HAL open science

Simulation data for the estimation of numerical constants for approximating pairwise evolutionary distances between amino acid sequences

Thomas Bigot, Julien Guglielmini, Alexis Criscuolo

To cite this version:

Thomas Bigot, Julien Guglielmini, Alexis Criscuolo. Simulation data for the estimation of numerical constants for approximating pairwise evolutionary distances between amino acid sequences. Data in Brief, 2019, 25, pp.104212. 10.1016/j.dib.2019.104212 . pasteur-03265225

HAL Id: pasteur-03265225

https://pasteur.hal.science/pasteur-03265225v1

Submitted on 19 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Data in brief 25 (2019) 104212

Contents lists available at ScienceDirect

Data in brief

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

Data Article

Simulation data for the estimation of numerical constants for approximating pairwise evolutionary distances between amino acid sequences

Thomas Bigot [1], Julien Guglielmini [1], Alexis Criscuolo*

Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France

ARTICLE INFO

Article history:
Received 30 April 2019
Received in revised form 5 June 2019
Accepted 25 June 2019
Available online 8 July 2019

Keywords:
Amino acid
Evolutionary model
Corrected distance
Uncorrected distance
Computer simulation
Nonlinear regression

ABSTRACT

Estimating the number of substitution events per site that have occurred during the evolution of a pair of amino acid sequences is a common task in phylogenetics and comparative genomics that often requires quite slow maximum-likelihood procedures when taking into account explicit evolutionary models. Data presented in this article are large sets of numbers of substitution events and associated numbers of observed differences between pairs of aligned amino acid sequences that have been generated through a simulation procedure of sequence evolution under a broad range of evolutionary models. These data are available at https://zenodo.org/record/2653704 (doi: 10.5281/zenodo.2653704). They are accompanied in this paper by figures showing the strong relationship between the corresponding evolutionary and uncorrected distances, as well as estimated numerical constants that determine non-linear functions that fit the simulated data. These numerical constants can be useful to quickly estimate pairwise evolutionary distances directly from uncorrected distances between aligned amino acid sequences.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* Corresponding author.
E-mail address: alexis.criscuolo@pasteur.fr (A. Criscuolo).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.dib.2019.104212
2352-3409/© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Specifications table

Subject area	Computational biology, Bioinformatics
More specific subject area	Phylogenetics
Type of data	Text files, Images, Tables
How data was acquired	Computer simulation, nonlinear regression
Data format	Simulated, Analyzed
Experimental factors	Simulated data from publicly available phylogenetic trees
Experimental features	Amino acid sequence evolution simulation, evolutionary and uncorrected distance estimations, and nonlinear regression
Data source location	Institut Pasteur, Paris, France
Data accessibility	Simulation data and scatter plot figures are available at https://zenodo.org/record/2653704 (doi:10.5281/zenodo.2653704), numerical data and graphical representations of the nonlinear fitting are with this article

Value of the data

- The data proposed here should aim at enhancing the estimation of pairwise evolutionary distances between any pairs of amino acid sequences from a methodological, practical or educational point of view.
- Available simulated data can be used to develop new methods and algorithms for more accurate or faster estimates of pairwise evolutionary distances.
- Numerical data can be used to perform faster evolutionary distance estimates directly from the proportion of observed differences.
- Associated figures can be used for educational purposes to illustrate the strong relationship between evolutionary and uncorrected distances.

1. Data

Given a pair of homologous amino acid sequences, there exists a strong positive monotonic relationship between the number d of substitution events per site that have occured during their evolution and the proportion p of observed differences (often called uncorrected distance or p-distance) between the two aligned sequences [1–9]. For estimating the (unknown) evolutionary distance d from the observed value p, analytical formulae of the following form (often called gamma distance) have been proposed:

$$d = a\, b\, [\, (1 - p\, /\, b)^{-1/a} - 1\,] \tag{1}$$

where a and b are two positive numerical parameters depending on the heterogeneity of the replacement rate among amino acid pairs and sites, and on the equilibrium frequencies of amino acid residues, respectively [2,10–15]. In line with previous attempts [2,4,5,14,16], data presented here are estimations of a and b as obtained through computer simulations for 27 empirical models of amino acid substitution [1,17–37] (see names and associated references in Tables 1 and 2), as well as the associated text files containing simulation datasets (https://zenodo.org/record/2653704) and figures showing the relationship between p and d (Figs. 1–6, and image files available at https://zenodo.org/record/2653704).

Table 1
Poisson correction (PC) gamma distance: estimated values and associated statistics of the numerical constants a for 27 empirical models of amino acid substitution.

