Phylogenomics using Compression Distances: Incorporating Rate Heterogeneity and Amino Acid Properties

Edward L. Braun
Department of Biology, University of Florida,, Gainesville, FL, USA
ebraun68@ufl.edu

ABSTRACT

Efforts to reconstruct the tree of life have a long history, but the field has changed fundamentally in the genomic era. Phylogenomics examines evolutionary relationships using very large datasets, so a major problem in the field is the development of unbiased computational methods for tree inference. Sources of bias include sequence alignment errors, discordance among gene trees, and long branch attraction. Distances based on data compression can address sequence alignment errors and analyses of distances may be robust to a major source of discordance among gene trees (incomplete lineage sorting). However, compression distances appear to be susceptible to long branch attraction. This study tested the hypothesis that compression distances can be modified to be more resistant to long branch attraction and found that correcting compression distances for multiple substitutions improved their behavior. Calculating distances after grouping amino acids based on their physicochemical properties incorporated more biological information. The modified compression distances used in this study also made it possible to estimate tree support using a method that closely resembles the bootstrap, the most popular support metric in phylogenomics.

CCS CONCEPTS

• **Applied computing** → **Molecular evolution**; **Molecular sequence analysis**; **Computational genomics**.

KEYWORDS

phylogenetics, protein evolution, data compression, Kolmogorov complexity, multispecies coalescent, incomplete lineage sorting, long branch attraction

ACM Reference Format:
Edward L. Braun. 2023. Phylogenomics using Compression Distances: Incorporating Rate Heterogeneity and Amino Acid Properties. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, Article 111, 6 pages. https://doi.org/10.1145/3584371.3612985

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09
https://doi.org/10.1145/3584371.3612985

1 INTRODUCTION

Understanding the historical relationships among living organisms is an major goal in the fields of comparative genomics and phylogenetics. In the current "phylogenomic era" the primary source of tree estimation error is systematic rather than stochastic in nature. Adding data ameliorates stochastic error but it exacerbates systematic error [20]. Standard methods of phylogenetic estimation (e.g., maximum likelihood [ML] with commonly-used models of sequence evolution) will yield incorrect estimates of topology if those sources of error are strong enough. Long-branch attraction (LBA) is the best characterized source of systematic error [12], but changes in state frequencies (i.e., shifts in nucleotide or amino acid frequencies) is also important [20]. However, discordance among gene trees has come to the forefront as a source of systematic error in phylogenomics [11]. This discordance reflects processes like incomplete lineage sorting (ILS) that result in genuine conflict among true gene trees [11, 29].

The simplest analytical approach in phylogenomics is ML analysis of concatenated multiple sequence alignments for many different loci. However, this "ML concatenation" approach can yield an incorrect topology when there is ILS [22, 35]. Summary coalescent analyses, which estimate individual gene trees (typically using ML) and combine those trees [29], are the most commonly used solution to the ILS problem. Summary coalescent analyses have many advantages but errors in the estimated gene trees can be problematic for summary coalescent methods [26, 40]. Perhaps surprisingly, analyses of concatenated data using certain distance methods were recently shown to be statistically consistent (i.e., to converge on the true tree in the limit of infinite data) [1, 8].

Genetic distance calculations typically use pairs of aligned sequences, but data compression methods can be used to alignment-free calculate distances [7, 23, 24]. Compression distances are based on the idea that compressing the concatenation of two very similar files results in a smaller file than compressing the concatenation of two very different files. In phylogenetics, the *normalized compression distance* (*NCD*) would be calculated using equation 1 with files of sequences represented as one-letter codes (i.e., nucleotides ∈ {A, C, G, T} or the 20 amino acids):

$$NCD(x,y) = \frac{C(x,y) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}} \quad (1)$$

In equation 1, $C(x)$ and $C(y)$ are compressed file sizes for sequences from taxa x and y and $C(x,y)$ is the compressed file size for their concatenation. *NCD* exhibited good performance in a simulation study that assumed ILS [45]. However, it has seldom been used in empirical studies; most *NCD* trees have used mammalian mitogenomic data (Fig. 1). Unfortunately, those trees strongly suggest the *NCD* is susceptible to LBA. Compression distances must be less

sensitive to LBA to be useful phylogenomic tools. The goal of this study is to determine whether it is possible to modify the *NCD* to incorporate biological information and if doing so makes it less sensitive to sources of systematic bias, like LBA.

