A Likelihood Ratio Test of Genetic Relationship among Languages

V.S.D.S. Mahesh Akavarapu and **Arnab Bhattacharya**
Dept. of Computer Science and Engineering
Indian Institute of Technology Kanpur
maheshak@cse.iitk.ac.in, arnabb@cse.iitk.ac.in

Abstract

Lexical resemblances among a group of languages indicate that the languages could be genetically related, i.e., they could have descended from a common ancestral language. However, such resemblances can arise by chance and, hence, need not always imply an underlying genetic relationship. Many tests of significance based on permutation of wordlists and word similarity measures appeared in the past to determine the statistical significance of such relationships. We demonstrate that although existing tests may work well for bilateral comparisons, i.e., on pairs of languages, they are either infeasible by design or are prone to yield false positives when applied to groups of languages or language families. To this end, inspired by molecular phylogenetics, we propose a likelihood ratio test to determine if given languages are related based on the proportion of invariant character sites in the aligned wordlists applied during tree inference. Further, we evaluate some language families and show that the proposed test solves the problem of false positives. Finally, we demonstrate that the test supports the existence of macro language families such as Nostratic and Macro-Mayan.

1 Introduction

Languages that descend from a common ancestral language are termed to be *genetically related*. The existence of lexical resemblances between the two languages is a preliminary indication that they could be related. Such resembling lexicons that truly have a common origin are called *cognates*. For instance, Sanskrit *nāma* and English *name* are cognates that can be traced to Proto-Indo-European *$*h_3$nómn*. However, such resemblances can also occur out of sheer chance. For instance, Persian *bad* and *behtar* accidentally resemble English *bad* and *better* respectively, but are not true cognates[1]. Hence, it is necessary to show *statistical significance* on any appropriate measure that captures the lexical relatedness before arguing for a genetic relationship among any group of languages or language families (Campbell, 2013).

Several significance tests appeared in the past to address this problem, with the majority of them based on permutation tests, starting from Oswalt (1970). Given wordlists of a group of languages to be evaluated for a genetic relationship, these tests obtain the null distribution of a certain measure capturing similarity between word pairs by random permutations of the wordlists. Such tests either act *bilaterally*, i.e., on a pair of languages or proto-languages, or *multilaterally* on a group of languages. Among these, the multilateral comparison, which was made famous by Greenberg (1963, 1971, 1987, 2000) in traditional historical linguistics, has been a subject of much criticism (Poser and Campbell, 2008). Hence, the preferred way of comparing two language families has been to compare their reconstructed proto-forms bilaterally. However, Greenberg (2005) argues that genetic classification should precede proto-language reconstruction. Moreover, there is often a lack of agreement on reconstructed proto-forms both in terms of phonology and semantics which gives room for sufficient manipulation of wordlists that can in turn alter the results of significance tests (Kessler, 2015). Further, we demonstrate that multilateral permutation tests (Kessler and Lehtonen, 2006; Kessler, 2007) yield false negatives even after incorporating complex word similarity metrics such as SCA and LexStat (List, 2010, 2012).

To overcome these issues, we turn to *phylogenetic analysis* (Wiley and Lieberman, 2011) that is known to approximately capture the ancestral states and has been applied to phonological reconstruction tasks such as proto-language and cognate

[1]Persian *bad* is of uncertain origin while *behtar* ultimately derives from PIE *$*h_1$wésus*. On the other hand, English *better* derives from PIE *$*b^h$edrós* and is cognate with Sanskrit *bhadrá*

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2559–2570

reflex prediction tasks (Jäger, 2019, 2022) with reasonably good results. Specifically, we propose a *likelihood ratio test* (LRT) where we expect the difference in likelihoods of the best trees under null and alternate hypotheses to capture genetic relatedness. The null hypothesis assumes negligible proportion of invariant sites while the alternate hypothesis assumes significant proportion of invariant sites. Intuitively, related languages should have more positions where a character or a sound class is invariant than unrelated languages. Hence, we essentially capture the notion of relatedness as possessing a relatively high proportion of invariant sites. Further in this test, reconstructed proto-forms are not required and at the same time, the evolutionary tree structure is strictly imposed by design, unlike the multilateral model, thereby effectively circumventing the aforementioned methodological problems. Although inspired by similar tests from molecular phylogenetics, the test we propose is novel in the sense that the problem of testing common descent never arises in biology since monogenesis is accepted as a fact therein (Kessler, 2008). We further evaluate the test on various language families and demonstrate that the test does not misclassify unrelated languages as related.

We finally show that the test supports the existence of the macro-families Nostratic (Bomhard and Kerns, 1994) and Macro-Mayan (Campbell, 1997). While such an attempt to justify the existence of macro-families using bootstrap analysis of distance-based phylogeny is found in Jäger (2015), expressing statistical significance in terms of likelihood ratio is preferred over bootstrap support values whose interpretation is debated in molecular phylogenetics (Anisimova and Gascuel, 2006).

Our contributions are summarized as follows.

- We have proposed a *likelihood ratio test* to determine the *genetic relatedness* of a group of languages based on *invariant site proportions*.
- We have demonstrated by applying various language sets that the test does not exhibit the problem of false positives nor requires reconstructed proto-forms, unlike the previously proposed tests.
- We have found through the test some supporting evidence for the existence of macro-families namely Nostratic and Macro-Mayan

The rest of the paper is summarized as follows. Related work is discussed in §2. The methodology of the test is presented in §3. Evaluation details such as datasets and details of previous methods and variants are discussed in §4. The results are discussed in §5. The application of the method on long-range comparisons is discussed in §6. The paper is concluded in §7.

2 Related Work

Permutation test for bilateral language relationship comparisons was introduced by Oswalt (1970). The significance of sound correspondences by brute force probability calculation was proposed by Ringe (1992, 1996). This approach was however criticized for not being able to show significance for known related pairs of languages like Latin-English and also for accounting phonologically implausible sound correspondences (Kessler, 2001). Multilateral permutation tests were proposed by (Kessler and Lehtonen, 2006; Kessler, 2007). Several applications of permutation tests exist such as (Turchin et al., 2010; Kassian et al., 2015).