Evolutionary model	b	a			
		Estimate	95% confidence interval		Mean squared error
Dayhoff [1]	1.00000	1.99924	1.99850	1.99997	0.00121
BLOSUM62 [17]	1.00000	3.24334	3.24188	3.24481	0.00064
JTT [18]	1.00000	2.57163	2.57057	2.57270	0.00089

Table 1 (*continued*)

Evolutionary model	*b*	*a*			
		Estimate	95% confidence interval		Mean squared error
mtREV [19]	1.00000	1.23867	1.23812	1.23922	0.00496
mtMam [20]	1.00000	0.90348	0.90324	0.90372	0.00365
cpREV [21]	1.00000	1.98628	1.98556	1.98699	0.00119
VT [22]	1.00000	3.41801	3.41628	3.41975	0.00072
WAG [23]	1.00000	2.69788	2.69665	2.69910	0.00096
WAG* [23]	1.00000	2.80430	2.80305	2.80555	0.00084
rtREV [24]	1.00000	2.08011	2.07936	2.08087	0.00107
PMB [25]	1.00000	3.45924	3.45765	3.46084	0.00059
DCMut-Dayhoff [26]	1.00000	2.01070	2.00996	2.01144	0.00120
DCMut-JTT [26]	1.00000	2.55191	2.55086	2.55295	0.00088
HIVb [27]	1.00000	1.83588	1.83529	1.83646	0.00110
HIVw [27]	1.00000	1.62839	1.62776	1.62902	0.00210
MtArt [28]	1.00000	0.93628	0.93602	0.93653	0.00345
LG [29]	1.00000	2.21046	2.20952	2.21140	0.00129
MtZoa [30]	1.00000	1.05466	1.05439	1.05492	0.00235
cpREV64 [31]	1.00000	2.63503	2.63381	2.63625	0.00103
FLU [32]	1.00000	1.52820	1.52775	1.52865	0.00144
gcpREV [33]	1.00000	1.76147	1.76090	1.76205	0.00128
stmtREV [34]	1.00000	2.03813	2.03719	2.03908	0.00184
AB [35]	1.00000	1.71521	1.71480	1.71562	0.00075
mtInv [36]	1.00000	1.57997	1.57919	1.58076	0.00373
mtMet [36]	1.00000	1.40469	1.40420	1.40518	0.00240
mtVer [36]	1.00000	1.15596	1.15558	1.15634	0.00330
DEN [37]	1.00000	2.12834	2.12753	2.12915	0.00111

Table 2
Equal-input (EI) gamma distance: estimated values and associated statistics of the numerical constants *a* and *b* for 27 empirical models of amino acid substitution.

Evolutionary model	*b*	*a*			
		Estimate	95% confidence interval		Mean squared error
Dayhoff [1]	0.93993	3.14582	3.14550	3.14613	0.00005
BLOSUM62 [17]	0.94151	6.32690	6.32599	6.32782	0.00002
JTT [18]	0.94191	4.39688	4.39633	4.39744	0.00004
mtREV [19]	0.92467	1.95601	1.95578	1.95623	0.00024
mtMam [20]	0.92473	1.30527	1.30514	1.30539	0.00040
cpREV [21]	0.93916	3.14971	3.14940	3.15002	0.00005
VT [22]	0.94092	6.96847	6.96714	6.96980	0.00003
WAG [23]	0.94055	4.81653	4.81579	4.81726	0.00005
WAG* [23]	0.94055	5.01598	5.01518	5.01679	0.00005
rtREV [24]	0.94024	3.30578	3.30545	3.30612	0.00005
PMB [25]	0.94195	7.10575	7.10459	7.10691	0.00002
DCMut-Dayhoff [26]	0.93993	3.16983	3.16951	3.17015	0.00005
DCMut-JTT [26]	0.94193	4.36663	4.36607	4.36719	0.00004
HIVb [27]	0.94179	2.77572	2.77550	2.77594	0.00004
HIVw [27]	0.93819	2.45611	2.45584	2.45639	0.00012
MtArt [28]	0.92743	1.35206	1.35186	1.35226	0.00095
LG [29]	0.94051	3.56820	3.56767	3.56873	0.00009
MtZoa [30]	0.92686	1.57251	1.57227	1.57275	0.00068
cpREV64 [31]	0.93948	4.64357	4.64279	4.64436	0.00006
FLU [32]	0.94110	2.22717	2.22704	2.22731	0.00005
gcpREV [33]	0.93745	2.72778	2.72755	2.72800	0.00005
stmtREV [34]	0.92778	3.77358	3.77322	3.77395	0.00004
AB [35]	0.93407	2.78549	2.78473	2.78625	0.00058
mtInv [36]	0.92211	2.85866	2.85835	2.85897	0.00011
mtMet [36]	0.92546	2.34419	2.34387	2.34451	0.00024
mtVer [36]	0.92052	1.91274	1.91241	1.91307	0.00067
DEN [37]	0.94143	3.34672	3.34632	3.34712	0.00006

