Journal of Language Evolution, 2021, 1–6
https://doi.org/10.1093/jole/lzab007
Commentary

Correcting a bias in TIGER rates resulting from high amounts of invariant and singleton cognate sets

Johann-Mattis List *

Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

*E-mail: mattis.list@lingpy.org

Abstract

In a recent issue of the Journal of Language Evolution, Syrjänen et al. (2021) investigate the suitability of computing Cummins and McInerney's (2011) TIGER rates for estimating the tree-likeness of linguistic datasets compiled for phylogenetic reconstruction. The authors test the TIGER rates on a diverse sample of simulated data, which by and large confirms the usefulness of TIGER rates as an analytic tool for investigating linguistic data, but they test them only on one real-world dataset of Uralic languages which turns out to behave quite differently from the simulated data. When testing the TIGER rates on additional datasets, I detected a bias in the computation which leads to an unnatural increase in those cases where a dataset contains many characters with invariant or singleton states. To overcome this problem, I suggest a modified variant of TIGER rates, which is provided in the form of a freely available Python package. Testing the modified TIGER scores on the simulated data of Syrjänen et al. shows that the corrected TIGER rates still readily distinguish between different degrees of tree-likeness. Testing them on a dataset in which the number of singletons and invariants was artificially increased further shows that the corrected TIGER rates are not influenced by the bias. A final tests on seven linguistic datasets show the usefulness of the corrected TIGER rates on a larger variety of linguistic datasets and illustrate the importance to take specific aspects of linguistic data into account when using biological methods in the domain of language evolution.

Key words: phylogenetic reconstruction; tree-likeness; cognate sets; wordlists; phylogenetic characters

1. Introduction

When I saw the recent study by Syrjänen et al. (2021), in which the authors presented how the TIGER scores—a way to score the compatibility of phylogenetic characters in order to assess their tree-likeness originally introduced by Cummins and McInerney (2011)—can be applied to linguistic data, I was very intrigued by this approach, since it was very straightforward both regarding the conceptualization and the implementation. Assuming that the computation of TIGER scores would be very useful for my own work, I immediately checked how they could be applied to additional linguistic datasets, which we had compiled in past research that aimed at unifying cross-linguistic datasets (List et al. 2021).

Since the implementation of the authors did not offer the possibility to analyse a given dataset directly from within a Python script, I decided to follow the instructions by both Syrjänen et al. (2021) and Cummins and

McInerney (2011) to write a very short Python package that would allow to compute the TIGER scores and also provide full coverage with respect to unit tests. This package, which is curated on GitHub at https://github.com/pylogeny/tiger and archived with PyPi (https://pypi.org/projects/pylotiger) in Version 1.0 and Zenodo (https://doi.org/10.5281/zenodo.5812242) is freely available and published under a permissive license and was tested to yield identical scores with the implementation by Syrjänen et al. (2021).

When testing this package on additional datasets, however, I realized that the original TIGER scores show a certain bias that makes them very vulnerable when cognate sets are sparsely or densely distributed across a given set of words. TIGER scores estimate how well the multi-states of one character conform in their structure with the multi-states in another character by comparing the number of sets of identical character states in the second character that are contained in the sets of identical character states in the first character. Since no pruning of characters is carried out before computing the TIGER scores, the current computation yields a direct bias towards those cases in which one character is represented by a single state across all taxonomic units (so-called 'invariants') and towards those cases in which all taxonomic units show distinct character states for a given character (so-called 'singletons'). Both invariants and singletons are classical examples for parsimony-uninformative characters, and while it makes sense to include them in Bayesian phylogenetic analyses, their role is at least dubious when it comes to computing the tree-likeness of the characters in a dataset.

