COMPARATIVE ANALYSIS OF PHYLOGENIC ALGORITHMS [1]

Valery Solovyev, Renat Faskhutdinov

Abstract *The paper is dedicated to comparative analysis of phyogenetic algorithms used for linguistics tasks. At present there are a lot of phylogenetic algorithms; however, there is no unanimous opinion on which of them should be used. The paper suggests the model of language evolution trees and introduces a parameter to characterize the topology of trees. The comparison of the main algorithms is made on the trees of various topology. The paper displays that the UPGMA algorithm gives better results on the trees close to balanced ones. It provides the explanation for a number of contradictive results, described in published works.*

The problem of the input data choice and the relation between results and the number and type of parameters is under consideration. The results obtained are also ambiguous. Typological databases "Jazyki mira" and WALS as well as the method of computer modeling are used in the paper.

Keywords: I*anguage evolution, phylogenetic algorithms*

Introduction

In a number of papers [Nakhlen *et al.,* 2005-1, Nakhlen *et al.,* 2005-2, Cysouw and Comrie, 2009, Atkinson *et al.,* 2005, Donwey *et al.,* 2008, Wichmann and Saunders, 2007] attempts have been made to apply approaches developed in biology for reconstructing trees of species evolution to linguistic data. Recently compiled large databases like WALS [2005] and "Jazyki mira" [2011], ASJP [Müller *et al.,* 2010], which have introduced a great deal of new data for comparative research, hold the promise of producing new results in historic linguistics. The three databases are compared in [Polyakov *et al.,* 2009].

The phylogenetics suggests different algorithms for constructing evolutionary trees. Meanwhile the questions of better algorithm and better data are still open. The most popular phylogenetic algorithms in include UPGMA (Unweighted Pair Group Method with Arithmetic mean), NJ (Neighbour Joining), MP (Maximum Parsimony), and MrBayes.

The results described in published papers are contradictive. In the paper [Wichmann and Saunders, 2007] the NJ, MP, and Bayes algorithms were compared, and the last is considered to be the most suitable. In [Nakhlen *et al.,* 2005-1] the evolution of Indo-European family was studied and it was ascertained that NJ provides best

[1] The research was supported by Russian Foundation of Basic Research (grant № 10-06-00087-a.)

result. In fact the NJ algorithm has been recently used in linguistic researches. The belief in advantages of NJ algorithm is based on the paper [Saitou and Nei, 1987]. However, in [Donwey *et al.,* 2008] it was proved on the material of Sumba languages that UPGMA has better results. According to [Solovyev, 2011], algorithm NJ yields serious mistakes while applying the ASJP database. We compare these two algorithms as the most popular ones.

Another problem is data selection. The problem of choosing features for comparison is not trivial. In glottochronology the approach has been to only consider the most stable lexical items. A similar approach should be applied also to typological features. Attempts to define relative stabilities for WALS features are presented in [Wichmann and Kamholz, 2008] and, with improved methods, in [Wichmann and Holman, 2009].

The paper considers dependence on a number of used features and their type (i.e. what part of grammar they belong to). Besides, dependence of the results on stability of features is analyzed with the use of the "Languages of the World" database.

Comparison of algorithms

Careful analysis of the argumentation given in paper [Saitou and Nei, 1987] shows that NJ provides better results on the trees of a certain topology (= structure). As a matter of fact the authors of the paper tested only two very specific topologies of trees. Besides, the research in [Saitou and Nei, 1987] was initially oriented to the studies of biological evolution, but not a language one. The trees of a language family level are not usually like these ones. That is why the task of systematic comparison of the algorithms on the trees of different configuration is of vital importance as well as the constructing the realistic model of language evolution trees.

We analyze different cases of using the algorithms NJ and UPGMA, showing that UPGMA often gives better results than NJ in the certain cases. The influence of the tree topology on the result is being studied. Comparison of trees from papers [Nakhlen *et al.,* 2005-1] and [Donwey *et al.,* 2008] let us hypothesize that if a reconstructed tree is close to the balanced one (all branches have the same number of edges) UPGMA can be more accurate than NJ.

