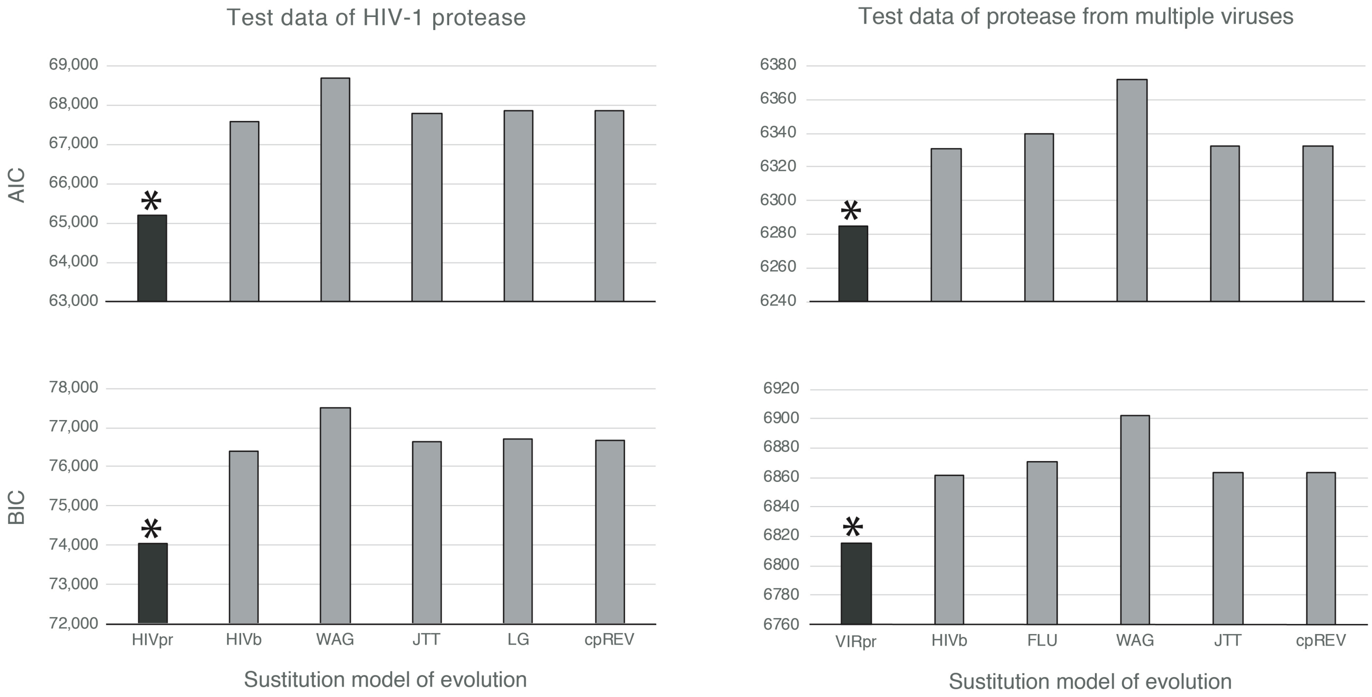

Figure 3. Likelihood-based evaluation of the HIVpr, VIRpr and currently available best-fitting substitution models. For the HIV PR (left plots) and viral PR (right plots) test datasets, the plots show the AIC (top plots) and BIC (bottom plots) scores obtained with the HIVpr and VIRpr substitution models inferred in this study and the top five currently available best-fitting substitution models with the corresponding test dataset. In all of the cases, the models developed in this study produced AIC and BIC scores (black bars) significantly lower than the currently available best-fitting substitution models (*p*-values = 0.00013 and 0.00014 for HIVpr and VIRpr, respectively and illustrated with * in the plots).