Some notable likelihood ratio tests in molecular phylogenetics, mostly on topologies, include (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996; Goldman et al., 2000; Anisimova and Gascuel, 2006) where bootstrap analysis is argued to be not so optimal to establish statistical significance on phylogenies. Otherwise, support for macro-families through bootstrap analysis for distance-based trees is shown in Jäger (2015). Comparisons of various methods of phylogenetic reconstruction such as distance-based and binary-character-based are given by Jäger (2018). Sound-class character-based phylogenetic analysis is found in (Jäger, 2019, 2022). Usually, Bayesian phylogenetic inference on binary cognate encodings gives good results (Rama et al., 2018; Rama and List, 2019).

Although the likelihood ratio metric is common for both past and present-day language models, the utility of this test using invariant sites outside computational historical linguistics is unknown.

3 Methodology

The key concept revolves around the idea that any hypothesis, in this case, a hypothesis on a phylogeny, is preferred over a competing null hypothesis if it is significantly more likely, i.e., has a higher likelihood than the latter. Given the wordlist data encoded as an aligned character matrix, related languages are expected to have a higher number of *invariant* columns. Thus, our null hypothesis con-

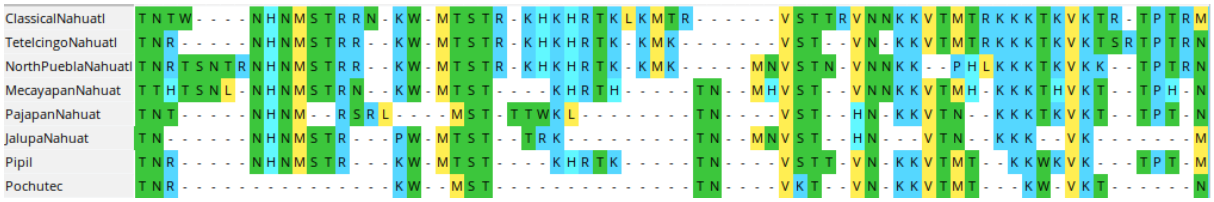


Figure 1: A section of character matrix for Uto-Aztecan family consisting of concatenated Multiple Sequence Alignments (MSAs) of consonant classes, one from each concept

sists of a phylogeny with a small proportion (fixed at 1%) of invariant sites, whereas the alternative hypothesis consists of a phylogeny with a larger but reasonable proportion (fixed at 6%) of invariant sites. The observed difference in their likelihood of real data is compared with that of data simulated from the null hypothesis through parametric bootstrapping and, accordingly, one of the hypotheses is rejected. The steps are elaborated next.

3.1 Character Matrix

The wordlists of a given group of languages, as mentioned previously, are encoded in the form of a *character matrix*. It consists of concatenated aligned words per concept, i.e., meaning. Thus, each row represents a language or *taxon*, and each column, also referred to as *site* in this paper, consists of phoneme classes, e.g., Dolgopolsky classes. Formally, let the input language set be $\{L_1, \ldots, L_m\}$, whose genetic relatedness is to be verified statistically. Let there be n concepts $C_1, \ldots, C_n$ in the wordlists. Each language L_i should have for each concept C_j a single word, say w_{ij}. If a language has multiple words for a single semantic slot, only the one with fundamental or core meaning is retained, following the recipe by Kessler (2001). For instance, if the meaning 'dull' has words *dull* and *unsharp*, *dull* is of core or fundamental meaning. Another example would be for the meaning 'belly', Latin *venter* is more fundamental than *abdōmen*. If it so happens that it still remains unresolved after this step, a single word is randomly picked up. In case a language has no word for a semantic slot, it is represented as a gap '–'. For each concept C_j and alphabet set $\mathbb{A}$, let $W^j \in \mathbb{A}^{m \times l_j}$ represent a multiple sequence alignment (MSA) of words where l_j is the length or the number of phonemes with vowels removed[2] in each word. The final character matrix $X \in \mathbb{A}^{m \times N}$ is concatenation of W^j, i.e., $[W^1 \ldots W^n]$ across columns and $N = \sum_{j=1}^{n} l_j$.

Greek_Anc	K	R	-	S	
Latin	K	R	N	-	-
English	H	R	N	-	-
Sanskrit	S	R	N	K	-

Table 1: Example of a Multiple Sequence Alignment (MSA) of consonant classes for a single concept 'horn'.

For example, consider a cognate set meaning 'horn' from a few Indo-European languages namely, Ancient Greek *keras*, Latin *cornu*, English *horn*, and Sanskrit *śṛṅga*. The resultant character matrix for this single meaning is a multiple sequence alignment with vowels removed and consonants encoded as Dolgopolsky classes as illustrated in Table 1. The final character matrix is the concatenation of such matrices across all the concepts. For an illustration of a final character matrix, see Figure 1, which is generated by MEGA11 (Tamura et al., 2021). In general, multiple sequence alignment is a fundamental step in several state-of-the-art methods in computational historical linguistics (Akavarapu and Bhattacharya, 2023, 2024).

3.2 Substitution Model

A *substitution model* describes the evolution of a character at a site assuming a Markovian process. Various substitution models have been described for various alphabets such as nucleotides, amino acids, etc. In this paper, we assume the simplest possible model where substitution rates are assumed to be equal between all the pairs of distinct characters. The resultant model is known as the Jukes-Cantor model (Jukes et al., 1969) in case of nucleotide substitutions and as Poisson (Bishop and Friday, 1987) in case of amino-acid substitutions. Formally, let the number of characters in the alphabet $\mathbb{A}$ be N. An element q_{ij} of the rate matrix Q, which denotes the rate at which character i

[2]Since the root form CVC is universal, including vowels results in spurious relationships. Further, languages of Caucasus like Georgian are rich in consonant clusters and, as a result, comparing them to others becomes difficult when vowels are considered.

mutates to character j is defined as follows:

$$q_{ij} = \mu \cdot \pi_i \, , i \neq j \text{ (equal rates)} \quad (1)$$

where π_i denotes the frequency of character i at the site and μ is the rate of mutation. The diagonal element should satisfy the normalization constraint:

$$q_{ii} = -\sum_{j \neq i} q_{ij} \quad (2)$$

The probability of transition $i \rightarrow j$ in time t is given by the matrix $P(t) = \{p_{ij}\} = e^{Qt}$. Likelihood of an evolutionary tree with topology T can be, thus, calculated from the substitution matrix where branch lengths V would denote the time.