2. Original TIGER rates

As a concrete example, consider the case of five language varieties A, B, C, D, and E, which have all the same character state *a* for a character X and states *a*, *b*, *c*, *d*, and *e* for a character Y. In order to compare the TIGER scores of character X against character Y, we first determine the sets of language varieties (so-called *set partitions*) defined by both characters. The character X yields one set {A, B, C, D, E} while character Y yields five sets containing one variety each {A} {B} {C} {D} {E}. If we now compute the *partition agreement score*, which is the crucial component of the TIGER scores, we iterate over each partition in character Y and check if this partition is contained in any partition in our character X. Since the set partitions for character Y consists of five sets with one variety each, while character X consists of one sole partition, all five set partitions appear as subsets of the set partition in character X. The resulting partition agreement score is therefore 1, since we count the number of partitions that appear as subsets (including identical partitions) and divide it by the total number of partitions in character Y, which yields $5/5 = 1$. If we compare character Y against character Z, however, we can see that there is only one partition in character X which does not appear as subset of any partition in character Y, and as a result, the partition agreement score will be $0/1 = 0$.

Adding a third character Z with states *a*, *a*, *a*, *b*, and *b* for our five varieties shows that the scores of 1 and 0 are no coincidence. In *all* cases in which we compare *any* character *against* our character Y, we will receive a partition agreement score of 0 (which would indicate full tree-likeness). In the same way, in *all* cases in which we compare our character X *against any* other character, we will receive a score of 0 as well. As a result, any dataset that contains many invariant or singleton characters will necessarily give the impression of having a high number of compatible characters that suggest a high tree-likeness of the underlying data. Since neither character X nor character Y yield any phylogenetically interesting information, this behaviour of the TIGER scores points to a bias in those cases where the *density* of cognates is either high or low. If there are many singletons in a dataset, TIGER scores will increase *overall*, since every other character in the data is compatible with these singletons. Similarly, if there are many invariant cognate sets, they will be compatible with all other characters and add a 'bubble' on the top of the violin plots that Syrjänen et al. (2021) used to visualize the distributions of the TIGER rates.

An additional bias that we can observe in the TIGER rates are those cases in which we have partially overlapping set partitions. As an example, consider the character L with states *a*, *a*, *a*, *b*, *b* and the character M with states *d*, *d*, *c*, *c*, *c*. The partition agreement score based on the TIGER rates would yield 0.5 for both cases, since we divide the number of sets in one partition (two in both cases) by the number of compatible characters (1 in both cases, state *b* when comparing L with M and state *d* when comparing M with L). This result seems unsatisfying, it would be the same if we compared a character J with states *e*, *e*, *e*, *f*, *g* with our character L, where we have again two sets in L and one of them being compatible with J, although intuitively, one would assume that the compatibility of J with L should be higher. The reason for this problem is that the TIGER scores do not take intersections in partitions into account. When comparing L with M, we find three sets in M which share at least one taxonomic unit with sets in L (state *d* in M with taxonomic units AB shares the units AB with state *a* in L with its taxonomic units ABC, state *c* in L with its taxonomic units ABC shares unit C with state *a* in M and units BC with state *b* in M). To count the compatibility of L compared with M, it would therefore be more consequent to divide by the number of sets in M that share at least one taxonomic unit with a set in L, which would be three in our case, yielding the score 0.33 (one-third) both for L compared with M and for M compared with L .

3. Corrected TIGER rates

We can avoid the bias by means of an extended, *corrected* calculation for the partition agreement score. This corrected partition agreement score accounts for the intersection bias mentioned in the previous paragraph and additionally ignores singleton and invariant cognate sets. Since characters can have states that are polymorphic, resulting from synonymous word forms, the correction for invariants and singletons is limited to characters with identical cognate sets for all words in a given meaning slot, but has to be calculated by checking the set partitions for each character, excluding either those which comprise only one taxonomic unit, or those which comprise the full set of taxa in a given sample. The corrected partition agreement scores are also available from the new Python implementation of the TIGER rates and can be invoked by changing the function that computes the partition agreement scores.