4. Conclusions

Substitution models of protein evolution are required for the most accurate phylogenetic reconstruction methods. However, the currently available set of empirical substitution models is highly limited, and mostly includes generalist models that are based on huge protein groups (i.e., nuclear or mitochondrial proteins) or on all of the proteins of a particular taxonomic level; they thus lack specificity when studying a particular protein family. Here, we show that there is a need for protein-specific empirical substitution models of evolution because they can provide accurate likelihood-based phylogenetic inferences, and we demonstrate this with the development and evaluation of four new empirical substitution models that mimic the substitution process of the PR and IN of HIV and other viruses. Of course, the accurate inference of protein-specific empirical substitution models of evolution requires a large number of protein sequences, but we believe that with the current exponential increase of protein sequences being deposited in databases, this limitation will be greatly reduced with time. Altogether, we conclude that, in order to obtain more accurate phylogenetic inferences for protein families, protein-specific empirical substitution models should be developed and applied. Indeed, we believe that the new empirical substitution models that we present in this study could be useful for evolutionary studies of viral PR and IN, which are some of the main targets of current antiretroviral drug-based treatments.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/genes13010061/s1, Table S1: Empirical substitution model of HIV PR evolution (HIVpr); Table S2: Empirical substitution model of viral PR evolution (VIRpr); Table S3: Empirical substitution model of HIV IN evolution (HIVin); Table S4: Empirical substitution model of viral IN evolution (VIRin); Figure S1. Comparison of VIRpr and HIVb empirical substitution models concerning their relative substitution rates among amino acids; Figure S2: Comparison of HIVin and HIVb empirical substitution models concerning their relative substitution rates among amino acids; Figure S3: Comparison of VIRin and WAG empirical substitution models concerning their relative substitution rates among amino acids; Figure S4: Likelihood-based evaluation of the HIVin, VIRin and the currently available best-fitting substitution models.

Author Contributions: M.A. conceived the study; R.D.A. collected the data and performed the analyses; M.A. and R.D.A. interpreted the results, and wrote and approved the final version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: We thank CESGA (Centro de Supercomputación de Galicia) for the use of the computer resources. This work was supported by the Spanish Ministry of Economy and Competitiveness [grant number RYC-2015-18241] and Xunta de Galicia [grant number ED431F 2018/08]. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The input data and the developed empirical substitution models are available at the Zenodo repository from the URL https://doi.org/10.5281/zenodo.5763867 (accessed on October 2021) [52].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Arenas, M. Trends in Substitution Models of Molecular Evolution. *Front Genet* **2015**, *6*, 319. [CrossRef] [PubMed]
2. Yutin, N.; Puigbò, P.; Koonin, E.V.; Wolf, Y.I. Phylogenomics of Prokaryotic Ribosomal Proteins. *PLoS ONE* **2012**, *7*, e36972. [CrossRef]
3. Shi, M.; Lin, X.-D.; Chen, X.; Tian, J.-H.; Chen, L.-J.; Li, K.; Wang, W.; Eden, J.-S.; Shen, J.-J.; Liu, L.; et al. The Evolutionary History of Vertebrate RNA Viruses. *Nature* **2018**, *556*, 197–202. [CrossRef] [PubMed]
4. Furukawa, R.; Toma, W.; Yamazaki, K.; Akanuma, S. Ancestral Sequence Reconstruction Produces Thermally Stable Enzymes with Mesophilic Enzyme-like Catalytic Properties. *Sci. Rep.* **2020**, *10*, 15493. [CrossRef]
5. Arenas, M.; Bastolla, U. ProtASR2: Ancestral Reconstruction of Protein Sequences Accounting for Folding Stability. *Methods Ecol. Evol.* **2020**, *11*, 248–257. [CrossRef]
6. Koshi, J.M.; Mindell, D.P.; Goldstein, R.A. Using Physical-Chemistry-Based Substitution Models in Phylogenetic Analyses of HIV-1 Subtypes. *Mol. Biol. Evol.* **1999**, *165*, 173–179. [CrossRef]
7. Bruno, W.J. Modeling Residue Usage in Aligned Protein Sequences via Maximum Likelihood. *Mol. Biol. Evol.* **1996**, *13*, 1368–1374. [CrossRef]
8. Thorne, J.L. Models of Protein Sequence Evolution and Their Applications. *Curr. Opin. Genet. Dev.* **2000**, *10*, 602–605. [CrossRef]
9. Liberles, D.A.; Teichmann, S.A.; Bahar, I.; Bastolla, U.; Bloom, J.; Bornberg-Bauer, E.; Colwell, L.J.; de Koning, A.P.; Dokholyan, N.V.; Echave, J.; et al. The Interface of Protein Structure, Protein Biophysics, and Molecular Evolution. *Protein Sci.* **2012**, *21*, 769–785. [CrossRef]
10. Arenas, M.; Sanchez-Cobos, A.; Bastolla, U. Maximum Likelihood Phylogenetic Inference with Selection on Protein Folding Stability. *Mol. Biol. Evol.* **2015**, *32*, 2195–2207. [CrossRef]
11. Parisi, G.; Echave, J. The Structurally Constrained Protein Evolution Model Accounts for Sequence Patterns of the LbetaH Superfamily. *BMC Evol. Biol.* **2004**, *4*, 41. [CrossRef] [PubMed]
12. Bordner, A.J.; Mittelmann, H.D. A New Formulation of Protein Evolutionary Models That Account for Structural Constraints. *Mol. Biol. Evol.* **2013**, *31*, 736–749. [CrossRef]
13. Echave, J. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Mol. Biol. Evol.* **2019**, *36*, 613–620. [CrossRef] [PubMed]
14. Bastolla, U.; Arenas, M. The Influence of Protein Stability on Sequence Evolution: Applications to Phylogenetic Inference. *Methods Mol. Biol.* **2019**, *1851*, 215–231.