3.3 Maximum Likelihood Tree (ML-tree)

For any phylogenetic tree with topology T, branch lengths V, other parameters such as shape parameter of heterogeneous rate, the proportion of invariant sites denoted by Θ, and with the observed data i.e., character matrix X, the *likelihood* is defined as the product of likelihoods at each site as given by the following equation, assuming independence for simplicity:

$$\mathcal{L}(T, V, \Theta | X) = \prod_{i=1}^{N} P(X_i | T, V, \Theta) \quad (3)$$

The site independence assumption also restricts the number of parameters. Given the limited amount of data, which is restricted to 100-200 wordlists, this is, thus, more suitable. Complex models such as bigram-based ones may be employed if sufficient data is available.

The parameters that maximize the likelihood, $\hat{T}, \hat{V}$, and $\hat{\Theta}$, define the *maximum likelihood tree* which is usually obtained by heuristic search in the parameter space. Typically, a tree is initialized either randomly or by some heuristic means, and from there, the tree space is explored through tree modifying operations to get the "best" tree. For a given tree, likelihood is computed using the well-known Felsenstein's pruning algorithm from phylogenetics (Felsenstein, 1973, 1981).

3.4 Invariant Sites

Invariant sites are those sites that are constant or evolve very slowly. These can be estimated through a maximum likelihood search along with other parameters. The proportion of invariant sites, P_{inv} may be known beforehand or estimated. Given the invariant sites, the likelihood defined in §3.3 is only the product of likelihoods across the variant sites.

Our observation is that estimated P_{inv} is higher ($>$0.06) among related languages while lower ($\approx$0.01) among (possibly) unrelated languages. Based on this observation and preliminaries, we now describe the likelihood ratio test.

3.5 Likelihood Ratio Test (LRT)

Given a null hypothesis H_0 and a competing alternative hypothesis H_a, the latter is preferred if it is more likely than the former i.e., $\mathcal{L}_{H_a} > \mathcal{L}_{H_0}$. In our case, the hypotheses consist of respective phylogenetic tree parameters estimated for ML-trees, i.e., H_0 consists of $\hat{T}_0, \hat{V}_0, \hat{\Theta}_0$ and H_a consists of $\hat{T}_a, \hat{V}_a, \hat{\Theta}_a$. The likelihood ratio test defines the following metric to decide whether to reject the null hypothesis:

$$\delta = 2 \cdot \ln \left(\frac{\mathcal{L}(\hat{T}_a, \hat{V}_a, \hat{\Theta}_a)}{\mathcal{L}(\hat{T}_0, \hat{V}_0, \hat{\Theta}_0)} \right) \quad (4)$$

The *Likelihood Ratio Test* (LRT) metric δ was shown to asymptotically follow a chi-squared distribution when the null hypothesis is assumed with the degrees of freedom $p - q$, where p and q respectively are the numbers of free parameters in the alternate and the null hypotheses (Wilks, 1938). However, it was argued that this may not hold in general for phylogenetic problems due to the discrete nature of tree topology (see (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996; Anisimova and Gascuel, 2006) for relevant work). As a result, the distribution of δ is determined by a parametric bootstrapping method where it is measured on the data simulated by the parameters estimated assuming the null hypothesis H_0 to hold, i.e, using the parameters $\hat{T}_0$, $\hat{V}_0$ and $\hat{\Theta}_0$.

As mentioned in §3.4, we propose LRT to test the relatedness of a group of languages using varying proportions of invariant sites. In other words words the null hypothesis H_0 consists of invariant site proportion P^0_{inv} and alternate hypothesis H_a consists of P^a_{inv} where $P^0_{inv} < P^a_{inv}$ as per the observations discussed in §3.4.

The typical way of obtaining the distribution for δ under H_0 involves finding the parameters $\{\hat{T}_0, \hat{V}_0, \hat{\Theta}_0\}$ and $\{\hat{T}_a, \hat{V}_a, \hat{\Theta}_a\}$ for the best trees respectively under H_0 and H_a along with observed δ, say $\hat{\delta}$. Further, several, say k, bootstrap replicates are generated from the topology, branch lengths, and other parameters defined by $\{\hat{T}_0, \hat{V}_0, \hat{\Theta}_0\}$, i.e.,

Family	Abbrv.	Languages	Concepts	Words
Afrasian	AfA	21	39	770
Dravidian	Drav	4	183	716
Indo-European	IE	12	185	2209
Kartvelian	Kart	1	180	180
Lolo-Burmese	LoBur	15	39	565
Mayan	May	30	94	2667
Mixe-Zoque	MZ	10	94	905
Mon-Khmer	MKh	9	199	1701
Mon-Khmer	MKh	16	94	1332
Munda	Mun	4	199	759
Uto-Aztecan	UAz	9	94	803

Table 2: Language families considered in this study.

assuming H_0. Next, the maximum likelihood search is run again on these replicates to obtain several samples for δ, say $\{\delta_1, \ldots, \delta_k\}$. However, we found considerable variation in $\hat{\delta}$, since the maximum likelihood search is only a heuristic and is affected by initialization. As a result, we obtain several samples for $\hat{\delta}$, say $\{\hat{\delta}_1, \ldots, \hat{\delta}_k\}$ by running the search k times and based on the null parameters, a single bootstrap replicate is generated for each search to consequently obtain $\{\delta_1, \ldots, \delta_k\}$ for corresponding k searches. Finally the *p-value* for $\mathbb{E}[\delta] < \mathbb{E}[\hat{\delta}]$ is obtained by one-sided paired t-test. If the p-value is less than a threshold (usually 0.05), we conclude that H_a may hold or, in other words, there are at least P^a_{inv} proportions of sites that are significantly invariant and, thus, the languages under consideration are likely to be related.

4 Experimental Setup

The section discusses the details of the experiments including datasets, baseline models, and implementation details.

4.1 Datasets

The data for evaluating the tests consists of wordlists from multiple language (sub-)families and their combinations. Combinations of related sub-families serve as positive examples while those of unrelated serve as negative examples. Evaluating the macro-families also consists of language groups whose relationship is only distantly suggested such as Nostratic (Bomhard and Kerns, 1994).