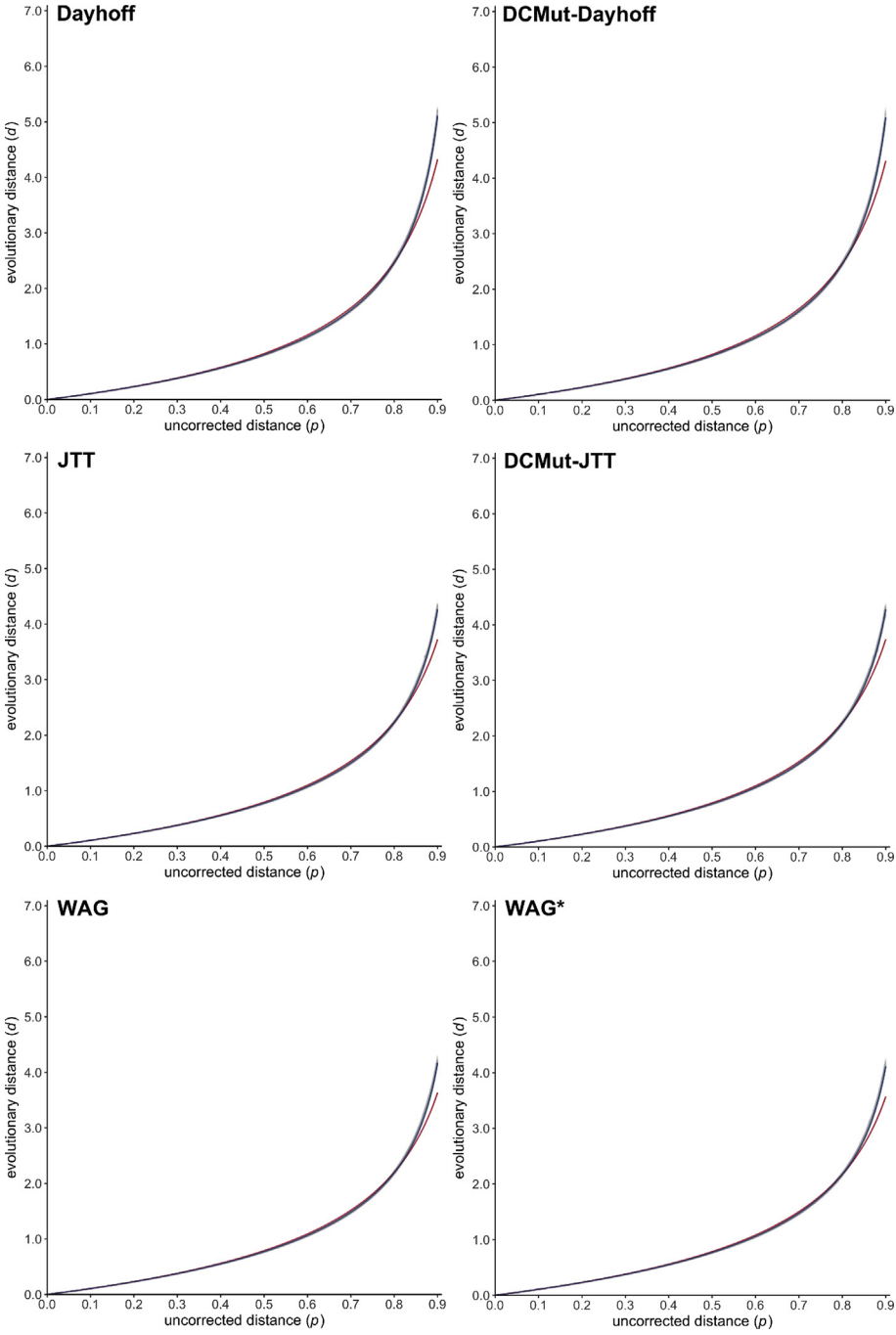

Fig. 1. Scatter plots (gray dots) representing the relationship between the uncorrected distance p (x-axis) and the evolutionary distance d (y-axis) for the three general amino acid substitution models Dayhoff [1], JTT [18] and WAG [23] (left), and for their variants DCMut-Dayhoff, DCMut-JTT [26] and WAG* [23] (right). Estimated Poisson correction (PC) and equal-input (EI) gamma distance functions are drawn in red and blue, respectively.