15. Pupko, T.; Bell, R.E.; Mayrose, I.; Glaser, F.; Ben-Tal, N. Rate4Site: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Evolutionary Determinants within Their Homologues. *Bioinformatics* **2002**, *18*, S71–S77. [CrossRef] [PubMed]

16. Yang, Z.; Nielsen, R.; Hasegawa, M. Models of Amino Acid Substitution and Applications to Mitochondrial Protein Evolution. *Mol. Biol. Evol.* **1998**, *15*, 1600–1611. [CrossRef]
17. Arenas, M.; Dos Santos, H.G.; Posada, D.; Bastolla, U. Protein Evolution along Phylogenetic Histories under Structurally Constrained Substitution Models. *Bioinformatics* **2013**, *29*, 3020–3028. [CrossRef]
18. Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: A Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference. *Bioinformatics* **2019**, *35*, 4453–4455. [CrossRef]
19. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **2007**, *24*, 1586–1591. [CrossRef]
20. Kosakovsky Pond, S.L.; Frost, S.D.; Muse, S.V. HYPHY: Hypothesis Testing Using Phylogenies. *Bioinformatics* **2005**, *21*, 676–679. [CrossRef]
21. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Comput Appl Biosci* **1992**, *8*, 275–282. [CrossRef]
22. Whelan, S.; Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699. [CrossRef] [PubMed]
23. Adachi, J.; Hasegawa, M. Model of Amino Acid Substitution in Proteins Encoded by Mitochondrial DNA. *J Mol Evol* **1996**, *42*, 459–468. [CrossRef] [PubMed]
24. Nickle, D.C.; Heath, L.; Jensen, M.A.; Gilbert, P.B.; Mullins, J.I.; Kosakovsky Pond, S.L. HIV-Specific Probabilistic Models of Protein Evolution. *PLoS One* **2007**, *2*, e503. [CrossRef] [PubMed]
25. Dang, C.C.; Le, Q.S.; Gascuel, O.; Le, V.S. FLU, an Amino Acid Substitution Model for Influenza Proteins. *BMC Evol Biol* **2010**, *10*, 99. [CrossRef] [PubMed]
26. Kim, T.L.; Cao, C.D.; Le, V.S. Building a Specific Amino Acid Substitution Model for Dengue Viruses. In Proceedings of the 2018 10th International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, 1 November 2018; pp. 242–246.
27. Le, T.K.; Vinh, L.S. FLAVI: An Amino Acid Substitution Model for Flaviviruses. *J. Mol. Evol.* **2020**, *88*, 445–452. [CrossRef] [PubMed]
28. Lemmon, A.R.; Moriarty, E.C. The Importance of Proper Model Assumption in Bayesian Phylogenetics. *Syst Biol* **2004**, *53*, 265–277. [CrossRef]
29. Minin, V.; Abdo, Z.; Joyce, P.; Sullivan, J. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. *Syst. Biol.* **2003**, *52*, 674–683. [CrossRef]
30. Yang, Z.; Goldman, N.; Friday, A. Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation. *Mol. Biol. Evol.* **1994**, *11*, 316–324.