The details of data from each family are shown in Table 2. Out of these, Mon-Khmer and Munda (200 wordlists) are extracted from the Austro-Asiatic data from Rama et al. (2018). Data for Old languages of Nostratic comprising Indo-European, Dravidian, and Kartvelian are prepared by us from the Swadesh 200-wordlists available at Wiktionary[3]. Data for all the other families are obtained from Rama (2018) which were, in turn, collected from various publicly available sources. The datasets are the same as those found in related tasks such as automated cognate detection and proto-language reconstruction.

Family	Abbrv.	Languages	Concepts	Words
Austro-Asiatic	AA	58	200	11001
Austronesian	AN	45	210	8309
Indo-European	IE	42	208	8478
Pama-Nyungan	PN	67	183	11503
Sino-Tibetan	ST	64	110	6762

Table 3: Language family datasets for tree construction.

In the Nostratic grouping, we considered the languages that are surviving or have surviving descendants and were attested by the 10th century CE. The motivation behind this choice is that older languages should be closer to the ancestral language and each other if at all there is any relationship. Several languages including literary Dravidian languages, Georgian, and Armenian are mostly conservative and deviate little from their old forms. The data is pre-processed by excluding motivated word forms including onomatopoeia, and nursery forms, listed in Kessler (2001). Short forms, i.e., words consisting of single syllables are also excluded. Such cleaning is necessary to avoid the appearance of spurious relationships. In the case of Nostratic, we were also careful to exclude borrowings by tracing etymologies from Wiktionary[3]. This step could not be extended to other language families due to a lack of readily available etymological information.

All the methods employed in this work, including both the proposed one and baseline ones described in §4.2, involve the construction of a phylogenetic tree. Hence, we also compare the methods on a tree construction task where we see how well the trees match the golden truth trees wherever available. The data for this task is taken from Rama et al. (2018) as summarized in Table 3.

4.2 Multilateral Permutation Test

As mentioned in §1, most previous methods compare languages bilaterally, i.e., a pair at a time. As a result, the only possible way to compare the language families in this approach is to compare their reconstructed proto-languages. However, proto-forms of a proto-language are not often universally agreed which leads to considerable allowance of

[3] https://en.wiktionary.org/wiki/Category:Swadesh_lists_by_language

manipulation that can affect the results (Kessler, 2015). An alternate solution to determine the significance of the relationship among multiple languages was proposed by Kessler and Lehtonen (2006) and Kessler (2007) who employ a permutation test based on multilateral comparison. This has been well received in historical linguistics (Ringe and Eska, 2013).

The test is based on nearest-neighbour hierarchical clustering where at any point two closest clusters are lumped into one cluster. The basic distance measure, $\hat{d}(A, B)$, between any two clusters A and B is the average of distances between all possible pairs of languages in these clusters, i.e.,

$$\hat{d}(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (5)$$

where the distance $d(a, b)$ between any two languages a and b is the mean distance between the pairs of words over all concepts. Following the notations of §3.1 where w_{aj} and w_{bj} are words in languages a and b respectively from concept C_j,

$$d(a, b) = \frac{\sum_{C_j, w_{aj} \neq \emptyset, w_{bj} \neq \emptyset} d(w_{aj}, w_{bj})}{|\{C_j : w_{aj} \neq \emptyset, w_{bj} \neq \emptyset\}|} \quad (6)$$

Taking an average over all languages essentially enforces multilateral comparison, i.e., multiple languages are being considered equally to compute the outcome. Further, the algorithm thus described is the same as UPGMA tree construction method (Sokal and Michener, 1958) where at any bifurcating node, a uniform rate of evolution is assumed across daughter clades. The final similarity metric $\hat{s}(A, B)$ is determined by the following statistic that is computed based on a random permutation of words across each column (taxon) which yields random distances $d(A, B)$:

$$\hat{s}(A, B) = \frac{\mathbb{E}[d(A, B)] - \hat{d}(A, B)}{\mathbb{E}[d(A, B)]} \quad (7)$$

The *p-value* of two language clusters A and B is the frequency of the event $\hat{d}(A, B) \geq d(A, B)$ relative to the total number of random permutations. Language clusters A and B are considered to be *related* if the p-value is less than 0.05. The given languages are termed *related* if the final two clusters that are merged at the root are related (Kessler and Lehtonen, 2006).

Kessler (2007) ran this test using various word similarity metrics which almost give similar results. Among these metrics, we ran on P1-dolgo which is a binary metric that determines whether the consonant class of the word's initial consonant matches or not. Additionally, we employ the binary similarity measure introduced by Turchin et al. (2010) to test the significance of the Altaic family where the first two consonants are considered. We further test continuous word distances introduced by List (2010) (SCA) and List (2012) (LexStat) that are based on sequence alignment techniques which were introduced in the context of automated cognate detection.

4.3 Implementation

We mapped the consonant classes to the protein alphabet since phylogenetic software expects input as either nucleotide or amino acid sequences. Moreover, most of the amino acid letters and Dolgopolsky classes are identical. In this regard, there is only one exception, namely, 'J' which is absent in the former but present in the latter and is, hence, simply replaced with 'I', which is in turn absent in Dolgopolsky classes. The multiple sequence alignments are obtained from CLUSTALW2 (Larkin et al., 2007) while the best trees and their corresponding likelihoods were computed using IQ-TREE (Nguyen et al., 2015). As described in §3.4 and §3.5, the proportions of invariant sites P^0_{inv} and P^a_{inv} are set to 0.01 and 0.06 respectively for null (H_0) and alternate (H_a) hypotheses. The parametric bootstrap replicates are generated using AliSim (Ly-Trong et al., 2022), an extension of IQ-TREE. To replicate as closely as possible, gaps present in the original character matrices are retained in the replicates. We calculate the p-value based on a sample size of $k = 15$. The outcomes are observed to be stable beyond this size. The word similarity metrics used in the baseline models are computed by using Lingpy (List and Forkel, 2021). For the phylogenetic tree construction task, MEGA11 (Tamura et al., 2021) was used to deduce the maximim likelihood tree (ML-tree) with the aforementioned model with an additional gamma rate heterogeneity parameter with two distinct rates whose shape is estimated. We name this method *ML-P+I+G2*.