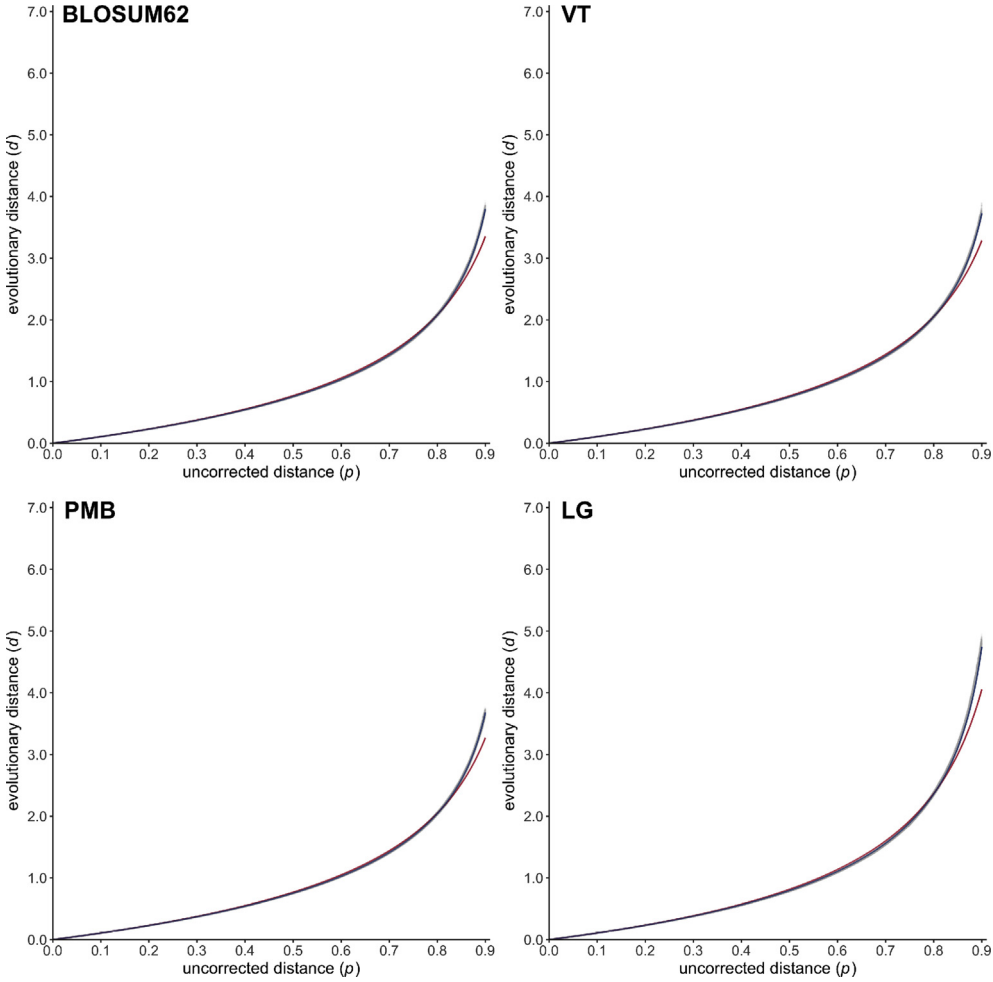

Fig. 2. Scatter plots (gray dots) representing the relationship between the uncorrected distance p (x-axis) and the evolutionary distance d (y-axis) for the four general amino acid substitution models BLOSUM62 [17], VT [22], PMB [25] and LG [29]. Estimated PC and EI gamma distance functions are drawn in red and blue, respectively.

2. Experimental design, materials, and methods

To simulate the evolution of amino acid sequences along reliable real-case phylogenetic trees, the 1,903,844 available ones on the ftp repository of PhylomeDB v4 (ftp://phylomedb.org/phylomedb) were considered, as they have been inferred by a workflow including homologous sequence clustering and alignment from a broad range of genes and phyla (eukaryota, bacteria and archaea; see details at http://phylomedb.org) followed by maximum likelihood phylogenetic inference [38]. A reduced subset of these trees was built to obtain a wide array of induced patristic distances that are quite evenly distributed over [0, 20] (see y-axis ranges in Figs. 1–6): for m growing from 0.0001 to 20 (step = 0.001), one tree (at least 25 taxa) was picked out such that its diameter (i.e. maximum patristic distance) was as close as possible to m. Following this procedure, 20,000 real-case phylogenetic trees representative of a comprehensive range of evolutionary events and distances were selected. For each considered evolutionary model (see Tables 1 and 2), the evolution of a