4. Experiments

4.1 Testing corrected TIGER rates on simulated data

The benefits of the corrected TIGER rates can be illustrated in three experiments. In the first experiment, we look at the simulated data provided by Syrjänen et al. (2021) and contrast the original TIGER rates with the corrected TIGER rates to see if the major discrimination between different data types (many borrowings, dialect chains, etc.) can be preserved. In order to do so, I first ran the code by Syrjänen et al. (2021) to compute the simulations for different types of phylogenies, ranging from a pure tree via different degrees of borrowings up to dialect chains, since the authors themselves have not shared the actual simulations in their supplement. I then computed the TIGER rates both in the original and the corrected version in order to check that the corrected TIGER rates preserve the rough discriminative function reported by Syrjänen et al. (2021).

Table 1. Comparing original and corrected TIGER rates on simulated data.

Datasets	TIGER	C-TIGER
pure_tree	0.80 ± 0.02	0.58 ± 0.06
borrowing_05	0.78 ± 0.02	0.50 ± 0.05
borrowing_10	0.76 ± 0.02	0.44 ± 0.05
borrowing_15	0.75 ± 0.01	0.40 ± 0.04
borrowing_20	0.73 ± 0.01	0.37 ± 0.04
Dialect	0.65 ± 0.02	0.19 ± 0.04
swamp	0.58 ± 0.01	0.07 ± 0.01

From the results in Table 1, we can see that the corrected TIGER rates preserve the distinctive function of the original TIGER rates, while at the same time leading to larger differences between the individual subsets. This experiment thus illustrates that the key function of the TIGER scores that Syrjänen et al. (2021) reported based on their simulated data is preserved with the corrected TIGER rates.

4.2 Testing corrected TIGER rates on artificially modified data

In the second experiment, we look at the degree by which TIGER rates and corrected TIGER rates can be influenced by singleton and invariant character states. Since—as I have shown before—singletons and invariants will both lead to an increase of TIGER rates, it is interesting to investigate how far this influence can go if we artificially increase the number of singletons and invariants in a dataset. In a first run, we thus test what happens if we turn a certain proportion of the characters in a linguistic datasets into singleton cognate sets. Using the Uralex data presented by Syrjänen et al. (2021), I designed an experiment which starts from 20% and then proceeds in steps of 20% until 80% is reached, turning the respective proportion of characters into singleton cognates and measuring the TIGER rates both in their original and their corrected version. For each run, 100 trials were carried out. The results are shown in Table 2. As can be seen from this table, the original TIGER values tend to increase as the amount of singletons in the data increases, while the corrected TIER values remain stable, although the number of valid characters that are considered in the computation shrinks (as expected).

We can test the influence of invariant cognate sets in the same way by systematically turning certain proportions of cognate sets in our data into invariants. The results for this test are shown in Table 3. As can be seen from the table, the invariants do not lead to an increase of the classical TIGER scores at first. On the opposite, the scores decrease at first, reaching their lowest point at a proportion of 0.6, before they grow again with

Table 2. Comparing original and corrected TIGER rates on data with artificially increased proportions of singletons.

Proportion	Characters	CS-Size	TIGER	C-TIGER
0	313/307	11.09 ± 5.18	0.68 ± 0.13	0.30 ± 0.20
0.2	313/246.00	13.75 ± 7.42	0.71 ± 0.13	0.30 ± 0.20
0.4	313/186.50	17.08 ± 8.10	0.76 ± 0.11	0.29 ± 0.20
0.6	313/121.50	19.63 ± 8.01	0.81 ± 0.10	0.29 ± 0.19
0.8	313/60.50	22.65 ± 6.49	0.87 ± 0.10	0.28 ± 0.19

Notes: Column 'Proportion' points to the amount of singletons added in the respective run. Column 'Characters' shows the number of characters considered, which is stable for the original TIGER rates but decreases when using the corrected ones. Column 'CS-Size' refers to the average size of the cognate sets in the data along with the standard deviation. Columns 'TIGER' and 'C-TIGER' provide the original and corrected TIGER rates along with the standard deviation.

Table 3. Comparing original and corrected TIGER rates on data with artificially increased proportions of invariants.