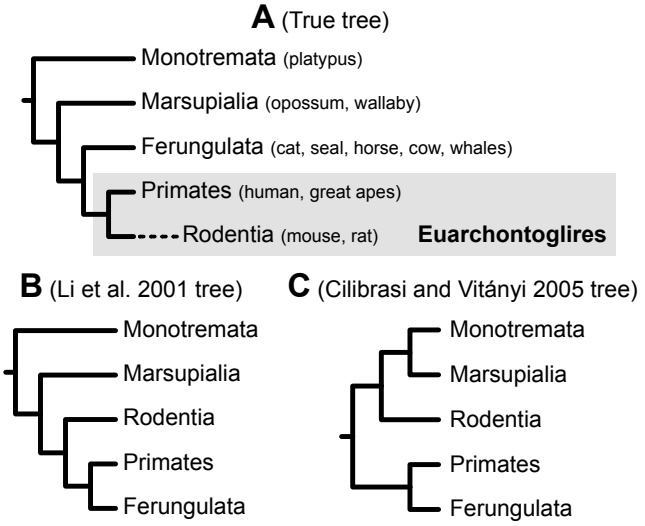

Figure 1: Published compression distance trees for mammals appear to be affected by LBA. (A) Consensus mammal tree based on many phylogenomic studies (reviewed by Murphy et al. [30]; this tree can be viewed as true. The rodent branch is extended by a dashed line to emphasize the high rate of molecular divergence in muroid rodents (e.g., rat and mouse); the high rate creates the potential for LBA that disrupts the "indicator clade" (Euarchontoglires). (B) Li et al. [23] compression tree rooted rooted between monotremes (e.g., platypus) and other mammals. (C) Cilibrasi and Vitányi [7] compression tree, rooted to a fish outgroup (not shown). Note that the only rodents in the compression distance trees [7, 23] were the rat and mouse.

2 METHODS

NCD is an approximation to an idealized but uncomputable distance based on Kolmogorov complexity [25]. The quality of this approximation depends on the properties of the data compression program. Comparing *NCD*(*x*,*y*) to *NCD*(*x*,*x*) and *NCD*(*y*,*y*), as shown in equation 2, should correct for compressor imperfections (to some degree).

$$NCD_c = \frac{NCD(x,y) - \frac{NCD(x,x)+NCD(y,y)}{2}}{1 - \frac{NCD(x,x)+NCD(y,y)}{2}} \quad (2)$$

Correcting *NCD* for the behavior of the compressor does not yield additive distances and non-additivity is the likely basis for LBA in distance analyses. This issue can be addressed using a correction analogous to the Poisson and Γ distances for aligned data [31], as shown in equations 3 and 4.

$$NCD_{\text{Poisson}} = -\ln(1 - NCD_c) \quad (3)$$

$$NCD_\Gamma = \alpha \left[(1 - NCD_c)^{-\frac{1}{\alpha}} - 1 \right] \quad (4)$$

Both equations correct the NCD_c for multiple substitutions, but equation 4 also accommodates variation in substitution rates across sites using an adjustable parameter (α) related to that variation. This is desirable because among-sites rate variation is pervasive in proteins [10]. However, different types of substitutions also accumulate at different rates in proteins (see Braun [4] for details). This can be addressed by using alternative amino acid alphabets (Table 1) might address the second issue. Four of the alphabets group amino acids based on physicochemical properties and the fifth groups captures the GC content of codons; the GC alphabet was used because amino acid frequencies are correlated with genomic base composition [37, 39].

Table 1: Amino acid alphabets

Alphabet	States	Groups
Standard	20	n/a
Dayhoff [9]	6	(C),(AGPST),(NDEQ),(RHK),(ILMV),(FWY)
Hanada [14]	4	(ANCGPST),(ILMV),(RQHKFWY),(DE)
HP	2	(RNDCQEHKSTWY),(AGILMFPV)
Size	2	(RQEHILKMFWY),(ANDCGPSTV)
GC	3	(FYMINK),(LVSTHQDECW),(GARP)

NCD values were calculated using gzip with best compression (`gzip -9`). Concatenated sequence files were "tiled"; each line in file *x*,*y* is a protein from taxon *x* immediately followed by its taxon *y* ortholog (proteins were reiterated twice in files *x*,*x* and *y*,*y*). This strategy requires ortholog identification but it does not require multiple sequence alignment. Support values analogous to the bootstrap, the support metric using in most phylogenetic studies [16]), were calculated by randomly sampling orthologs (100 orthologs for each replicate). Least squares trees were estimated from the distances using PAUP* 4.0b166 [44] and consensus trees were generated using SumTrees.py from DendroPy [43]. Programs and links to additional information (including the data used for these analyses) are available from https://github.com/ebraun68/compdisttest.