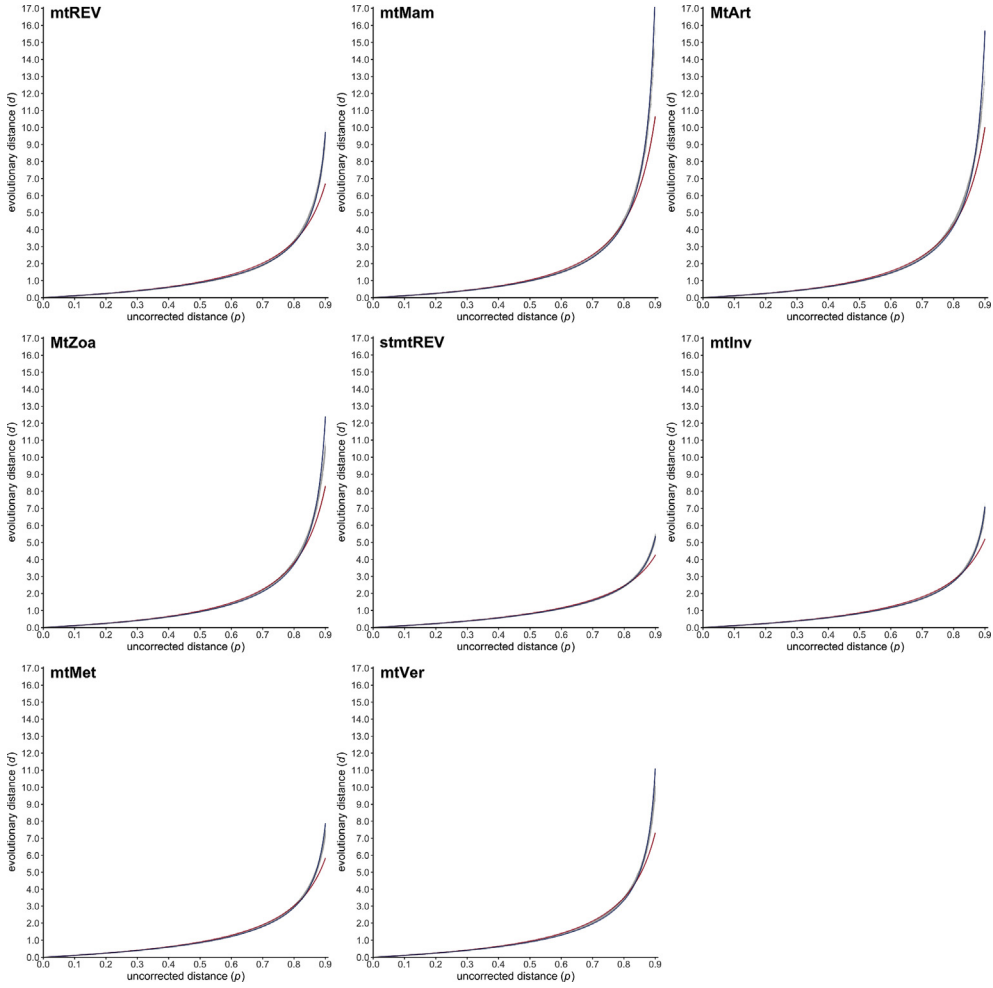

Fig. 3. Scatter plots (gray dots) representing the relationship between the uncorrected distance *p* (x-axis) and the evolutionary distance *d* (y-axis) for the eight mitochondrial amino acid substitution models mtREV [19], mtMam [20], MtArt [28], MtZoa [30], stmtREV [34], mtInv, mtMet and mtVer [36]. Estimated PC and EI gamma distance functions are drawn in red and blue, respectively.

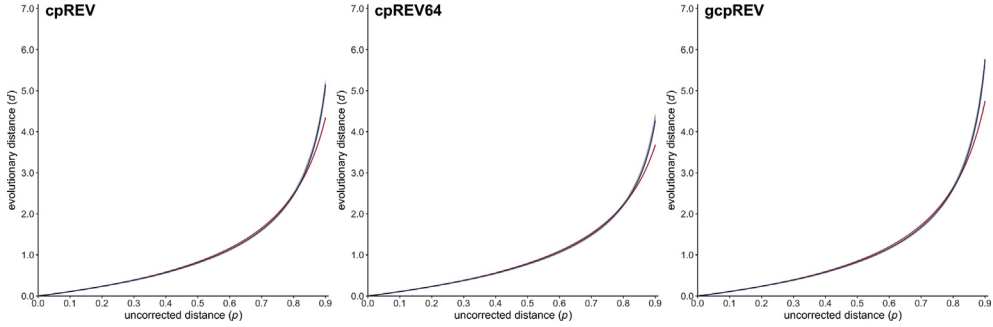

Fig. 4. Scatter plots (gray dots) representing the relationship between the uncorrected distance *p* (x-axis) and the evolutionary distance *d* (y-axis) for the three amino acid substitution models cpREV [21], cpREV64 [31] and gcpREV [33] dedicated to plastid-encoded protein sequences. Estimated PC and EI gamma distance functions are drawn in red and blue, respectively.