The *generalized quartet distances* (GQD) (Pompei et al., 2011) between the predicted and the gold trees are computed from quartet distances obtained using qdist (Mailund and Pedersen, 2004). The *quartet distance* between two trees measures the number of four-leaf-subsets that have dissimilar

Method	MKh	Mun	MKh-Mun	IE	Drav	May	MZ	UAz	MKh-May	MKh-UAz	AfA-LoBur
Related	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
P1-Dolgo	0.123 (<0.001)	0.243 (<0.001)	0.080 (<0.001)	0.071 (<0.001)	0.440 (<0.001)	0.228 (<0.001)	0.412 (<0.001)	0.572 (<0.001)	0.007 (<0.001)	0.005 (0.063)	0.017 (<0.001)
Turchin	0.019 (<0.001)	0.124 (<0.001)	0.019 (<0.001)	0.028 (<0.001)	0.292 (<0.001)	0.126 (<0.001)	0.256 (<0.001)	0.402 (<0.001)	0.003 (<0.001)	0.003 (0.005)	0.004 (<0.001)
LexStat	0.065 (<0.01)	0.138 (<0.01)	0.048 (<0.01)	0.036 (<0.01)	0.197 (<0.01)	0.129 (<0.01)	0.244 (<0.01)	0.306 (<0.01)	0.028 (<0.01)	0.018 (<0.01)	0.033 (<0.01)
SCA	0.087 (<0.01)	0.187 (<0.01)	0.074 (<0.01)	0.056 (<0.01)	0.296 (<0.01)	0.177 (<0.01)	0.304 (<0.01)	0.400 (<0.01)	0.015 (<0.01)	0.006 (<0.01)	0.031 (<0.01)
LRT	9.205 (<0.001)	1.58 (<0.001)	14.18 (<0.001)	26.154 (<0.001)	1.78 (<0.001)	68.212 (<0.001)	7.192 (<0.001)	10.448 (<0.001)	-14.359 (0.280)	-12.188 (0.065)	-10.768 (0.979)

Table 4: Significance testing on various existent and non-existent families. The values indicate the similarity measure $\hat{s}$ in the case of permutation tests and in the case of LRT they indicate the mean of statistic $\hat{\delta}$. Values in parentheses indicate p-value. False positives are marked in red.

topologies. Unlike biological phylogenetic trees, language trees are often multifurcated. Hence, GQD excludes penalties over the order of bifurcations. The code and relevant data have been made publicly available[4]. Further implementation details can be found in `README.md` therein.

5 Results

The primary results of the paper are tabulated in Table 4, where the results of LRT (last row) are compared with those of the multilateral permutation tests. Except for LRT, the column 'Method' indicates the distance metric employed in the permutation test. The row 'Related' indicates the current consensus about the relatedness of the language families. For the permutation test, the values indicate the similarity metric $\hat{s}$ defined in Eq. (7), as measured at the root. On the other hand, for LRT the values indicate the mean of observed $\hat{\delta}$ (see §3.5). The p-values are indicated in parentheses. The standard threshold of 0.05 is assumed for p-values. Please refer to Table 2 and Table 3 for abbreviations of various language families.

One can observe that false positives, indicated in red, are absent for LRT, in contrast with multilateral permutation tests which exhibit false positives in all cases (except P1-Dolgo for MKh-UAz). However, we note that the similarity scores of the Turchin measure are consistently small (< 0.005) for negatives irrespective of the significance implied by the p-value. Hence, it may be noted that Turchin could be a good measure for permutation tests when similarity scores are taken into consideration.

Further, one can observe from Table 4 that mean $\hat{\delta}$ values are small for valid families such as Mun and Drav. This has to do with the fact that the data for these families consists of a lower number of taxa (see Table 2). Hence, althought the $\hat{\delta}$ measure need not imply strength, its sign implies which hypothesis is to be preferred, i.e., the one with a larger proportion of invariant sites in case of a positive value and the one with a smaller proportion of invariant sites in case of a negative value.

[4] https://github.com/mahesh-ak/PhyloVal

Method	AA	AN	IE	PN	ST	Avg
P1-Dolgo	0.060	0.208	0.033	0.175	0.188	0.133
Turchin	0.069	0.195	0.058	0.175	0.275	0.154
LexStat	0.051	0.178	**0.020**	0.164	0.096	0.102
SCA	0.049	0.119	0.025	0.166	**0.087**	0.089
ML-P+I+G2	**0.026**	**0.065**	0.033	**0.145**	0.125	**0.079**

Table 5: Comparison of the methods on phylogenetic tree construction task provided as GQD scores. The best results are in **bold**.

5.1 Tree Construction

As mentioned in §4.1, both the methods output a tree, and, therefore, the methods have been evaluated on the tree construction task. The purpose of this task is to ensure that the proposed methods have indeed a good sense of phylogenetic inference and are, hence, appropriate to carry out significance tests over phylogenies. The results are tabulated in Table 5. By comparing with the mean scores of state-of-the-art language phylogeny inference methods on this data, ML-P+I+G2 (0.079) is a few steps behind Bayesian inferred tree (0.066) (Rama et al., 2018) and maximum a posteriori tree (0.051) (Rama and List, 2019). Hence, it can be concluded that consonant-class-based character matrix encoding is almost as good as cognate-based binary character matrix encoding while probabilistic methods based on character matrices are superior to distance-based methods for this task. Among the distance-based approaches, one with the SCA metric performs best. A similar situation was ob-

Method	Drav-IE	Drav-IE-Kart	May-MZ	May-UAz	May-MZ-UAz
P1-Dolgo	0.046 (<0.001)	0.038 (<0.001)	0.033 (<0.001)	0.046 (<0.001)	0.036 (<0.001)
Turchin	0.017 (<0.001)	0.002 (0.197)	0.012 (<0.001)	0.012 (<0.001)	0.008 (<0.001)
LexStat	0.024 (<0.01)	0.014 (<0.01)	0.033 (<0.01)	0.027 (<0.01)	0.024 (<0.01)
SCA	0.024 (<0.01)	0.007 (0.01)	0.019 (<0.01)	0.024 (<0.01)	0.015 (<0.01)
LRT	24.882 (<0.001)	0.316 (<0.001)	20.988 (<0.001)	-1.035 (<0.001)	-9.819 (<0.001)

Table 6: Results of evaluation of macro families. Parentheses contain p-values.

served in Rama et al. (2018) and Rama and List (2019) where SCA-based cognates yield the best performance. However, it should be noted that SCA and LexStat-based measures yield false positives on significance testing (Table 4) despite their performance on this task.

