Numerical methods from bioinformatics applied to the diachronic study of the Chibchan languages

Métodos numéricos de la bioinformática aplicados al estudio diacrónico de las lenguas chibchas

David Jiménez
Universidad de Costa Rica, Costa Rica
david.jimenezlopez@ucr.ac.cr

Haakon S. Krohn
Universidad de Costa Rica, Costa Rica
hkrohn@gmail.com

Ernesto García-Estrada
Universidad de Costa Rica, Costa Rica
er.garciaestrada@gmail.com

Viviana Solís Solís
Universidad de Costa Rica, Costa Rica
viviana9501@gmail.com

Original recibido: 23/09/22
Dictamen enviado: 31/01/23
Aceptado: 14/03/23

Abstract

In this study we applied methods from the area of bioinformatics to generate a genealogical classification of 17 languages of the Chibchan family based on a list of core vocabulary. We aligned all the words from all the possible language pairs by means of the Needleman–Wunsch algorithm, originally developed for genetic sequence alignment. Then, we calculated a normalized distance between the aligned words, taking into account the number of distinctive features that diverged between the phonemes. This procedure resembles the traditional lexicostatistical method, but it differs in the fact that it does not involve a binary labeling of cognates. Based on the mean distance between every language pair, we generated a binary tree. The results suggest that manual identification of cognates can be omitted in genealogical classification of languages and that, on the contrary, a calculation of phonological distances in terms of distinctive features can produce more precise groupings.

Keywords: alignment, bioinformatics, Chibchan languages, historical linguistics, Needleman–Wunsch algorithm

Resumen

En este estudio se aplicaron métodos de la bioinformática para generar una clasificación genealógica de 17 lenguas de la familia chibcha a partir de una lista de vocabulario básico. Todas las palabras de todos los posibles pares de lenguas se alinearon mediante el algoritmo Needleman–Wunsch, originalmente elaborado para la alineación de secuencias genéticas. Luego, se calculó una distancia normalizada entre los vocablos alineados, tomando en cuenta la cantidad de rasgos distintivos que diferían entre los fonemas. Este procedimiento se asemeja al método léxico-estadístico tradicional, pero se diferencia por el hecho de que no involucra un registro binario de cognados. Con base en la distancia promedio entre cada par de lenguas, se generó un árbol binario. Los resultados indican que la identificación manual de cognados puede omitirse en la clasificación genealógica de lenguas y que, en cambio, el cálculo de distancias fonológicas en términos de rasgos distintivos puede producir agrupaciones más precisas.

Palabras clave: *algoritmo Needleman–Wunsch, alineación, bioinformática, lenguas chibchas, lingüística histórica*

Introduction

For more than two centuries, few scholars in the areas of comparative linguistics and phylogenetics have resisted the temptation to draw parallels between the two disciplines. The first steps towards a formalization of evolution were taken by Jean-Baptiste Lamarck, who in his 1801 book established that, at least among the invertebrates, there were sufficient morphological similarities between clearly differentiated species to assume a common origin. Three decades later, Charles Darwin explicitly developed representations that today are known as phylogenetic trees. In the case of historical linguistics, Sir William Jones speculated already in 1786 about the existence of a language that would be the common ancestor of Sanskrit, Latin and Greek. This idea makes up the foundation of the concept of language families, represented by the same type of tree structures as biological species.

More than a century later, Watson and Crick (1953) published the discovery of the molecular structure of the nucleic acids, giving rise to the modern study of genetics. This is the point where a new analogy between the two disciplines that concern us in this study emerged: in essence, the vehicle through which living organisms pass on their characteristics to their descendants is a string of symbols,

which is similar to the material analyzed in historical linguistics. As regards DNA, the string in question consists of a combination of four[1] characters: the nucleic bases adenine (A), cytosine (C), guanine (G) and thymine (T). Nevertheless, Crick, Barnett, Brenner and Watts-Tobin (1961) discovered that one of the main functions of DNA is to encode proteins, which have many similarities with DNA and can also be modeled as a string of symbols; in this case, using an alphabet of 20 characters, or amino acids.[2] The replacements of one amino acid with another, as a result of transcription and codification errors, are not equal and depend little on context. These changes could therefore be considered context-free.

One of the main techniques that have been used for comparing genetic and proteomic material is called sequence alignment and has been studied for at least fifty years. Starting with the graphic representation of two genetic sequences that should match (similarly to word pairs considered cognates in linguistics), e.g., GCATGCU and GATTACA, the goal is to deduce whether the transcription errors that resulted in the divergence correspond to substitutions or deletions. In this example, one possibility for optimal alignment would be GCATG-CU and G-ATTACA, where the dash means that the base in that position in the other sequence was possibly deleted at some point.

In this research we adapt some of these techniques in a preliminary reconstruction of the family tree of the Chibchan languages, spoken in Honduras, Nicaragua, Costa Rica, Panama, Colombia, and Venezuela. This type of analysis has never been applied to this language family. However, despite certain aspects of this analysis being novel, such as the incorporation of distinctive features, some of the same techniques have been used in the past in the field of historical linguistics, at least dating back to Covington (1996), and have more recently been used by Steiner, Cysouw and Stadler (2011) and List (2014), among others.

It is important to emphasize that the objective of this exploratory study is not to propose a new subdivision of the Chibchan languages, but rather to test the algorithm on this family in order to compare the results with those from previous, more manual, research. Based on the results from this work, the algorithm can subsequently be improved with the goal of creating an effective tool for further historical analyses of both Chibchan languages and other families. A possible

1 In reality, there is a total of five canonical nucleic bases, including uracil in addition to the ones mentioned in the text. However, uracil is not found in DNA, being exclusive to RNA, where thymine is not found. Thymine and uracil are almost identical molecules, except for a methyl group that is present in the former but absent in the latter.