The mammalian dataset comprised 1453 orthologs from the OrthoMaM v. 10c [38]; 19 taxa that resemble the taxon sample in Fig. 1 of Li et al. [23] were analyzed. The avian dataset comprised 2590 aligned orthologs from Jarvis et al. [19]; 21 taxa were selected for analysis; TAPER [46] was used to identify homology errors in the Jarvis data, which have been noted earlier [41]. Analyses were limited to loci with all selected taxa and no TAPER masking. ML concatenation analyses used IQ-TREE, [28] and support was calculated using the ultrafast bootstrap [15]. Patristic distances for the ML trees (the sum of branch lengths connecting each pair of taxa) were calculated using the T-REX server [3].

3 RESULTS AND DISCUSSION

3.1 Transformed compression distances are less susceptible to systematic bias

The concatenated ML tree (Fig. 2) was congruent with recent genome-scale mammalian phylogenies [13, 30]; those phylogenies were generated using a variety of methods, including some that explicitly incorporate discordance among gene trees due to ILS. The only node with <100% support was Euungulata (Perissodactyla [horse and rhinoceros] + Cetartiodactyla [cow and whales]); that clade is poorly supported in other phylogenomic studies. The raw NCD_c tree broke up the LBA "indicator clade" (Euarchontogilres, see Fig. 3A), suggesting that raw NCD_c trees are susceptible to LBA under conditions where standard ML is robust. NCD_c branch lengths had proportionally longer terminal branches than the ML tree and they exhibited less heterogeneity in root to tip distances.

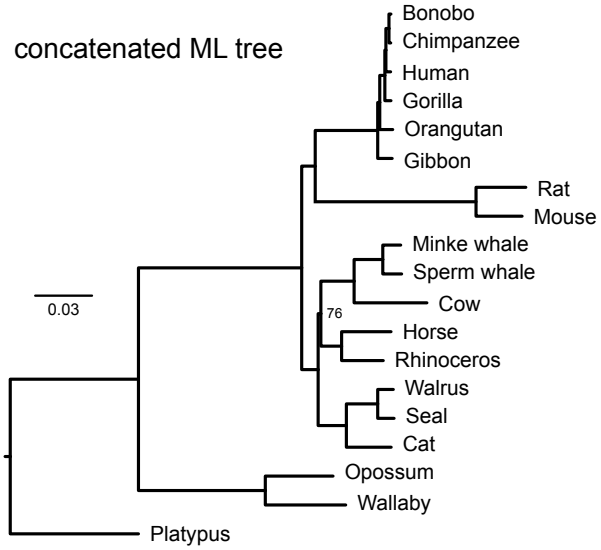

Figure 2: Concatenated ML tree for mammals. IQ-TREE topology obtained using the best-fitting model (Q.bird+F+I+G4 [27]) identified using the `-m TEST` option [21]. Support values reflect the ultrafast bootstrap [15]; unlabeled nodes had complete support.

Using NCD_Γ (equation 4) to calculate distances appeared to ameliorate LBA (Fig. 3B). Relative branch length differences between ML tree and the NCD_Γ tree were evident, suggesting that the two methods extract different information from the data. The most striking branch length difference involves the wallaby. Provocatively, NCD_Γ trees with α<0.7 had, at most, weak support for marsupial monophyly. NCD_Γ with α=0.7 is likely reasonable since the ML concatenation estimate of α for variable sites in these sequences was 0.7531. Moreover, support for Euarchontoglires was >50% for all alphabets when α was between 0.3 and 0.8 (see github). There is no ideal way to estimate α; the α parameter for aligned Γ distances is non-identifiable given pairwise comparisons [42]. ML estimates of the α parameter are themselves imperfect because ILS can lead to sites that appear artificially fast (see Fig. 5 in Houde et al. [17]). Estimating rate heterogeneity parameters is fertile area for further research if compression distances are to become a mature tool in phylogenomics.

