Evaluating distance-based pyhlogenetic algorithms for automated language classification*

Gerhard Jäger
University of Tübingen, Institute of Linguistics
Wilhelmstr. 19
72074 Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

Abstract

The abstract reports a comparison of various algorithms for the task of inducing a phylogenetic tree of languages from pairwise distances. The distances used are obtained from the 5,000+ Swadesh lists of the *Automated Similarity Judgment Program* with the help of a weighted string alignment method. Automatically induced phylogenetic trees are evaluated by comparing them to several expert classifications. The main finding is that both the *Unweighted Neighbor Joining* and the BIONJ algorithm clearly outperform the widely used *Neighbor Joining* algorithm. Additionally, a post-processing of phylogenetic tree using *Nearest Neighbor Interchange* leads to a slight improvement.

1 Introduction

Recent years have seen a substantial number of publications dealing with the problem of inducing a family tree of languages from cross-linguistic data by algorithmic means. This problem is analogous to *phylogenetic inference* in computational biology, where an evolutionary tree is induced from biomolecular or phenotyical features of a collection of organisms. Computational biologists have developed a wide range of algorithmic tools for this task, which are standardly being used in computational historical linguistics as well now.

There are basically two approaches to phylogenetic inference. *Character based* methods represent each organism/language as a vector of character values for a given set of discrete characters. In linguistic applications, these characters are mostly Swadesh concepts, with cognate classes as values. Phylogenetic inference amounts to the construction of a phylogenetic tree where branches are annotated with character mutations. Phylogenetic algorithms search for a tree that optimizes a certain criterion, such as the minimal number of mutations or maximal (posterior) likelihood of all mutations combined.

Distance based methods start from a matrix of pairwise distances between organisms/languages. The algorithm computes a phylogenetic tree where branches are annotated with a length. The optimal tree is the one where the predicted distances — i.e., the total path lengths between two leafs — have the best fit to the observed distances. Algorithms differ with regard to the precise definition of optimality as well as with regard to the search procedure.

Character based methods are usually more reliable and more informative as they generate a full evolutionary history rather than just a plain tree. However, they require a classification of raw data into character classes, and this information is not always readily available. In contradistinction, distance-based methods only require an estimate of the evolutionary distance between any pair of data points. Also, distance based methods are computationally much more efficient than character-based methods and are thus better suited to explore large data sets.

Applications of character-based methods in computational historical linguistics include Gray and

*This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP, which is gratefully acknowledged.

Atkinson (2003) and Dunn et al. (2005). Distance based methods have been used, *inter alia*, in Brown et al. (2008) and Jäger (2013).

In this abstract, various efficient distance-based algorithms will be compared with regard to their suitability of automated language classification.

2 Obtaining pairwise distances from the ASJP data base

The study is carried out using version 15 of the the *Automated Similarity Judgment Project* data base (Wichmann et al. 2012), a collection of Swadesh lists for more than 5,800 languages and dialects which are phonetically transcribed in a uniform way. Only the 40 most stable Swadesh concepts were used in this paper. Retaining only word lists from languages which are alive or recently extinct and excluding creoles, 5,481 word lists were kept.

The pairwise distances between word lists were computed according to the procedure described in Jäger (2013). For reasons of space, the presentation here is kept to a minimum and the interested reader is kindly asked to consult the original literature for details.

Suppose we want to estimate the distance between two word lists L_1 and L_2. The similarity between two word forms is computed using global string alignment according to the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). In a first step, a 40×40 similarity matrix between the items in L_1 and L_2 is computed. The entries along the diagonal represent similarities between synonymous words, while the off-diagonal entries provide a sample of the similarity of unrelated words from L_1 and L_2. The aggregated distance between L_1 and L_2 is computed by means of a non-parametric one-sided test for the null hypothesis that the mean similarity between synonymous word pairs is less or equal the mean similarity between unrelated word pairs.

3 Quartet fit to expert trees

A suitable way to assess the quality of an automatically obtained phylogenetic tree is to compare it to an expert classification. Following the methodology in Jäger (2013), the gold standards used are

- the two-level classification according to the *World Atlas of Language Structures* (Haspelmath et al. 2008), abbreviated as **WALS** in the sequel,
- the classification according to *Ethnologue* (Lewis et al. 2013), abbreviated as **Ethn**, and
- the classification according to Hammerström (2010), abbreviated as **Hstr**.

The degree of fit of a phylogenetic tree to an expert tree is computed using the *quartet distance* between the two trees (Estabrook et al. 1985). Given an unrooted tree and four leaves A, B, C, and D, the tree induces the *butterfly* $(AB|CD)$ if and only if one of the bipartitions that is induced by its internal branches separates AB from CD. If there is no such internal branch, the tree induces a *star* on the quartet of leaves.

Given two unrooted trees over the same set of leaves, their quartet distance is the proportion of quartets over their leaves that have different topologies in the two trees.