First of all we propose the model of language evolution trees. We studied the question of edges length variations in the real trees of language evolution. One of the most completely described trees is the evolution tree of the Turkic family, given in paper [Sravnitel'no-istoricheskaja, 2002]. The lengths of all edges in the tree (there are 77 of them) have been calculated and located in the order of increasing. The results are represented in Diagram 1.

It turned out that there are several super long edges. The longest, which is of 2130 years, corresponds to the initial separation of the Chuvash language from proto-Turkic language. The next longest edges (1330 and 1270 years) demonstrate separating the Yakut language from the Siberian branch and the Salar language from the Oguz branch. There is one abnormally short edge of 30 years that is the edge in evolution tree of Kypchat languages. The lengths of the majority of edges excluding the shortest and the ten longest edges can be strictly

put on the direct line. The fact that the lengths of the majority of edges except some of them can be put on the direct line means that the edge lengths can be considered as a random value with an even distribution.

The lengths vary from 90 to 650 years. Thus, the average meaning of an edge length is 370 years. The declination is ± 280 years that equals 75% average length. Similar results are obtained for other language families. This data is a basis for the algorithms of random tree generation below.

We conducted an experiment with generation of random binary trees of arbitrary topology to check the hypothesis. The trees were generated with a given number of leaves and the length of each edge was determined as a random number on a given interval. Then, matrixes of distances between leaves were made for every generated tree T. After that, trees T-UPGMA and T-NJ were determined by methods UPGMA and NJ.