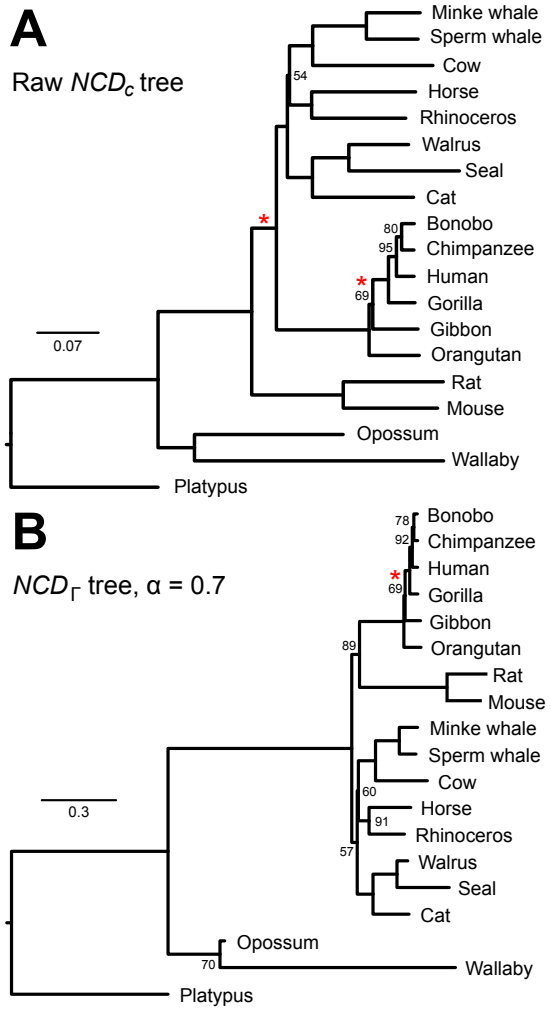

Figure 3: Comparison of uncorrected and corrected NCD_c trees for mammals. (A) Least squares tree for uncorrected ("raw") NCD_c values. Support was calculated using 100 random samples of the full ortholog set (see Methods); unlabeled nodes had complete support. Red asterisks indicate conflicts with the ML tree. (B) Least squares tree for corrected NCD_c values with α=0.7. Support and conflicts with the ML tree are indicated as in part A. Distances calculations for both trees used the standard amino acid alphabet.

3.2 Different alphabets capture distinct evolutionary information

Pairwise NCD_c values calculated using different alphabets (Table 1) were quite similar, but they were not identical (Fig. 4). Nevertheless, there were some general patterns; uncorrected NCD_c values for the two- and three-state alphabets showed consistently higher slopes than the standard 20-state alphabet whereas the uncorrected values for the four- and six-state alphabets had shallower slopes. Support

for some nodes also differed based on the alphabet. NCD_Γ trees had relatively high support for Euarchontoglires when the value of α was between 0.3 and 0.7 (Fig. 5A). However, the behavior of the Hanada and Dayhoff alphabets differed from the other alphabets (Table 1); both of those alphabets provided higher support for Euarchotoglires for α>0.7 and for α=0.2. Support for marsupial monophyly in NCD_Γ trees exhibited a pattern that was essentially the opposite of the pattern for Euarchontoglires. As described above, NCD_Γ trees α<0.7 had greatly reduced support for Marsupialia (Fig. 5B). However, the different patterns of support for Euarchontoglire and Marsupialia also extended to the alphabets; calculating NCD_Γ using the Hanada and Dayhoff alphabets always resulted in lower support marsupial monophyly, regardless of the α value. For additional details, see the github page for this project.

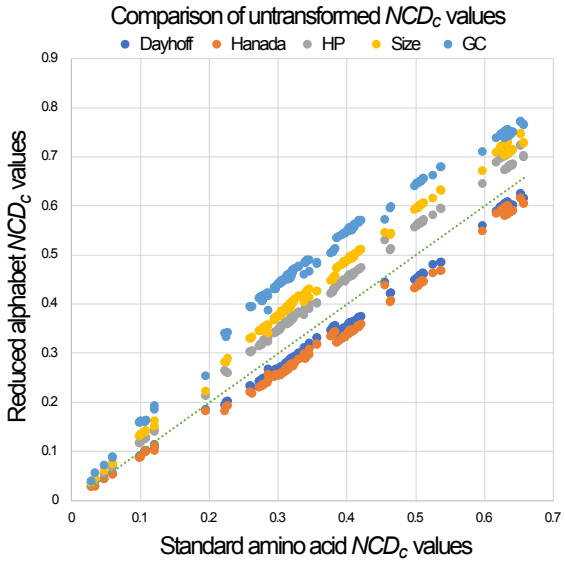

Figure 4: Different amino acid alphabets capture distinct information about evolutionary divergence. (A) Uncorrected NCD_c values for each alphabet plotted against NCD_c for the standard alphabet. The dashed green line has a slope of one.

3.3 Compression distances exhibit similar behaviors with mammal and bird proteins

This overall pattern of increase for NCD_c observed for avian proteins (see github) was essentially identical to that observed for mammalian proteins (Fig. 4). This is not especially surprising given that models of protein sequence evolution estimated from aligned data are very similar for birds and mammals [32]. However, avian coding regions are known to exhibit a high degree of variation in GC-content [5, 6, 18, 34]. Moreover, this variation is correlated, at least to some degree, with evolutionary rate [5, 18]. This prompted us to compare uncorrected NCD_c values for the standard amino acid alphabet and the GC alphabet to patristic distances for the ML tree of birds.