As the above mentioned expert trees are strongly multiple branching while automatically generated trees are necessarily binary, the plain quartet distance is not an ideal measure of fit though. In the sequel we will use the *quartet fit* (defined in Jäger 2013), which is the proportion of butterflies of the expert tree that also occur within the automatically generated tree.

4 The *FastME* algorithms

The *FastME* software package (Desper and Gascuel 2002; software is available from `http://www.atgc-montpellier.fr/fastme/usersguide.php`) implements several distance based phylogenetic algorithms. *FastME* operates in two stages: (1) an initial tree is constructed using one of several possible distance-based algorithms, and (2) this tree is optimized by *Nearest Neighbor Interchange* (NNI), using one of several possible optimality criteria. In total, *FastME* offers five algorithms for constructing the initial tree:

1. *balanced Greedy Minimum Evolution* (bGME),
2. *ordinary least square Greedy Minimum Evolution* (OLS-GME),

3. *Neighbor-Joining* (NJ),

4. *BIONJ* (a variant of NJ, see Gascuel 1997a), and

5. *UNJ* (another variant of NJ, see Gascuel 1997b).

Additionally, there are three options for post-processing:

1. no post-processing,

2. *balanced NNI* (bNNI), or

3. *ordinary least square NNI* (OLS-NNI).

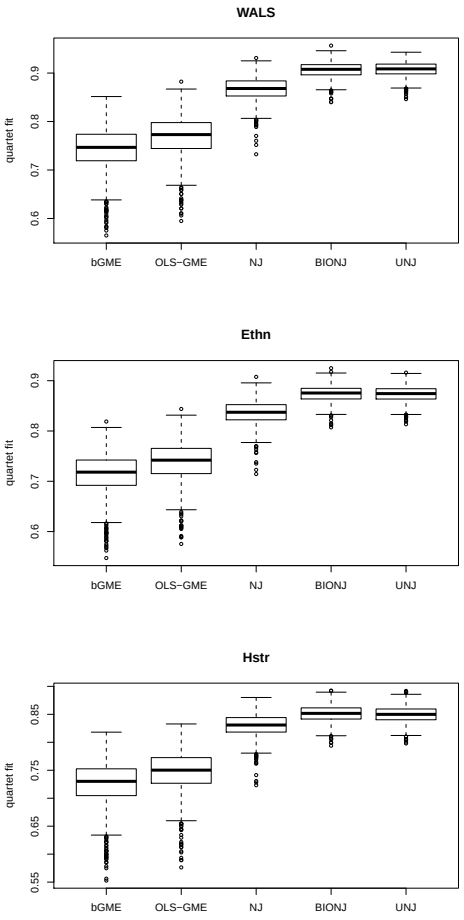

Figure 1: Quartet fits without post-processing

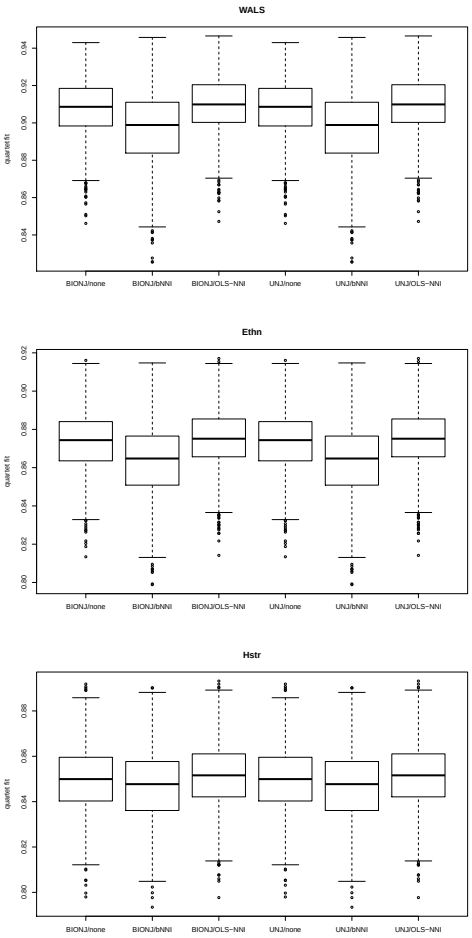

Figure 2: Quartet fits for BIONJ and UNJ with and without post-processing

5 Evaluation

To assess the suitability of these 15 algorithms for automated language classification, we followed the following procedure (cf. Jäger 2013):

1. Select a random sample of 1,000 word list from the total 5,481 word lists,

2. compute the pairwise distance matrix,

3. infer phylogenetic trees using each of the 15 *FastME* algorithms, and

4. compute the quartet distance of each of these 15 trees to the three expert trees.