31. Zhang, J.; Nei, M. Accuracies of Ancestral Amino Acid Sequences Inferred by the Parsimony, Likelihood, and Distance Methods. *J Mol Evol* **1997**, *44* (Suppl. 1), S139–S146. [CrossRef]
32. Zhang, J. Performance of Likelihood Ratio Tests of Evolutionary Hypotheses under Inadequate Substitution Models. *Mol. Biol. Evol.* **1999**, *16*, 868–875. [CrossRef]
33. Abascal, F.; Zardoya, R.; Posada, D. ProtTest: Selection of Best-Fit Models of Protein Evolution. *Bioinformatics* **2005**, *21*, 2104–2105. [CrossRef] [PubMed]
34. Keane, T.M.; Creevey, C.J.; Pentony, M.M.; Naughton, T.J.; McLnerney, J.O. Assessment of Methods for Amino Acid Matrix Selection and Their Use on Empirical Data Shows That Ad Hoc Assumptions for Choice of Matrix Are Not Justified. *BMC Evol. Biol.* **2006**, *6*, 29. [CrossRef]
35. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
36. Capella-Gutierrez, S.; Silla-Martinez, J.M.; Gabaldon, T. TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* **2009**, *25*, 1972–1973. [CrossRef]
37. Abascal, F.; Posada, D.; Zardoya, R. MtArt: A New Model of Amino Acid Replacement for Arthropoda. *Mol. Biol. Evol.* **2007**, *24*, 1–5. [CrossRef] [PubMed]
38. Minh, B.Q.; Dang, C.C.; Vinh, L.S.; Lanfear, R. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Syst. Biol.* **2021**, *70*, 1046–1060. [CrossRef]
39. Arenas, M.; Villaverde, M.C.; Sussman, F. Prediction and Analysis of Binding Affinities for Chemically Diverse HIV-1 PR Inhibitors by the Modified SAFE_p Approach. *J. Comput. Chem.* **2009**, *30*, 1229–1240. [CrossRef]
40. Arenas, M. Genetic Consequences of Antiviral Therapy on HIV-1. *Comput. Math. Method Med.* **2015**, *2015*, 9. [CrossRef]
41. Ghosh, A.K.; Osswald, H.L.; Prato, G. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. *J. Med. Chem.* **2016**, *59*, 5172–5208. [CrossRef]
42. Hazuda, D.J. HIV Integrase as a Target for Antiretroviral Therapy. *Curr. Opin. HIV AIDS* **2012**, *7*, 383–389. [CrossRef] [PubMed]
43. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. ProtTest 3: Fast Selection of Best-Fit Models of Protein Evolution. *Bioinformatics* **2011**, *27*, 1164–1165. [CrossRef]
44. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*; Petrov, B.N., Csaki, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
45. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]

46. Weber, I.T.; Wang, Y.-F.; Harrison, R.W. HIV Protease: Historical Perspective and Current Research. *Viruses* **2021**, *13*, 839. [CrossRef]
47. Craik, C.S.; Roczniak, S.; Largman, C.; Rutter, W.J. The Catalytic Role of the Active Site Aspartic Acid in Serine Proteases. *Science* **1987**, *237*, 909–913. [CrossRef] [PubMed]
48. Engelman, A.; Craigie, R. Identification of Conserved Amino Acid Residues Critical for Human Immunodeficiency Virus Type 1 Integrase Function in Vitro. *J. Virol.* **1992**, *66*, 6361–6369. [CrossRef] [PubMed]
49. Kulkosky, J.; Jones, K.S.; Katz, R.A.; Mack, J.P.; Skalka, A.M. Residues Critical for Retroviral Integrative Recombination in a Region That Is Highly Conserved among Retroviral/Retrotransposon Integrases and Bacterial Insertion Sequence Transposases. *Mol. Cell. Biol.* **1992**, *12*, 2331–2338. [CrossRef] [PubMed]
50. Parera, M.; Fernandez, G.; Clotet, B.; Martinez, M.A. HIV-1 Protease Catalytic Efficiency Effects Caused by Random Single Amino Acid Substitutions. *Mol Biol Evol* **2007**, *24*, 382–387. [CrossRef]
51. Ribeiro, A.J.M.; Tyzack, J.D.; Borkakoti, N.; Holliday, G.L.; Thornton, J.M. A Global Analysis of Function and Conservation of Catalytic Residues in Enzymes. *J. Biol. Chem.* **2020**, *295*, 314–324. [CrossRef] [PubMed]
52. HIV Protease and Integrase Empirical Substitution Models of Evolution: Protein-Specific Models Outperform Generalist Models. Available online: https://zenodo.org/record/5763867#.YcWbnx17mjQ (accessed on 7 December 2021).