The concatenated ML tree for the avian dataset exhibited substantial branch length heterogeneity (Fig. 6A) . The deepest divergence

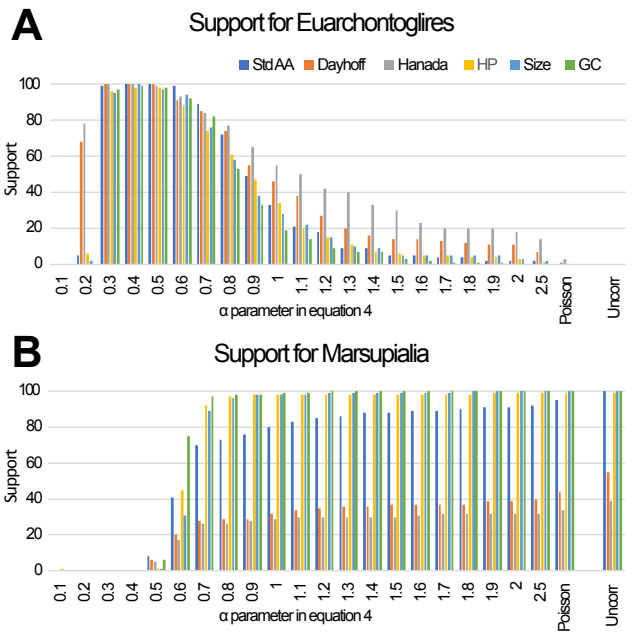

Figure 5: Support for selected clades in phylogenetic analyses of NCD_c values. (A) Support for Euarchontoglires (rodents+primates). (B) Support for marsupial monophyly. Both graphs show different transformations of the NCD_c as sets of columns. Column colors indicate the alphabet. Each set of columns indicates the transformation used to correct the NCD_c; Uncorr indicates no transformation.

within Neoaves (the clade comprising most extant birds) was between Strisores (hummingbirds, nightjars, and allies [36]) and all other Neoaves in this tree. For this taxon sample, the correct root of Neoaves is likely to be between Mirandornithes (flamingos and grebes [36]) and other Neoaves [5, 6, 18]. This conflict was not completely unexpected; several studies [6, 34] have noted that analyses of coding data often place Strisores sister to all other Neoaves (e.g., the coding exon trees in Jarvis et al. [18] and the primary tree in Prum et al. [33]). Perhaps surprisingly, given the presumed role of GC-content variation across taxa in biased estimation of the bird tree [5, 6, 18, 34], we observed very similar patterns of increase for NCD_c values calculated using the standard and GC alphabets relative to patristic distances (Fig. 6B). Although the two alphabets clearly extract different information from the data, the most noticeable difference between the alphabets was the higher values for the GC alphabet, which was expected based on the mammal data (Fig. 4). A recent study showed that the patterns of base composition change across the bird tree are more complex than expected for simple shifts in GC content [2]; it might be more difficult to detect those complex changes using the relatively simple GC alphabet.

Estimates of the bird tree based on transformed NCD_c values had limited support regardless of the value of α. However, we note the ML concatenation tree for birds exhibits several differences from the likely true species tree (Fig. 6A). In addition to the unexpected

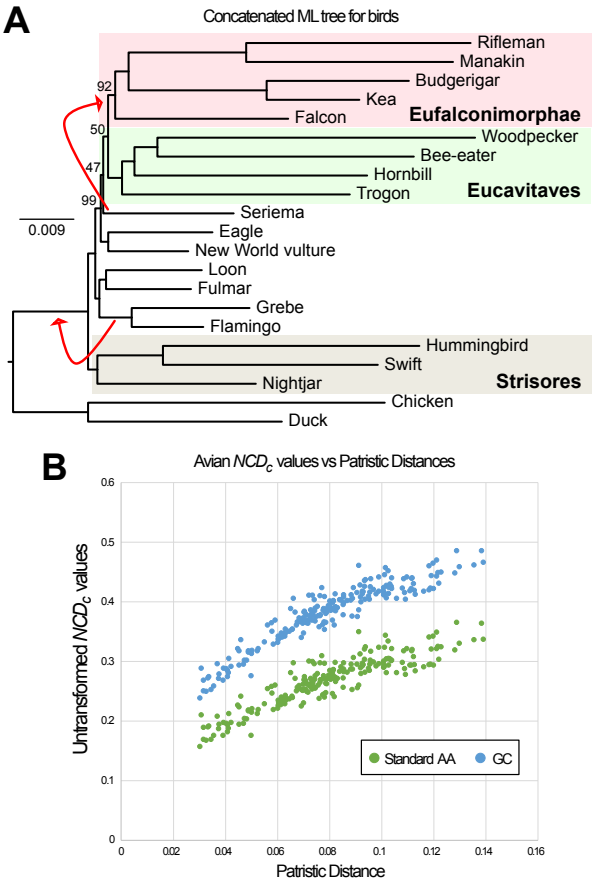

Figure 6: NCD_c values for the standard amino acid alphabet and the GC alphabet exhibit similar patterns of increase in birds. (A) Concatenated ML tree for birds used to calculate patristic distances. This topology was obtained using IQ-TREE topology with the best-fitting model (Q.bird+F+I+G4 [27]) identified using the `-m TEST` option [21]. Red arrows indicate rearrangements relative to the likely true tree; moving the relevant clades to the positions indicated using the arrows would yield the best available estimate of the true tree (see Braun [5] for review). Avian clade names are from Sangster et al. [36]. Unlabeled nodes had complete support. (B) Uncorrected NCD_c values the standard and GC alphabets calculated for birds and compared to patristic distances.