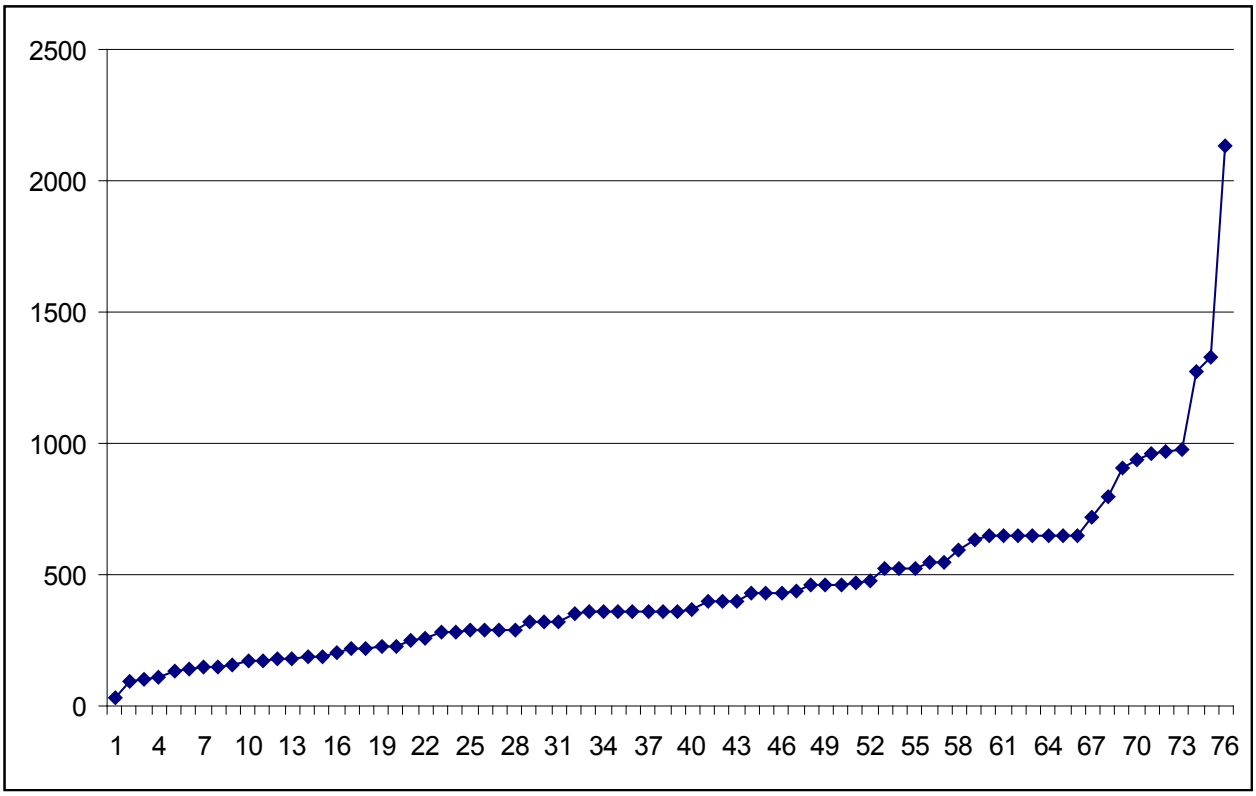

Diagram 1. Lengths of edges of the evolution tree of Turkic languages

In order to assess how big the difference between two trees is we used the Robinson-Foulds distance [Pattengale *et al.,* 2007] between them, which stands for the number of elementary transformations needed for conversion of one tree into another.

The measure of branching, as the sum of levels of inner nodes, is introduced to characterize numerically the degree of closeness of the tree to the balanced tree. In this case the root level equals 0 and the level of an ancestor is greater than the level of a descendant by 1. It is obvious that the closer a tree to the balanced tree, the smaller the measure of its branching.

To be more precise one can describe the whole algorithm as the following:

1. A random binary tree T with a given number of leaves r is generated, and all the edges are of equal length 1.
2. The measure of branching of tree T is calculated.
3. In the generated tree the length of each edge is changed by a random number from interval [– p, + p], where p = 0.75.
4. The matrix of distances between leaves is constructed by the data.
5. Trees T-UPGMA and T-NJ are constructed by the distance matrix by methods UPGMA and NJ.
6. The Robinson-Foulds distance between the obtained trees and tree T is calculated.

We made calculations for two cases, when number of leaves is equal to 15 and 50. 1000 random trees have been generated and the results have been averaged. The branching measure for the trees with 15 leaves is from 31 to 105 and for the trees with generated random sample it was from 33 to 58. It is convenient to divide all the trees by the measure of their branching into several groups in order to analyze the data obtained. We chose four groups approximately equal by the number of trees with the following values of measure: 31-36, 37-40, 41-45, 46-105. For each group we calculated the averaged Robinson-Foulds distances, given in the Table 1.

Table 1. Averaged distances, r = 15 leaves

Measure of branching	UPGMA	NJ
31 - 36	4,31	5,04
37 - 40	6,41	5,72
41 - 45	8,11	6,42
46 - 105	9,04	7,43

It is clear that the efficiency of the algorithms depends on the topology of trees. For trees with a small measure of branching, which are close to a balanced one, better results are provided by UPGMA algorithm. The similar result is obtained for r = 50.

Thus, it has been proved that NJ algorithm is not undoubtedly the best one. Both real examples and modeling by generation method of random trees shows that UPGMA is preferable in a number of cases.

Data selection

We begin handling the problem of selection of features from the WALS-based investigation of a number and type of features for algorithms UPGMA, NJ, and MP.

We used the following six pairs Americas languages from six different families (also considered in [Wichmann and Saunders, 2007]):

1. Athapaskan: Slave, Navajo
2. Uto-Aztecan: Yaqui, Comanche
3. Chibchan: Ika, Rama,
4. Aymaran: Aymara, Jaqaru
5. Otomanguean: Chalcatongo Mixtec, Lealao Chinantec
6. Carib: Hixkaryana, Carib.

We tried to reveal the dependence on the number and the type of features using this language set. Having used all the set of WALS structural features (142 feature) as well as 60 randomly chosen features (i.e. a bit less than a half of them all) we obtained the following results. Random features were chosen three times, and the average data are represented. Following [Wichmann and Saunders, 2007], we use also the 17 best features.