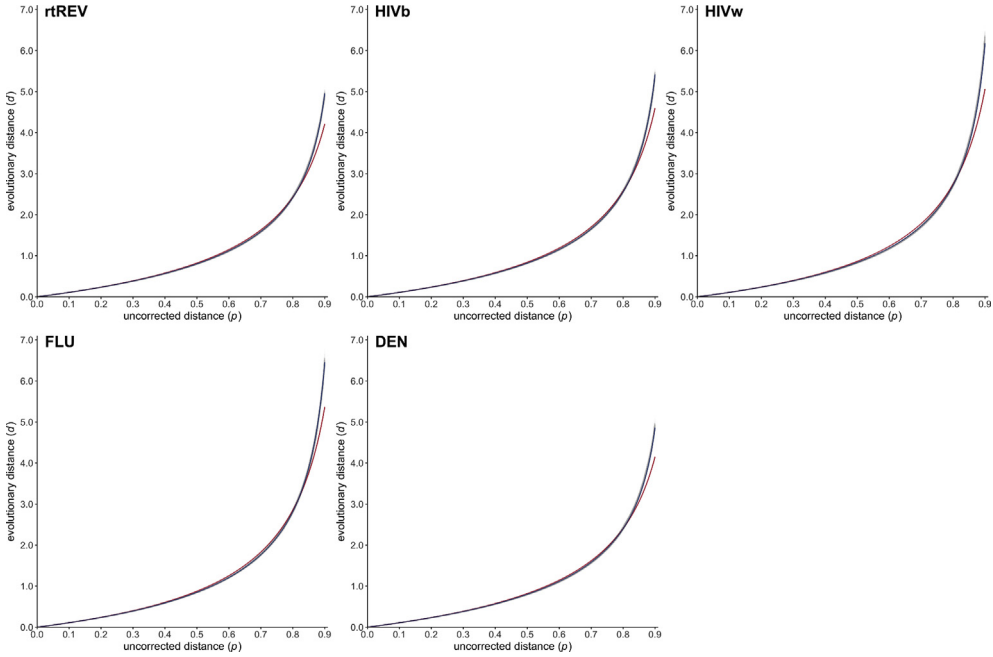

Fig. 5. Scatter plots (gray dots) representing the relationship between the uncorrected distance *p* (x-axis) and the evolutionary distance *d* (y-axis) for five amino acid substitution models dedicated to retrovirus (rtREV [24]), HIV (HIVb, HIWw [27]), influenza (FLU [32]), and dengue (DEN [37]) protein sequences. Estimated PC and EI gamma distance functions are drawn in red and blue, respectively.

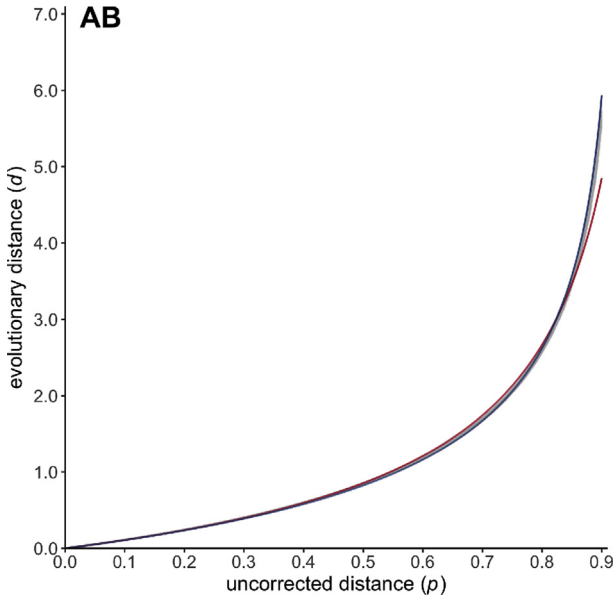

Fig. 6. Scatter plot (gray dots) representing the relationship between the uncorrected distance *p* (x-axis) and the evolutionary distance *d* (y-axis) for the antibody-specific model of amino acid substitution AB [35]. Estimated PC and EI gamma distance functions are drawn in red and blue, respectively.

sequence of 50,000 amino acid residues was simulated using INDELible v1.03 [39] along each of the 20,000 selected phylogenetic trees, and the matrix of observed p-distances was computed from each of the simulated multiple sequence alignments using FastME v2.1.5 [40]. Next, for each evolutionary model, a subset of simulated data (i.e. phylogenetic tree, simulated multiple sequence alignment, and corresponding p-distance matrix) was selected to obtain at least 500,000 values p that approximately follow a uniform distribution over [0, 0.9]. For each of those selected multiple sequence alignments, the branch lengths of the associated phylogenetic tree were refitted using RAxML-NG v0.8.1 BETA [41] with the corresponding evolutionary model, and the matrix of patristic distances d was computed using gotree v0.2.10 (https://github.com/evolbioinfo/gotree). Of note, for each of the 27 considered evolutionary models, INDELible and RAxML-NG were both used with the corresponding empirical replacement matrix file gathered from http://giphy.pasteur.fr/empirical-models-of-amino-acid-substitution. Finally, as each pair of distance matrices (i.e. uncorrected and evolutionary distances p and d) represents numbers of observed differences and occurred substitutions per site, respectively, each entry was multiplied by the total number of sites (i.e. 50,000) and rounded to the closest integer. This scaling and rounding step allows observing the same integer values than the ones obtained with alternative programs for branch length refitting (e.g. PhyML [42], IQ-TREE [43]) while each program leads to slightly different evolutionary distances d because of rounding errors or implementation choices (not shown).

Two versions of the nonlinear functional relationship between the evolutionary distance d and the uncorrected distance p were fitted separately to each simulated data. The first, called the Poisson correction (PC) gamma distance, is determined by fixing $b = 1$ in formula (1) [2,5,44]. The second, called the equal-input (EI) gamma distance, is determined with $b = 1 - \sum_r \pi_r^2$ in formula (1), where π_r is the equilibrium frequency of the amino acid residue r [12,13,15]. For each of the 27 considered evolutionary models, empirical values of π_r from the corresponding amino acid replacement matrix were used for computing b (Table 2). For each evolutionary model and each of the two PC and EI gamma distances, the numerical constant a was estimated by weighted nonlinear regression from the pairs of integer versions of uncorrected and evolutionary distances p and d gathered from the corresponding simulation file (see above) and divided by the number of simulated sites (i.e. 50,000). Each least-square estimation of the parameter a was performed using R v3.5.3 [45] with the function nls. Default Gauss-Newton algorithm was used with relative weighting (i.e. each d was weighted with d^{-2}) and starting value $a = 2$.