position of Strisores, the seriema is also misplaced relative to expectation (and that expectation is based on extensive data; see Braun [5] for review). Thus, neither ML concatenation nor analyses of compression distances were able to provide accurate estimates of phylogeny for birds. The part of the bird tree captured in this taxon sample is very challenging, with clear branch length differences, base compositional variation, and pervasive ILS (and potentially other sources of bias) [2, 5, 6, 17, 18, 34]. Thus, it is unsurprising that modified *NCD* trees for birds were inaccurate.

4 CONCLUSIONS

This study provided empirical evidence that phylogenetic analyses using unmodified compression distances (i.e., the *NCD*) are susceptible to LBA. However,, it also revealed that there are straightforward ways to transform the *NCD* to correct this problem; specifically, the NCD_Γ can be used to incorporate the potential for multiple that correct can this problem. This study also corroborated the hypothesis that calculating compression distances using alternative alphabets can reveal distinct signals in the data; using different alphabets affected support (Fig. 5). Although the transformation to incorporate among-sites substitution rate variation (i.e., the NCD_Γ) and the use of alternative alphabets both represent ways to incorporate biological information into compression distances, the Γ transformation is likely to be more important based on these analyses. Finally, the approach used in this study made it possible to generate support values similar to the bootstrap, which is a major benefit relative to other methods to calculate support that can be used with compression distance analyses (e.g., Li et al. [23]). This study did not show that analyses of compression distances (or analyses of any distances) were more robust to ILS than concatenated ML. However, overcoming LBA is a prerequisite for any useful phylogenomic method and that was the focus of this study. The demonstration that compression distances can be improved by adding biological information represents an important step in the development of those distances as useful phylogenomic tools.

ACKNOWLEDGMENTS

I am grateful to John Gatesy for helpful discussions regarding mammalian phylogeny. E.L.B. was supported by grant DEB-1655683 from the U.S. National Science Foundation.

REFERENCES

[1] Elizabeth S. Allman, Colby Long, and John A. Rhodes. 2019. Species Tree Inference from Genomic Sequences Using the Log-Det Distance. *SIAM Journal on Applied Algebra and Geometry* 3, 1 (Jan. 2019), 107–127. https://doi.org/10.1137/18m1194134

[2] Jacob S. Berv, Sonal Singhal, Daniel J. Field, Nathanael Walker-Hale, Sean W. McHugh, J. Ryan Shipley, Eliot T. Miller, Rebecca T. Kimball, Edward L. Braun, Alex Dornburg, C. Tomomi Parins-Fukuchi, Richard O. Prum, Benjamin M. Winger, Matt Friedman, and Stephen A. Smith. 2022. Molecular early burst associated with the diversification of birds at the K–Pg boundary. *bioRxiv* (Oct. 2022), 2022.10.21.513146. https://doi.org/10.1101/2022.10.21.513146

[3] Alix Boc, Alpha Boubacar Diallo, and Vladimir Makarenkov. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research* 40, W1 (June 2012), W573–W579. https://doi.org/10.1093/nar/gks485

[4] Edward L Braun. 2018. An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins. *Bioinformatics* 34, 13 (June 2018), i350–i356. https://doi.org/10.1093/bioinformatics/bty261

[5] Edward L. Braun, Joel Cracraft, and Peter Houde. 2019. Resolving the Avian Tree of Life from Top to Bottom: The Promise and Potential Boundaries of the Phylogenomic Era. In *Avian Genomics in Ecology and Evolution*, Robert H. S. Kraus (Ed.). Springer International Publishing, 151–210. https://doi.org/10.1007/978-3-030-16477-5_6

[6] Edward L. Braun and Rebecca T. Kimball. 2021. Data Types and the Phylogeny of Neoaves. *Birds* 2, 1 (Jan. 2021), 1–22. https://doi.org/10.3390/birds2010001

[7] R. Cilibrasi and P.M.B. Vitanyi. 2005. Clustering by Compression. *IEEE Transactions on Information Theory* 51, 4 (April 2005), 1523–1545. https://doi.org/10.1109/tit.2005.844059