Table 2. Dependence on a number of features

Algorithms	17 Features	142 Features	60 Features
UPGMA	5	4	3,7
NJ	3	3	4,3
MP	4	4	4,3

A complete set of features gives a slightly inferior result, but is comparable with the set of highly-informative features, selected in [Wichmann and Saunders, 2007]. A reduced number of features (up to 60) leads to sharp change for the worse of results for UPGMA. At the same time the results for NJ and MP algorithms improved. It means that the quality of the algorithm results strongly depend on a number of features that needs further investigation. The algorithms having been analyzed are strictly divided into two groups: UPGMA and NJ, MP. The latter group works better with an average number of features.

Table 3. Dependence on a type of features

	Phonetic features	Morphological features	Syntax features
UPGMA	3	1	4
NJ	2	4	6
MP	2	4	4

The next experiment was aimed at explanation of the contribution made to general classification by separate aspects of language such as phonetics, morphology and syntax. The data are given in Table 7. Phonetic features are the features 1-19 WALS, morphological features – 20-56, syntax features – 57-128 (other WALS features are not grammatical).

It was unexpected to some extent that good results were obtained for a set of syntax features. The great expectations were connected with morphological characteristics, since they are presumably less borrowable. That is why one could expect that they would be more useful for explanation of genetic relations. On the other hand, many syntax properties change very slowly. J. Nichols [2007] suggested using some of them for establishing genetic relations.

Let us consider the ways how feature stability influents the result. General information on grammatical features' stability is available from [Wichmann & Holman [14]. We use the database "Jazyki mira". 503 most informative features (that are found at least in 25 languages but no more than in 300 languages) were selected. 4 measures for feature stability were under consideration: [Maslova, 2004], [Nichols, 1995], Wichmann & Holman [14], [Solovyev and Faskhutdinov, 2009].

For every measure the features were divided into four approximately equal by number of features groups, from the maximum (group 1) to the minimum (group 4) degree of stability. For every feature group we constructed evolution trees by NJ algorithm. The Robinson-Foulds distances were calculated between consensus tree (for languages from "Jazyki mira") and the trees constructed by NJ for all stability groups. The results are in the table 4.

Table 4. Robinson-Foulds distances for different measures and stability groups

Stability measure/Group Number	Group 1	Group 2	Group 3	Group 4
Maslova's measure	52	54	50	44
Nichols's measure	48	46	54	46
Wichmann's measure	50	52	54	44
Solovyev's measure	50	50	52	40

Best trees are constructed in the fourth group (with the lowest degree of stability) for all stability measures. The obtained result can be explained by the fact that the features from the first three groups are less informative.

Conclusion

More and more wide application of philogenetic algorithms in linguistic studies calls for consideration of justification of choice of both algorithms and data. Despite the existence of a number of methodological publications, first of all, the abovementioned [Wichmann and Saunders, 2007], many open questions remain.

The paper suggests the model of a language evolution trees and introduces the measure of trees' balance. In a number of cases, namely, for almost balanced trees, based on the model comparison of NJ and UPGMA algorithms proved higher efficiency of UPGMA. This provides theoretical explanation for a number of previously published results.

Consideration of several ways of selection of features proved the expediency of an increased attention to syntactic features, which are moderately persistent. Far from being exhaustive, the conducted research hints at promising venues of future undertakings.

Bibliography

[Atkinson *et al.,* 2005] Atkinson Q., Nicholls G., Welch D., Gray R.: From words to dates: water into wine mathemagic or phylogenetic inference? Trans. of the Philological Society. V.103:2, 2005. p.193-219.

[Donwey *et al.,* 2008] Donwey S., Halmark B., Cox M., Norquest P., Lansing S. Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction. Journal of Quantitative Linguistics. V.15, N4, 2008, pp. 340-369.

[Maslova, 2004] Maslova E. Dinamika tipologicheskih raspredelenij i stabil'nost' jazykovyh tipov. Voprosy jazykoznanija. № 5. 2004. C. 3–16. (In Russian).