Proportion	Characters	CS-Size	TIGER	C-TIGER
0	313/308	11.73 ± 5.40	0.71 ± 0.09	0.27 ± 0.19
0.2	313/245.17	9.06 ± 6.17	0.60 ± 0.17	0.30 ± 0.20
0.4	313/184.42	7.04 ± 6.36	0.55 ± 0.24	0.30 ± 0.20
0.6	313/122.51	5.01 ± 5.92	0.53 ± 0.31	0.30 ± 0.20
0.8	313/61.87	3.04 ± 4.68	0.53 ± 0.36	0.30 ± 0.20

Note: For details on the columns, see Table 2.

proportion 0.8. The reason for this behaviour can be found in the fact that the Uralex data have already a large proportion of singleton character states, which are—of course—highly compatible among themselves, but will be successively deleted, when being replaced by invariants. This shows—what should not be surprising—that the influence of high amounts of singletons and invariants on the original TIGER scores also depends on the peculiarities of the overall distribution of cognate sets over concepts.

4.3 Testing corrected TIGER rates on linguistic data

In the last experiment, we look into 'real' linguistic data and the consequences of invariants and singletons for TIGER rates. For this purpose, I have compiled a collection of seven phylogenetic datasets of different cognate density, taken from the Lexibank repository (List et al. 2021), where lexical dataset suitable for phylogenetic analyses are collected and offered in the form of Cross-Linguistic Data Formats (CLDF), a format specification that increases the comparability and interoperability of cross-linguistic datasets (Forkel et al. 2018). These seven datasets include the Uralex data by Syrjänen et al. (2021), as well as data from Dravidian languages (Kolipakam et al. 2018), Mixe-Zoquean languages (Cysouw et al. 2006), Aslian languages (Dunn et al. 2013), Semitic languages (Feleke 2021), Japonic languages (Hattori 1973), and Palaungic languages (Deepadung et al. 2015). These datasets differ with respect to the number of languages, the number of concepts, and also with respect to the density of cognates, as reflected by the *diversity index* which I proposed earlier (List 2014, 188), and which divides the number of cognate sets minus the number of concepts by the number of words minus the number of concepts. For all datasets, I computed the classical TIGER rates, the corrected TIGER rates, the Delta Scores (Holland et al. 2002) in the implementation by Greenhill (2016), and the distribution of cognate set sizes.

The results of this test are shown in Table 4. As can be seen from this table, the corrected TIGER rates differ from the original TIGER rates in some important respects. While both rates rank the Uralic data highest with respect to tree-likeness, the Dravidian data receive a remarkably high score in the original TIGER rates, while the dataset is ranked last in the corrected TIGER rates, which corresponds well to the fact that the Dravidian data also receive the highest Delta scores in the sample.

In Fig. 1, the original and corrected TIGER rates were plotted along with the cognate set size distributions. As can be seen from this figure, comparing the high rates of singletons in the first five datasets and the extremely relatively high rates of invariants in Japonic and Palaungic, as shown in Table 4, finds a direct

Table 4. Comparing original and corrected TIGER rates on seven linguistic datasets.

Dataset	Languages	Concepts	Diversity	Chars	Singletons	Invariants	TIGER	C-TIGER	Delta
Uralic	27	313	0.37	313/307	0.64	0.00	0.68 ± 0.13	0.30 ± 0.20	0.17 ± 0.03
Dravidian	20	100	0.33	100/97	0.64	0.00	0.65 ± 0.10	0.20 ± 0.17	0.30 ± 0.04
Mixe-Zoquean	10	110	0.23	110/89	0.41	0.06	0.55 ± 0.24	0.28 ± 0.16	0.18 ± 0.03
Aslian	32	146	0.19	146/146	0.40	0.00	0.55 ± 0.10	0.20 ± 0.16	0.24 ± 0.02
Semitic	21	150	0.18	150/144	0.47	0.01	0.58 ± 0.13	0.21 ± 0.21	0.26 ± 0.03
Japonic	10	200	0.15	200/124	0.40	0.17	0.60 ± 0.32	0.26 ± 0.17	0.27 ± 0.07
Palaung	16	100	0.08	100/68	0.16	0.15	0.53 ± 0.34	0.28 ± 0.16	0.20 ± 0.02