This procedure is repeated 1,000 times.

	bGME	*OLS-GME*	*NJ*	*BIONJ*	*UNJ*
WALS	0.7441	0.7692	0.8671	0.9066	**0.9074**
Ethn	0.7158	0.7388	0.8362	**0.8740**	0.8737
Hstr	0.7271	0.7472	0.8303	**0.8514**	0.8498

Table 1: Quartet fits without post-processing

The mean quartet fits for the five algorithms without post-processing are given in Table 1. The corresponding distributions are visualized in Figure 1.

It is rather obvious that the three algorithms from the Neighbor-Joining family lead to massively better results than the two versions of GME. BIONJ and UNJ have a very similar performance, which is slightly better than NJ's.

In the next step we assess the effect of the post-processing options both for the outcome of BIONJ and of UNJ. The results are given in Table 2 (mean quartet fits) and visualized in Figure 2

It turns out that the effect of OLS-NNI improves the results of BIONJ and UNJ only slightly, while bNNI actually leads to a slight decrease in fit.

6 Conclusion

This abstract reported a systematic comparison of several distance-based phylogenetic algorithms and two post-processing algorithms with regard to their suitability for automated language classification. The main finding is that two variants of the popular Neighbor-Joining algorithm, namely BIONJ and unweighted NJ, both lead to a substantial improvement of the results. The quality of the results obtained with BIONJ and with UNJ are about equally good. Post-processing with Ordinary Least Square Nearest Neighbor Interchange leads to a further, if slightly, improved fit of the automatically obtained classification to expert classifications. These results hold for all three expert classifications considered here.

The general agenda pursued here — comparing different phylogenetic algorithms applied to the ASJP data with regard to how well they recover an expert classification — has to my knowledge first been explored in Pompei et al. (2011). Their methodology differs from the one used here in crucial respects though. Pompei et al. perform a separate comparison between the automatically generated trees and an expert tree for each language family separately. In the present paper, phylogenetic inference was performed for samples of word lists from separate language families. Therefore the mentioned study essentially assesses how well the internal classification of language families are algorithmically recovered, while the present investigation also factors in how well the algorithmically obtained trees separate language families. Therefore the results are by and large orthogonal to each other. Also, the mentioned authors use a different distance measure as input for phylogenetic inference. In future research, a more comprehensive comparison will be performed that takes all these aspects into account.

Acknowledgments

I thank Søren Wichmann for helpful comments on an earlier version of this paper.

References

Cecil H. Brown, Eric W. Holman, Søren Wichmann, and Viveka Velupillai, 2008. Automated classification of the world's languages: A description of the method and preliminary results. *STUF — Language Typology and Universals*, 4:285–308.

Richard Desper and Olivier Gascuel, 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology*, 9(5):687–705.

Michael Dunn, Angela Terrill, Ger Ressink, Robert A. Foley, and Stephen C. Levinson, 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

George F. Estabrook, F. R. McMorris, and Christopher A. Meacham, 1985. Comparison of undirected phylogenetic trees based on subtrees of

	BIONJ			*UNJ*		
	none	*bNNI*	*OLS-NNI*	*none*	*bNNI*	*OLS-NNI*
WALS	0.9066	0.8961	0.9082	0.9074	0.8967	**0.9092**
Ethn	0.8740	0.8632	**0.8749**	0.8737	0.8632	**0.8749**
Hstr	0.8514	0.8464	**0.8516**	0.8498	0.8470	0.8514

Table 2: Quartet fits for BIONJ and UNJ with and without post-processing

four evolutionary units. *Systematic Biology*, 34(2):193–200.

Olivier Gascuel, 1997a. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695.

Olivier Gascuel, 1997b. Concerning the NJ algorithm and its unweighted version, UNJ. In Boris Mirkin, F. R. McMorris, Fred S. Roberts, and Andrey Rzhetsky (eds.), *Mathematical Hierarchies and Biology*, DIMACS series in discrete mathematics and theoretical computer science, pages 149–171. American Mathematical Society.

Russell D. Gray and Quentin D. Atkinson, 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(27):435–439.

Harald Hammerström, 2010. A full-scale test of the language farming dispersal hypothesis. *Diachronica*, 27:197–213.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, 2008. The World Atlas of Language Structures online. Max Planck Digital Library, Munich. http://wals.info/.

Gerhard Jäger, 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. ms., University of Tübingen and Swedish Collegium of Advanced Study Uppsala.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig (eds.), 2013. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, Seventeenth edn. Online version: http://www.ethnologue.com.

Saul B. Needleman and Christian D. Wunsch, 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443453.

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

Søren Wichmann, André Müller, Viveka Velupillai, Annkathrin Wett, Cecil H. Brown, Zarina Molochieva, Julia Bishoffberger, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Harald Hammarström, Agustina Carrizo, Robert Mailhammer, Helen Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, and Anthony Grant, 2012. The ASJP Database (version 15). http://email.eva.mpg.de/∼wichmann/ASJPHomePage.htm.