6 Evaluation of Macro Families

We apply the tests on groupings of a few families from proposed macro families, namely Nostratic, Macro-Mayan, and Amerind. Under Nostratic, we test for groupings Dravidian-Indo-European (*Drav-IE*) and Dravidian-Indo-European-Kartvelian (*Drav-IE-Kart*) while we test Mayan-Mixe-Zoque (*May-MZ*) under Macro-Mayan and Mayan-Uto-Aztecan (*May-UAz*), Mayan-Mixe-Zoque-Uto-Aztecan (*May-MZ-UAz*) under Amerind. The results are tabulated in Table 6. While going by the p-values, the LRT test seems to support all of the mentioned families. However, the mean LRT statistic $\hat{\delta}$ is weak (negative or close to 0) for Drav-IE-Kart (Nostratic) and May-UAz, May-MZ-UAz (Amerind). In other words, by looking at Eq. (4), the alternate hypothesis H_a, i.e., having higher invariant sites is not preferred. Thus, it may be concluded that LRT is a highly sensitive test since the mere addition of a single language (Georgian) to a strongly supported group of 16 languages (Drav-IE) alters the outcome drastically. This is a desirable property since the presence of even a single anomaly, an unrelated language in this case, can be detected. Note that other combinations in Nostratic such as Drav-Kart or IE-Kart are much weaker and not well supported by the permutation test itself, which is elaborated as follows.

6.1 Analysis of Permutation tests on Nostratic

Bilateral significances on Nostratic grouping Drav-IE-Kart for various distance metrics are reported in Figure 2, where the pairwise relationships based on p-value (with threshold 0.05) are color-coded. The computation follows the same steps as defined in §4.2 except that distances and similarities are calculated over pairs of languages instead of language clusters. This indeed forms the first iteration of a complete multilateral test.

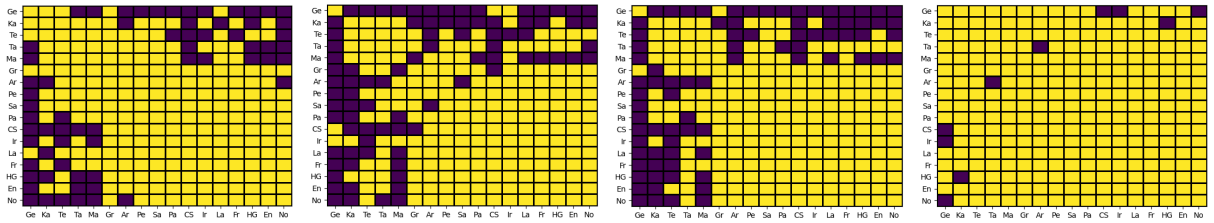
The languages are abbreviated in Fig. 2 as follows: Old Georgian (Ge), Old Kannada (Ka), Old Telugu (Te), Old Tamil (Ta), Old Malayalam (Ma), Ancient Greek (Gr), Old Armenian (Ar), Middle Persian (Pe), Sanskrit (Sa), Pali (Pa), Old Church Slavonic (CS), Old Irish (Ir), Latin (La), Old French (Fr), Old High German (HG), Old English (En) and Old Norse (No).

It is visible that for each metric, languages of the same family (IE and Drav) are almost always related pairwise. Secondly, many pairs from Drav-IE appear related. However, except for LexStat, Georgian shows to be related to at most two languages from the Drav-IE grouping. Yet, in the permutation tests for these metrics, except for Turchin (Table 6), Drav-IE-Kart appears significantly related with sometimes even good similarity scores (in the case of P1-Dolgo). All that can be concluded here is that, except for the LexStat metric, permutation tests are very sensitive to pairwise language comparisons and may not yield false positives. However, if Drav-IE-Kart is to be considered a valid grouping, these tests may be said to yield false negatives.

6.2 Analysis of ML-trees of Nostratic

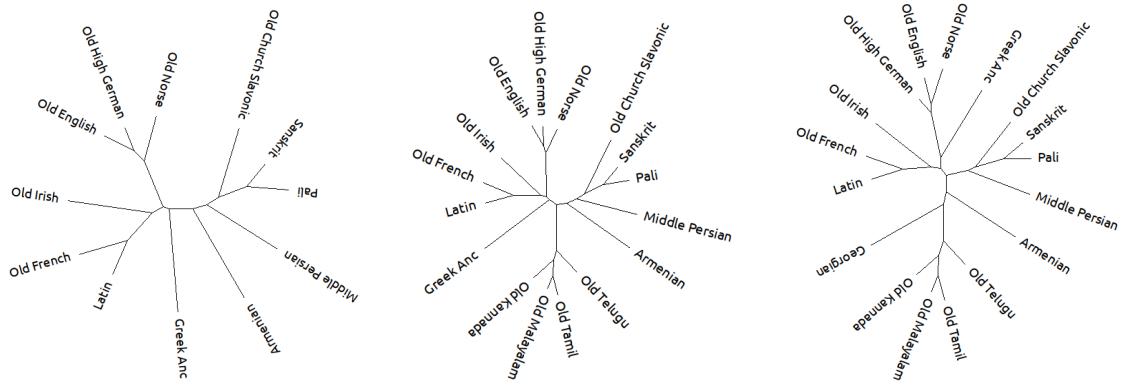
Unrooted maximum likelihood trees (ML-trees) are drawn in Figure 3 on various sub-groupings of Nostratic using MEGA11 assuming the Poisson+I model. For the IE tree (Figure 3(a)), the sub-families, except for the position of Old Church Slavonic, are highly faithful reflecting the existing notions. For instance, the topology of the Germanic family, i.e., (Old Norse, (Old English, Old High German)) contains the valid West-Germanic branch (Old English, Old High German). Similarly, the Italo-Celtic group (Old Irish, (Latin, Old French)) is visible. Also, one can distinguish a clear boundary between Western and Eastern IE languages reflecting the geographical distribution. However, the position of Old Church Slavonic intruded into Indo-Iranian appears problematic.