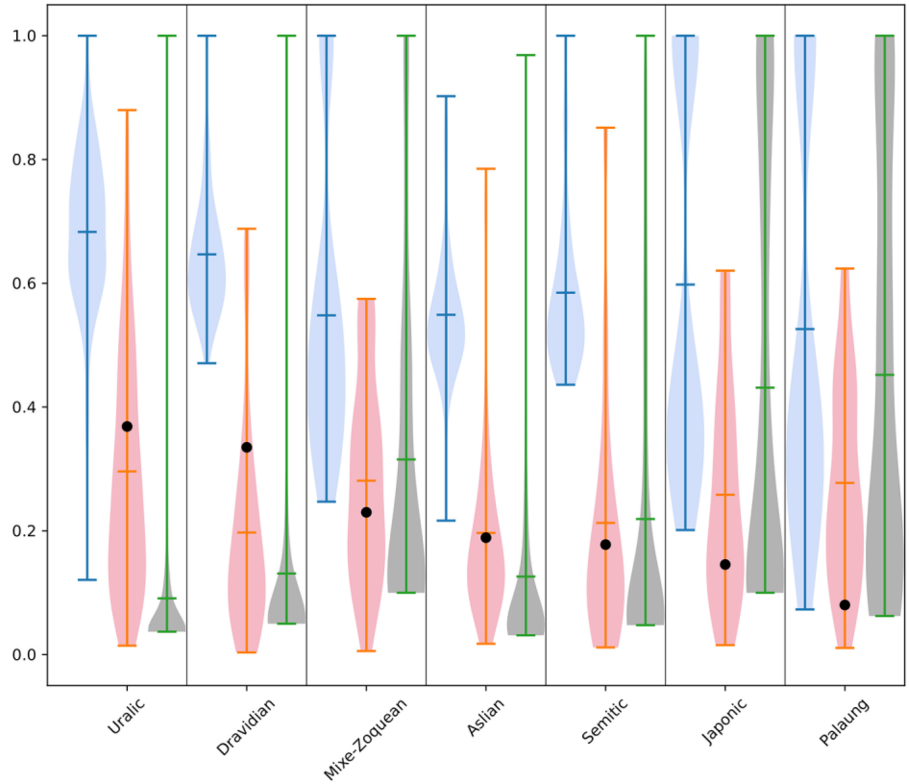

Figure 1. Comparing original and corrected TIGER rates on linguistic datasets. Blue violins reflect the original TIGER rates, red violins reflect the corrected TIGER rates, and dark gray violins reflect the proportion of cognate set sizes, which are calculated by dividing the size of a cognate set by the number of languages in the sample. Note that the minimal score for the cognate set size is 1/*n*, where *n* is the number of languages in the sample. For this reason, datasets with fewer languages have higher minimal values in the plots. The black dots reflect the calculated diversity indices described in the main text.

reflection in the distribution of the original TIGER rates, which is successfully handled in the corrected TIGER rates.

5. Summary

While the TIGER scores as presented by Syrjänen et al. (2021) already gave the impression of an intriguingly useful way to assess the tree-likeness of linguistic datasets, their value can even be increased more by correcting systematically for singletons and invariants. This does not mean that the last word on TIGER rates and other methods for assessing the tree-likeness of linguistic data has been spoken, and it is quite likely that scholars will find better solutions than the one proposed here in the future. What I think is important with respect to what I have outlined here is the role which the peculiarities of linguistic data play when applying methods originally designed for biological data in the linguistic domain. Since typical datasets in biology do not seem to suffer that much from singleton and invariant characters, it is quite likely that the TIGER rates as proposed by Cummins and McInerney (2011) are still the best choice. For the case of linguistics, however, I think my analyses show that—as long as no better methods will be proposed—the corrected TIGER rates are a more reliable choice to assess the tree-likeness of a given dataset.

Funding

This project was funded by the by the ERC Starting Grant 715618 “Computer Assisted Language Comparison” (https://digling.org/calc, 2017–2022).