All simulation datasets are available at https://zenodo.org/record/2653704 (doi:10.5281/zenodo.2653704). The 20,000 phylogenetic trees selected for simulating sequence evolution are available as a text file together with descriptive statistics summarizing the corresponding patristic distances. For each of the 27 evolutionary models (see Tables 1 and 2), blocks of simulation data (i.e. PhylomeDB identifiers, random seeds, trees with refitted branch lengths, integer values of p and d) are available as text files. Estimated values of a are given for the PC and EI gamma distances in Tables 1 and 2, respectively, together with the associated 95% confidence intervals and mean squared errors. Figs. 1–6 represent the 27 scatter plots of simulated d against p, as well as the regression curves for the two PC and EI gamma distance functions. Each scatter plot is also available with and without the regression curves at https://zenodo.org/record/2653704.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

The authors are obliged to the Bioinformatics and Biostatistics Hub of Institut Pasteur, Paris, France, for support. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris. The authors also thank one anonymous reviewer for its fruitful comments.

Conflict of interest

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

References

[1] M.O. Dayhoff, R. M Schwartz, B.C. Orcutt, A model of evolutionary change in proteins, in: M.O. Dayhoff (Ed.), Atlas of Protein Sequence and Structure, Natl. Biomed. Res. Found., Washington DC, 1978, pp. 345–352.
[2] T. Ota, M. Nei, Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites, J. Mol. Evol. 38 (1994) 642–643. https://doi.org/10.1007/BF00175885.
[3] N.V. Grishin, Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites, J. Mol. Evol. 41 (1995) 675–679. https://doi.org/10.1007/BF00175826.
[4] R.F. Doolittle, D.-F. Feng, S. Tsang, G. Cho, E. Little, Response: dating the cenancester of organisms, Science 274 (1996) 1751–1753. https://doi.org/10.1126/science.274.5293.1751.
[5] M. Nei, S. Kumar, Evolutionary change of amino acid sequences, in: Molecular Evolution and Phylogenetics, Oxford University Press, New York, 2000, pp. 17–32.
[6] R. Zardoya, A. Meyer, Vertebrate phylogeny: limits of inference of mitochondrial genome and nuclear rDNA sequence data due to an adverse phylogenetic signal/noise ratio, in: P.E. Ahlberg (Ed.), Major Events in Early Vertebrate Evolution, Taylor & Francis, London, 2001, pp. 135–155.
[7] C. Lauber, A.E. Gorbalenya, Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses, J. Virol. 86 (2012) 3890–3904. https://doi.org/10.1128/JVI.07173-11.
[8] B. Nabholz, N. Uwimana, N. Lartillot, Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds, Genom. Biol. Evol. 5 (2013) 1273–1290. https://doi.org/10.1093/gbe/evt083.
[9] M.R. Young, P.D.N. Hebert, Patterns of protein evolution in cytochrome c oxidase 1 (COI) from the class Arachnida, PLoS One 10 (2015) e0135053, https://doi.org/10.1371/journal.pone.0135053.
[10] T. Uzzell, K.W. Corbin, Fitting discrete probability distributions to evolutionary events, Science 172 (1971) 1089–1096. https://doi.org/10.1126/science.172.3988.1089.
[11] G.B. Golding, Estimates of DNA and protein sequence divergence: an examination of some assumptions, Mol. Biol. Evol. 1 (1983) 125–142. https://doi.org/10.1093/oxfordjournals.molbev.a040303.
[12] F. Tajima, M. Nei, Estimation of evolutionary distance between nucleotide sequences, Mol. Biol. Evol. 1 (1984) 269–285. https://doi.org/10.1093/oxfordjournals.molbev.a040317.
[13] F. Tajima, Unbiased estimation of evolutionary distance between nucleotide sequence, Mol. Biol. Evol. 10 (1993) 677–688. https://doi.org/10.1093/oxfordjournals.molbev.a040031.
[14] X. Gu, The age of the common ancestor of eukaryotes and prokaryotes: statistical inferences, Mol. Biol. Evol. 14 (1997) 861–866. https://doi.org/10.1093/oxfordjournals.molbev.a025827.
[15] K. Tamura, S. Kumar, Evolutionary distance estimation under heterogeneous substitution pattern among lineages, Mol. Biol. Evol. 19 (2002) 1727–1736. https://doi.org/10.1093/oxfordjournals.molbev.a003995.
[16] J. Zhang, M. Nei, Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods, J. Mol. Evol. 44 (1997) S139–S146. https://doi.org/10.1007/PL00000067.
[17] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. U.S.A. 89 (1992) 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.
[18] D.T. Jones, W.R. Taylor, J.M. Thornton, The rapid generation of mutation data matrices from protein sequences, Comput. Appl. Biosci. 8 (1992) 275–282. https://doi.org/10.1093/bioinformatics/8.3.275.
[19] J. Adachi, M. Hasegawa, Model of amino acid substitution in proteins encoded by mitochondrial DNA, J. Mol. Evol. 42 (1996) 459–468. https://doi.org/10.1007/BF02498640.
[20] Z. Yang, R. Nielsen, M. Hasegawa, Models of amino acid substitution and applications to mitochondrial protein evolution, Mol. Biol. Evol. 15 (1998) 1600–1611. https://doi.org/10.1093/oxfordjournals.molbev.a025888.
[21] J. Adachi, P.J. Waddell, W. Martin, M. Hasegawa, Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA, J. Mol. Evol. 50 (2000) 348–358. https://doi.org/10.1007/s002399910038.
[22] T. Muller, M. Vingron, Modeling amino acid replacement, J. Comput. Biol. 7 (2000) 761–776. https://doi.org/10.1089/10665270050514918.
[23] S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, Mol. Biol. Evol. 18 (2001) 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851.
[24] M.W. Dimmic, J.S. Rest, D.P. Mindell, R.A. Goldstein, rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny, J. Mol. Evol. 55 (2002) 65–73. https://doi.org/10.1007/s00239-001-2304-y.
[25] S. Veerassamy, A. Smith, E.R. Tillier, A transition probability model for amino acid substitutions from blocks, J. Comput. Biol. 10 (2003) 997–1010. https://doi.org/10.1089/106652703322756195.
[26] C. Kosiol, N. Goldman, Different versions of the Dayhoff rate matrix, Mol. Biol. Evol. 22 (2005) 193–199. https://doi.org/10.1093/molbev/msi005.
[27] D.C. Nickle, L. Heath, M.A. Jensen, P.B. Gilbert, J.I. Mullins, S.L. Kosakovsky Pond, HIV-specific probabilistic models of protein evolution, PLoS One 2 (2007) e503, https://doi.org/10.1371/journal.pone.0000503.
[28] F. Abascal, D. Posada, R. Zardoya, MtArt: a new model of amino acid replacement for Arthropoda, Mol. Biol. Evol. 24 (2007) 1–5. https://doi.org/10.1093/molbev/msl136.
[29] S.Q. Le, O. Gascuel, An improved general amino acid replacement matrix, Mol. Biol. Evol. 25 (2008) 1307–1320. https://doi.org/10.1093/molbev/msn067.