Further, the addition of the Dravidian family in Drav-IE does not alter the IE topology (Figure 3(b)). It is intriguing to note the western inclination of Dravidian given its eastern geographical location in the present day. However, this is in line with the observation of Caldwell (1875),



(a) P1-Dolgo (b) Turchin (c) SCA (d) LexStat

Figure 2: Bilateral (pairwise) significance among the languages of Nostratic grouping. The yellow shade implies that the relationship is statistically significant ($p < 0.05$), while the purple shade implies otherwise.



(a) IE (b) Drav-IE (c) Drav-IE-Kart

Figure 3: Comparison of unrooted ML-trees on various groupings of Nostratic language families

the founder of comparative Dravidian linguistics himself. Finally, the addition of Georgian invalidates the West-Germanic branch as well as pushes Old Greek problematically into the Western group (Figure 3(c)). However, much of the topology is undisturbed and one can also notice how the languages/families that are located south of the Caucasus namely, Armenian, Georgian, and Dravidian are grouped. Overall, it may be concluded that the addition of unrelated or weakly related languages can alter the actual topology.

Similar analyses in case of Macro-Mayan and Amerind families are provided in Appendix A where one can observe similar perturbations in topology (see Fig. 5) of one family (Mayan) in presence of others (Mixe-Zoque and Uto-Aztecan).

7 Conclusions

In this paper, we have presented a likelihood ratio test based on the proportions of invariant sites to determine the genetic relatedness of a group of languages. Our proposed test does not yield false positives, which is in contrast with previous permutation-based tests that proved to be good only for pairwise language comparisons and not for validating a language group. By applying this test, we have found strong supporting evidence for macro-families such as Dravidian-Indo-European, Macro-Mayan (for Mayan-Mixe-Zoque, and weak evidence for Nostratic (Dravidian-Indo-European-Kartvelian) and Amerind (for Mayan-Uto-Aztecan). Through secondary analyses, we have also shown that probabilistic-based methods are superior to distance-based ones based on tree construction and the correlation of topologies with geography. In this work we did not touch upon semantic shifts, i.e., words changing meaning over time; for example, the word *quick* initially meant 'lively'. While considering semantic shifts may provide room for data manipulation favoring any particular hypothesis, few semantic slots such as 'bark'-'skin' are often found to have common words. In such cases, the slots may be merged into one as suggested by Kessler (2001).

In summary, before constructing phylogenies of a group of languages, the relatedness of the group should be established through a significance test such as the one we have presented. Otherwise, the phylogenic grouping would not only be questionable but may also alter the topology of a related sub-group.

Limitations

The values of P^0_{inv} and P^a_{inv} (§3.5) are roughly decided based on the estimated ones from two examples, namely, Afrasian-Lolo-Burmese as a negative example and Indo-European as a positive example. The question of what should be the most appropriate values that should make the test optimal is not addressed here. Ideally, to address this question, more data is needed with several positive and negative examples to search for optimal values of these parameters. Also, the exact values may require calibration according to the phylogenetic software used since there could be significant differences in the implementations. Secondly, while analyzing Nostratic languages, Uralic, an important language family, has not been included due to the selection criteria (§4.1) that the languages should have been attested before 10th century CE. To include Uralic, the (Nostratic) languages that are attested around the same period as the earliest attested ones from Uralic (roughly 1300 CE onwards) should be considered to make 'fair' comparisons.

Ethics Statement

All the datasets are obtained from publicly available sources. Thus, there are no foreseen ethical considerations or conflicts of interest.

References

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023. Cognate Transformer for Automated Phonological Reconstruction and Cognate Reflex Prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6852–6862, Singapore. Association for Computational Linguistics.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024. Automated Cognate Detection as a Supervised Link Prediction Task with Cognate Transformer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 965–975, St. Julian's, Malta. Association for Computational Linguistics.

Maria Anisimova and Olivier Gascuel. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539–552.

M J Bishop and A E Friday. 1987. Tetrapod relationships: The molecular evidence. *Molecules and morphology in evolution: Conflict or compromise*, pages 123–139.

Allan R Bomhard and John C Kerns. 1994. *The Nostratic macrofamily: A study in distant linguistic relationship*. De Gruyter Mouton.

Robert Caldwell. 1875. *A comparative grammar of the Dravidian or South-Indian family of languages*. Trübner.

Lyle Campbell. 1997. *American Indian languages: The historical linguistics of Native America*, volume 4. Oxford University Press, USA.

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

Joseph Felsenstein. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249.

Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

Nick Goldman, Jon P Anderson, and Allen G Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49(4):652–670.

Joseph H Greenberg. 1963. The languages of Africa. *International Journal of American Linguistics*.

Joseph H Greenberg. 1971. The Indo-Pacific hypothesis. *Current Trends in Linguistics*, 8:807–871.

Joseph H Greenberg. 1987. *Language in the Americas*. Stanford University Press.

Joseph H Greenberg. 2000. *Indo-European and its closest relatives: The Eurasiatic language family, volume 1, grammar*, volume 1. Stanford University Press.

Joseph H Greenberg. 2005. *Genetic linguistics: Essays on theory and method*. OUP Oxford.

John P Huelsenbeck and JJ Bull. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology*, 45(1):92–98.

John P Huelsenbeck, David M Hillis, and Rasmus Nielsen. 1996. A likelihood-ratio test of monophyly. *Systematic Biology*, 45(4):546–558.

Gerhard Jäger. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757.

Gerhard Jäger. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5(1).

Gerhard Jäger. 2019. Computational Historical Linguistics. *Theoretical Linguistics*, 45(3-4):151–182.

Gerhard Jäger. 2022. Bayesian Phylogenetic Cognate Prediction. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 63–69, Seattle, Washington. Association for Computational Linguistics.

Thomas H Jukes, Charles R Cantor, et al. 1969. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132.

Alexei Kassian, Mikhail Zhivlov, and George Starostin. 2015. Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies*, 43(3-4):301–347.

Brett Kessler. 2001. The significance of word lists. *Stanford*.

Brett Kessler. 2007. Word Similarity Metrics and Multilateral Comparison. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 6–14, Prague, Czech Republic. Association for Computational Linguistics.