Data Availability

Data and code needed to replicate this study have been archived with Zenodo at https://doi.org/10.5281/zenodo.5812242.

Supplementary material

The supplementary material accompanying this study consists in the new Python package for the computation of TIGER rates and corrected TIGER rates and the data and code needed to replicate the three experiments reported here. The material is curated on GitHub at https://github.com/pylogeny/tiger (Version 1.0.0) and archived with Zenodo (https://doi.org/10.5281/zenodo.5812242). The Python package has also been uploaded tothe Python Package Index at https://pypi.org/project/pylotiger/. The experiments can be found in the folder *examples* along with instructions for replication.

Conflict of interest statement. None declared.

References

Cummins, C. A., and McInerney, J. O. (2011) ‘A Method for Inferring the Rate of Evolution of Homologous Characters That Can Potentially Improve Phylogenetic Inference, Resolve Deep Divergence and Correct Systematic Biases’, *Systematic Biology*, 60/6: 833–44. https://doi.org/10.1093/sysbio/syr064.

Cysouw, M., Wichmann, S., and Kamholz, D. (2006) ‘A Critique of the Separation Base Method for Genealogical Subgrouping, with Data from Mixe-Zoquean’, *Journal of Quantitative Linguistics*, 13/2–3: 225–64. [Data on Zenodo: 10.5281/zenodo.5126948]

Deepadung, S., Buakaw, S., and Rattanapitak, A. (2015) ‘A Lexical Comparison of the Palaung Dialects Spoken in China, Myanmar, and Thailand’, *Mon-Khmer Studies*, 44: 19–38. [Data on Zenodo: https://doi.org/10.5281/zenodo.5121402]

Dunn, M., Kruspe, N., and Burenhult, N. (2013) ‘Time and Place in the Prehistory of the Asian Languages’, *Human Biology*, 85/1–3: 383–400. [Data on Zenodo: https://doi.org/10.5281/zenodo.5121613]

Feleke, T. L. (2021) ‘Ethiosemitic Languages: Classifications and Classification Determinants’, *Ampersand*, 8: 100074.https://doi.org/10.1016/j.amper.2021.100074. [Data on Zenodo: https://doi.org/10.5281/zenodo.5126691]

Forkel, R. et al. (2018) ‘Cross-Linguistic Data Formats, Advancing Data Sharing and Re-use in Comparative Linguistics’, *Scientific Data*, 5/180205: 180205–10.

Greenhill, S. J. (2016) ‘Phylogemetric: A Python Library for Calculating Phylogenetic Network Metrics’, *Journal of Open Source Software*. https://doi.org/10.21105/joss.00028.

Hattori, S. (1973) ‘Japanese Dialects’. In: Hoenigswald, H. M. and Langacre, R. H. (eds) *Diachronic, Areal and Typological Linguistics*, pp. 368–400. Current Trends in Linguistics 11. The Hague: Paris. [Data on Zenodo: https://doi.org/10.5281/zenodo.5126845]

Holland, B. R. et al. (2002) ‘δ Plots: A Tool for Analyzing Phylogenetic Distance Data’, *Molecular Biology and Evolution*, 19/12: 2051–9. https://doi.org/10.1093/oxfordjournals.molbev.a004030.

Kolipakam, V. et al. (2018) ‘A Bayesian Phylogenetic Study of the Dravidian Language Family’, *Royal Society Open Science*, 5/3: 171504–17. [Data on Zenodo: https://doi.org/10.5281/zenodo.5121580]

List, J.-M. (2014) *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

—— et al. (2021) ‘Lexibank: A Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features [Preprint, Version 1]’. Research Square. https://doi.org/10.21203/rs.3.rs-870835/v1

Syrjänen, K. et al. (2021) ‘Crouching TIGER, Hidden Structure: Exploring the Nature of Linguistic Data Using TIGER Values’, *Journal of Language Evolution*, 6/2: 99–118. https://doi.org/10.1093/jole/lzab004. [Data on Zenodo: https://doi.org/10.5281/zenodo.4777568]