[8] Gautam Dasarathy, Robert Nowak, and Sebastien Roch. 2015. Data Requirement for Phylogenetic Inference from Multiple Loci: A New Distance Method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12, 2 (March 2015), 422–432. https://doi.org/10.1109/tcbb.2014.2361685

[9] M. O. Dayhoff, R. V. Eck, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure, vol. 5*, Margaret O. Dayhoff

(Ed.). National Biomedical Research Foundation, Silver Springs, MD, 345–352.
[10] Julian Echave, Stephanie J. Spielman, and Claus O. Wilke. 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics* 17, 2 (Jan. 2016), 109–121. https://doi.org/10.1038/nrg.2015.18
[11] Scott V. Edwards. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1 (Jan. 2009), 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x
[12] Joseph Felsenstein. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Zoology* 27, 4 (Dec. 1978), 401–410. https://doi.org/10.1093/sysbio/27.4.401
[13] Nicole M. Foley, Victor C. Mason, Andrew J. Harris, Kevin R. Bredemeyer, Joana Damas, Harris A. Lewin, Eduardo Eizirik, John Gatesy, Elinor K. Karlsson, Kerstin Lindblad-Toh, Zoonomia Consortium, Mark S. Springer, and William J. Murphy. 2023. A genomic timescale for placental mammal evolution. *Science* 380, 6643 (April 2023). https://doi.org/10.1126/science.abl8189
[14] Kousuke Hanada, Shin-Han Shiu, and Wen-Hsiung Li. 2007. The Nonsynonymous/Synonymous Substitution Rate Ratio versus the Radical/Conservative Replacement Rate Ratio in the Evolution of Mammalian Genes. *Molecular Biology and Evolution* 24, 10 (July 2007), 2235–2241. https://doi.org/10.1093/molbev/msm152
[15] Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 35, 2 (Oct. 2017), 518–522. https://doi.org/10.1093/molbev/msx281
[16] Susan Holmes. 2003. Bootstrapping Phylogenetic Trees: Theory and Methods. *Statist. Sci.* 18, 2 (May 2003). https://doi.org/10.1214/ss/1063994979
[17] Peter Houde, Edward L. Braun, and Lawrence Zhou. 2020. Deep-Time Demographic Inference Suggests Ecological Release as Driver of Neoavian Adaptive Radiation. *Diversity* 12, 4 (April 2020), 164. https://doi.org/10.3390/d12040164
[18] Erich D. Jarvis, Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, Brant C. Faircloth, Benoit Nabholz, Jason T. Howard, and 95 additional coauthors. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346, 6215 (Dec. 2014), 1320–1331. https://doi.org/10.1126/science.1253451
[19] Erich D. Jarvis, Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, Brant C. Faircloth, Benoit Nabholz, Jason T. Howard, Alexander Suh, Claudia C. Weber, Rute R. da Fonseca, Alonzo Alfaro-Núñez, Nitish Narula, Liang Liu, Dave Burt, Hans Ellegren, Scott V. Edwards, Alexandros Stamatakis, David P. Mindell, Joel Cracraft, Edward L. Braun, Tandy Warnow, Wang Jun, M. Thomas Pius Gilbert, and Guojie Zhang. 2015. Phylogenomic analyses data of the avian phylogenomics project. *GigaScience* 4, 1 (Feb. 2015). https://doi.org/10.1186/s13742-014-0038-1
[20] Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, and Hervé Philippe. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics* 22, 4 (April 2006), 225–231. https://doi.org/10.1016/j.tig.2006.02.003
[21] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14, 6 (May 2017), 587–589. https://doi.org/10.1038/nmeth.4285
[22] Laura Salter Kubatko and James H. Degnan. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology* 56, 1 (Feb. 2007), 17–24. https://doi.org/10.1080/10635150601146041
[23] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17, 2 (Feb. 2001), 149–154. https://doi.org/10.1093/bioinformatics/17.2.149
[24] Ming Li, Xin Chen, Xin Li, Bin Ma, and P.M.B. Vitanyi. 2004. The Similarity Metric. *IEEE Transactions on Information Theory* 50, 12 (Dec. 2004), 3250–3264. https://doi.org/10.1109/tit.2004.838101
[25] Ming Li and Paul Vitányi. 2019. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer International Publishing. https://doi.org/10.1007/978-3-030-11298-1
[26] Kelly A. Meiklejohn, Brant C. Faircloth, Travis C. Glenn, Rebecca T. Kimball, and Edward L. Braun. 2016. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. *Systematic Biology* 65, 4 (Feb. 2016), 612–627. https://doi.org/10.1093/sysbio/syw014