Brett Kessler. 2008. The Mathematical Assessment of Long-Range Linguistic Relationships. *Language and Linguistics Compass*, 2(5):821–839.

Brett Kessler. 2015. Response to Kassian et al., Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies*, 43(3-4):357–367.

Brett Kessler and Annukka Lehtonen. 2006. Multilateral comparison and significance testing of the Indo-Uralic question. *Phylogenetic methods and the prehistory of languages*, pages 33–42.

Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948.

Johann-Mattis List. 2010. SCA: Phonetic alignment based on sound classes. In *European Summer School in Logic, Language and Information*, pages 32–51. Springer.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

Johann-Mattis List and Robert Forkel. 2021. LingPy. A Python library for historical linguistics. Version 2.6.9.

Nhan Ly-Trong, Suha Naser-Khdour, Robert Lanfear, and Bui Quang Minh. 2022. AliSim: a fast and versatile phylogenetic sequence simulator for the genomic era. *Molecular Biology and Evolution*, 39(5):msac092.

Thomas Mailund and Christian NS Pedersen. 2004. QDist—Quartet distance between evolutionary trees. *Bioinformatics*, 20(10):1636–1637.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Robert L Oswalt. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior*, 3(3):117–129.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS One*, 6(6):e20109.

William Poser and Lyle Campbell. 2008. Language Classification: History and Methods.

Taraka Rama. 2018. Similarity Dependent Chinese Restaurant Process for Cognate Identification in Multilingual Wordlists. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 271–281, Brussels, Belgium. Association for Computational Linguistics.

Taraka Rama and Johann-Mattis List. 2019. An Automated Framework for Fast Cognate Detection and Bayesian Phylogenetic Inference in Computational Historical Linguistics. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6225–6235, Florence, Italy. Association for Computational Linguistics.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, New Orleans, Louisiana. Association for Computational Linguistics.

Donald A Ringe. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, 82(1):1–110.

Donald A Ringe. 1996. The mathematics of 'Amerind'. *Diachronica*, 13(1):135–154.

Donald A Ringe and Joseph F Eska. 2013. *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press.

Robert R. Sokal and Charles Duncan Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

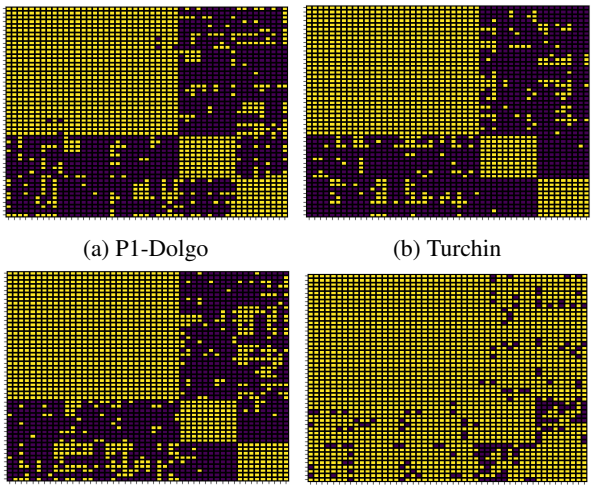
Koichiro Tamura, Glen Stecher, and Sudhir Kumar. 2021. MEGA11: molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution*, 38(7):3022–3027.

Peter Turchin, Ilia Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, (5 (48)):117–126.

Edward Orlando Wiley and Bruce S Lieberman. 2011. *Phylogenetics: Theory and practice of phylogenetic systematics*. John Wiley & Sons.

S. S. Wilks. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

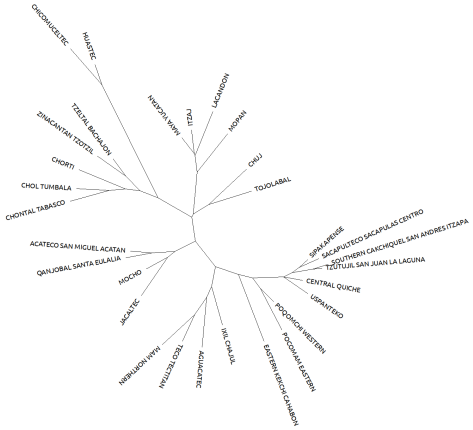
A Analysis of Macro-Mayan and Amerind



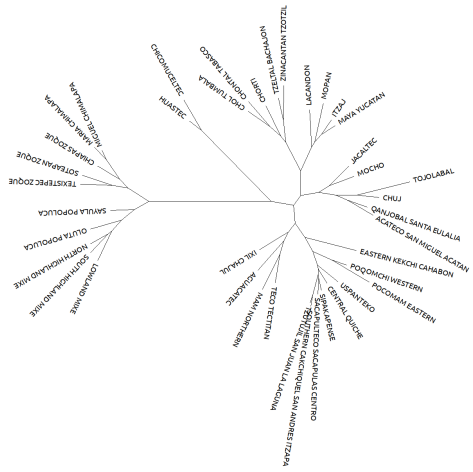
(a) P1-Dolgo (b) Turchin

(c) SCA (d) LexStat

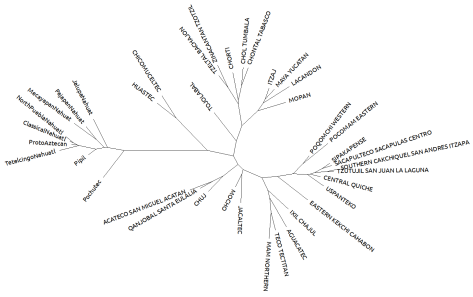
Figure 4: Bilateral (pairwise) significance among the languages of Macro-Mayan/Amerind grouping. The yellow shade implies that the relationship is statistically significant ($p < 0.05$), while the purple shade implies otherwise. While moving across the diagonal, the first cluster of significantly related languages is that of Mayan, the second is that of Mixe-Zoque and the thrid, Uto-Aztecan



(a) Mayan



(b) Mayan-Mixe-Zoque



(c) Mayan-Uto-Aztecan

(d) Mayan-Mixe-Zoque-Uto-Aztecan

Figure 5: Comparison of unrooted ML-trees on various groupings of Macro-Mayan/Amerind language families