[Müller *et al.,* 2010] Müller, André, Søren Wichmann, Viveka Velupillai, Cecil H. Brown, Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias Urban, Helen Geyer, and Anthony Grant. 2010. ASJP World Language Tree of Lexical Similarity: Version 3 (July 2010). http://email.eva.mpg.de/~wichmann/language_tree.htm.

[Nakhlen *et al.,* 2005-1] Nakhlen L., Ringe D., Warnow T.: Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. Language. v. 81, 2005. p. 382-420.

[Nakhlen *et al.,* 2005-2] L. Nakhleh, T. Warnow, D. Ringe, and S.N. Evans, A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset. The Transactions of the Philological Society, 3(2): 171-192, 2005.

[Nichols, 1995] Nichols J. Diachronically stable structural features. Andersen, Henning (ed.), Historical Linguistics. 1993. Selected Papers from the 11th International Conference on Historical Linguistics. Los Angeles 16–20 August 1993. Amsterdam/Philadelphia: John Benjamins Publishing Company. 1995. P. 337–355.

[Nichols, 2007] Nichols J. Typology in the service of classification. http://aalc07.psu.edu/ papers/jn_typol_class3.pdf. Stanford, 2007.

[Pattengale *et al.,* 2007] Pattengale N., Gottlieb E., Moret B. Efficiently Computing the Robinson-Foulds Metric. - Journal of Computational Biology. 2007, 14(6): 724-735.

[Polyakov *et al.,* 2009] Polyakov V., Solovyev V., Wichmann S., Belyaev O. Using WALS and Jazyki Mira. Linguistic typology. 2009. v. 13, № 1.

[Saitou and Nei, 1987] Saitou N., Nei M. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. V.4, N4, 1987. pp 406-425.

[Solovyev, 2011] Solovyev V. The Problem of Interpretation of Phylogenetic ASJP Trees. ASJP-project, 2011, http://email.eva.mpg.de/~wichmann/papers.htm

[Solovyev and Faskhutdinov, 2009] Solovyev V., Faskhutdinov R. Metodika otcenki stabil'nosti grammaticheskih svojstv. Izvestija RAN. V. 68., № 4., 2009. (In Russian).

[Sravnitel'no-istoricheskaja, 2002] Sravnitel'no-istoricheskaja grammatika tjurkskih jazykov. Red. E.R.Tenischev. Moscow: Nauka. 2002. (In Russian).

[Cysouw and Comrie, 2009] M. Cysouw, B. Comrie. How varied typologically are the languages of Africa? In: Rudie Botha & Chris Knight. (eds.) The Cradle of Language. Volume 2. Oxford, 2009.

[Jazyki mira, 2011] The Database "Jazyki mira". 2011, http://www.dblang.ru.

[WALS 2005] The World Atlas of Language Structures. Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). Oxford: Oxford University Press, 2005. 695 p.

[Wichmann and Holman, 2009] Wichmann S., Holman E. Temporal stability of linguistic typological features. Lincom Europa: Muenchen. 2009.

[Wichmann and Kamholz, 2008] Wichmann S., Kamholz D. A stability metric for typological features. STUF – Language Typology and Universals. 2008. v. 61. p.251-262.

[Wichmann and Saunders, 2007] Wichmann S., Saunders A. How to use typological database in historical linguistic research. Diachronica. 2007. v. 24, №2.

Author's information

Valery Solovyev *– Professor, Department of Computer Science, Kazan Federal University, Kremlevskaja, 18, 420008 Kazan, Russia; e-mail: maki.solovyev@mail.ru*

Major Fields of Scientific Research: Cognitive linguistics, Language evolution, Quantitative linguistics

Renat Faskhutdinov *– Junior Researcher, Department of Computer Science, Kazan Federal University, Kremlevskaja, 18, 420008 Kazan, Russia; e-mail: jvenal@mail.ru*

Major Fields of Scientific Research: Quantitative linguistics, Linguistic databases