[27] Bui Quang Minh, Cuong Cao Dang, Le Sy Vinh, and Robert Lanfear. 2021. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Systematic Biology* 70, 5 (Feb. 2021), 1046–1060. https://doi.org/10.1093/sysbio/syab010
[28] Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37, 5 (Feb. 2020), 1530–1534. https://doi.org/10.1093/molbev/msaa015
[29] Siavash Mirarab, Luay Nakhleh, and Tandy Warnow. 2021. Multispecies Coalescent: Theory and Applications in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 52, 1 (Nov. 2021), 247–268. https://doi.org/10.1146/annurev-ecolsys-012121-095340
[30] William J. Murphy, Nicole M. Foley, Kevin R. Bredemeyer, John Gatesy, and Mark S. Springer. 2021. Phylogenomics and the Genetic Architecture of the Placental Mammal Radiation. *Annual Review of Animal Biosciences* 9, 1 (Feb. 2021), 29–53. https://doi.org/10.1146/annurev-animal-061220-023149
[31] Masatoshi Nei and Jianzhi Zhang. 2006. Evolutionary Distance: Estimation. *Encyclopedia of Life Sciences* (July 2006), 1–4. https://doi.org/10.1038/npg.els.0005108
[32] Akanksha Pandey and Edward L. Braun. 2020. Protein evolution is structure dependent and non-homogeneous across the tree of life. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM. https://doi.org/10.1145/3388440.3412473
[33] Richard O. Prum, Jacob S. Berv, Alex Dornburg, Daniel J. Field, Jeffrey P. Townsend, Emily Moriarty Lemmon, and Alan R. Lemmon. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526, 7574 (Oct. 2015), 569–573. https://doi.org/10.1038/nature15697
[34] Sushma Reddy, Rebecca T. Kimball, Akanksha Pandey, Peter A. Hosner, Michael J. Braun, Shannon J. Hackett, Kin-Lan Han, John Harshman, Christopher J. Huddleston, Sarah Kingston, Ben D. Marks, Kathleen J. Miglia, William S. Moore, Frederick H. Sheldon, Christopher C. Witt, Tamaki Yuri, and Edward L. Braun. 2017. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. *Systematic Biology* 66, 5 (March 2017), 857–879. https://doi.org/10.1093/sysbio/syx041
[35] Sebastien Roch and Mike Steel. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology* 100 (March 2015), 56–62. https://doi.org/10.1016/j.tpb.2014.12.005
[36] George Sangster, Edward L. Braun, Ulf S. Johansson, Rebecca T. Kimball, Gerald Mayr, and Alexander Suh. 2022. Phylogenetic definitions for 25 higher-level clade names of birds. *Avian Research* 13 (2022), 100027. https://doi.org/10.1016/j.avrs.2022.100027
[37] Gabrielle E. Scolaro and Edward L. Braun. 2023. The Structure of Evolutionary Model Space for Proteins across the Tree of Life. *Biology* 12, 2 (Feb. 2023), 282. https://doi.org/10.3390/biology12020282
[38] Celine Scornavacca, Khalid Belkhir, Jimmy Lopez, Rémy Dernat, Frédéric Delsuc, Emmanuel J P Douzery, and Vincent Ranwez. 2019. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Molecular Biology and Evolution* 36, 4 (Jan. 2019), 861–862. https://doi.org/10.1093/molbev/msz015
[39] Gregory A. C. Singer and Donal A. Hickey. 2000. Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins. *Molecular Biology and Evolution* 17, 11 (Nov. 2000), 1581–1588. https://doi.org/10.1093/oxfordjournals.molbev.a026257
[40] Mark S. Springer and John Gatesy. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution* 94 (Jan. 2016), 1–33. https://doi.org/10.1016/j.ympev.2015.07.018
[41] Mark S. Springer and John Gatesy. 2017. On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity* 16, 3 (Dec. 2017), 210–228. https://doi.org/10.1080/14772000.2017.1401016
[42] Mike Steel. 2009. A basic limitation on inferring phylogenies by pairwise sequence comparisons. *Journal of Theoretical Biology* 256, 3 (Feb. 2009), 467–472. https://doi.org/10.1016/j.jtbi.2008.10.010
[43] Jeet Sukumaran and Mark T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26, 12 (April 2010), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228
[44] David L. Swofford. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), version 4. https://paup.phylosolutions.com/
[45] Deangelo Wilson and John D. Rogers. 2023. Evaluating Compression-Based Phylogeny Estimation in the Presence of Incomplete Lineage Sorting. *Journal of Computational Biology* 30, 3 (March 2023), 250–260. https://doi.org/10.1089/cmb.2022.0197
[46] Chao Zhang, Yiming Zhao, Edward L. Braun, and Siavash Mirarab. 2021. TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution* 12, 11 (Aug. 2021), 2145–2158. https://doi.org/10.1111/2041-210x.13696