[30] O. Rota-Stabelli, Z. Yang, M.J. Telford, MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies, Mol. Phylogenetics Evol. 52 (2009) 268–272. https://doi.org/10.1016/j.ympev.2009.01.011.
[31] B. Zhong, T. Yonezawa, Y. Zhong, M. Hasegawa, The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics, Mol. Biol. Evol. 27 (2010) 2855–2863. https://doi.org/10.1093/molbev/msq170.
[32] C.C. Dang, S.Q. Le, O. Gascuel, V.S. Le, FLU, an amino acid substitution model for influenza proteins, BMC Evol. Biol. 10 (2010) 99. https://doi.org/10.1186/1471-2148-10-99.
[33] C.J. Cox, P.G. Foster, A 20-state empirical amino-acid substitution model for green plant chloroplasts, Mol. Phylogenetics Evol. 68 (2013) 218–220. https://doi.org/10.1016/j.ympev.2013.03.030.
[34] Y. Liu, C.J. Cox, W. Wang, B. Goffinet, Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias, Syst. Biol. 63 (2014) 862–878. https://doi.org/10.1093/sysbio/syu049.
[35] A. Mirsky, L. Kazandjian, M. Anisimova, Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences, Mol. Biol. Evol. 32 (2015) 806–819. https://doi.org/10.1093/molbev/msu340.
[36] V.S. Le, C.C. Dang, S.Q. Le, Improved mitochondrial amino acid substitution models for metazoan evolutionary studies, BMC Evol. Biol. 17 (2017) 136. https://doi.org/10.1186/s12862-017-0987-y.
[37] T.K. Le, C.C. Dang, S.V. Le, Building a specific amino acid substitution model for Dengue viruses, in: T.M. Phuong, M.L. Nguyen (eds), Proceedings of 10th International Conference on Knowledge and Systems Engineering (KSE 2018), Ho Chi Minh City, Vietnam, pp. 242–246. https://doi.org/10.1109/KSE.2018.8573341.
[38] J. Huerta-Cepas, S. Capella-Gutiérrez, L.P. Pryszcz, M. Marcet-Houben, T. Gabaldón, PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome, Nucleic Acids Res. 42 (2014) D897–D902. https://doi.org/10.1093/nar/gkt1177.
[39] W. Fletcher, Z. Yang, INDELible: a flexible simulator of biological sequence evolution, Mol. Biol. Evol. 26 (2009) 1879–1888. https://doi.org/10.1093/molbev/msp098.
[40] V. Lefort, R. Desper, O. Gascuel, FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program, Mol. Biol. Evol. 32 (2015) 2798–2800. https://doi.org/10.1093/molbev/msv150.
[41] A.M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference, bioRxiv, 2018. https://doi.org/10.1101/447110.
[42] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel O, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Syst. Biol. 59 (2010) 307–321. https://doi.org/10.1093/sysbio/syq010.
[43] L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies, Mol. Biol. Evol. 32 (2015) 268–274. https://doi.org/10.1093/molbev/msu300.
[44] M. Nei, P. Xu, G. Glazko, Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 2497–2502. https://doi.org/10.1073/pnas.051611498.
[45] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. https://www.R-project.org.