2 There are 22 amino acids in nature, but DNA can encode only 20 of them.

benefit is a quicker and more detailed subdivision of language groups than what is obtained from the traditional lexicostatistical method, without the need for manual identification of cognates.

Previous work

This section on previous work is divided into two subsections, one for each of the two branches of knowledge we are dealing with: bioinformatics and historical linguistics.

Previous work in bioinformatics

During the 1960s, the need to conduct alignments of genetic sequences emerged. As a consequence, Needleman and Wunsch (1970) proposed what is today known as the Needleman–Wunsch algorithm, which for more than a decade was the most well-known algorithm for this purpose. The algorithms that have replaced the Needleman–Wunsch algorithm are, in principle, variations and tweaks of the original.

In order to obtain the global alignment of two genetic sequences A and B with lengths a and b, respectively, the algorithm requires a matrix with the dimensions $(a + 1) \times (b + 1)$, where a 0 is placed in the first cell. Additionally, three parameters are needed: a positive score (or reward) n if the step corresponds to a match between the two symbols, a negative score (or penalization) m if the symbols do not match, and a negative score (or penalization) k if a hole is found (where the transcription error was an omission or an insertion of a symbol). The cells are updated using the formula:

$$F_{ij} = \max\{F_{i-1,j-1} + S(a_i, b_j), \\ F_{i,j-1} + k, F_{i-1,j} + k\},$$

where F_{ij} is the score in the cell (i, j) and $S(a_i, b_j) = n$ if $a_i = b_j$, and otherwise $S(a_i, b_j) = m$.

Each cell is updated if every neighbor above, to the left, and diagonally towards the upper left cell has been updated. The matrix stores in memory which cell obtained the highest score; if more than one cell share the highest score, any of them can be chosen.

As an example, we will use the algorithm to align the sequences GATTACA and GCATGCU, where we declare the reward for character matching as $n = 1$, whereas the two penalizations are $m = k = -1$.

The matrix is initialized with 0 in the first entry, which corresponds to the coordinates (0, 0). The cells in the first column and the first row are automati-

cally filled since values can only be taken from the cells immediately above or to the left. At this point, the value of the cell on the coordinates (1, 1) is calculated. It should be noted that if the first symbol of both words is compared, this symbol is G, and we know that S (G, G) = 1. If this score is added to the value of the cell located diagonally from the current cell, the result is 1. On the other hand, $F_{0,1} = F_{1,0} = -1$; thus, the two other values are $F_{0,1} - 1 = F_{1,0} - 1 = -2$. The highest of those scores is retrieved diagonally, being 1 in this case. Repeating the algorithm, we fill out the rest of the matrix, as shown in Table 1.

Table 1. Example of a matrix filled out using the Needleman–Wunsch algorithm.

		G	A	T	T	A	C	A
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7
G	↑ -1	↖ 1	← 0	← -1	← -2	← -3	← -4	← -5
C	↑ -2	↑ 0	← -1	↖ -1	↖ -2	↖ -3	↖ -2	← -3
A	↑ -3	↑ -1	↖ 1	← 0	← -1	↖ -1	← -2	↖ -1
T	↑ -4	↑ -2	↑ 0	↖ 2	↖ 1	← 0	← -1	← -2
C	↑ -5	↑ -3	↑ -1	↑ 1	↖ 1	← 0	↖ 1	← 0
G	↑ -6	↖ -4	↑ -2	↑ 0	↖ 0	↖ 0	↑ 0	↖ 0
G	↑ -7	↖ -5	↑ -3	↑ -1	↖ -1	↖ -1	↖ -1	↖ 0

Next, the goal is to follow the path guided by the arrows, from the bottom right to the top left corner, which yields the result displayed in Table 2.

Table 2. The same example of a matrix filled out using the Needleman–Wunsch algorithm, showing only the path from the bottom right to the top left corner.

		G	A	T	T	A	C	A
	0							
G		↖ 1						
C		↑ 0						
A			↖ 1	← 0				
T					↖ 1	← 0		
C							↖ 1	
G							↑ 0	
G								↖

Wherever a diagonal arrow is present, the corresponding position must wait until the symbols fit. On the other hand, an upwards arrow means that, in the first sequence (in this case, GATTACA), a hole is left in the alignment, represented by a dash. If the arrow points to the left, the opposite occurs: a hole is left in the second sequence and is reintegrated in the third. Holes should not match mutually, meaning that the reconstruction ends up as follows:

```
G-ATTAC-A
GCA-T-CGG
```

The second progress that we find important to mention are the BLOSUM matrices introduced in Henikoff and Henikoff (1992). When alignment techniques from genetics started being used in proteomics, scholars realized that the results are not always optimal from the perspective of the properties of the compared proteins. Hence, it became necessary to find a way to apply different penalizations for the various possible nonmatches among the 20 symbols (amino acids) in the structure. The Henikoff couple decided to analyze each of the 210 possible pairs of non-directional substitutions (including the matches). For instance, threonine and methionine are two amino acids; according to this paradigm, a change from threonine to methionine has the same penalization as a change from threonine to methionine.

Previous work in historical linguistics

Classifications of the Chibchan languages

The first explicit recognitions of genealogical relations between some of the Chibchan languages were presented by Gabb (1875), Müller (1882) and Herzog (1886). Soon after, Uhle (1890), the first scholar to establish a link between the Chibchan languages of Central and South America, proposed a subgrouping of the languages of this family. For almost a hundred years after this, according to Constenla Umaña (1985a, p. 156), the attempts of internal classification of the Chibchan languages can the grouped into three types of methodologies: inspection, counting of related terms in word lists, and use of the comparative method.

Constenla Umaña (1985a) (and partially Constenla Umaña, 1985b) introduced the lexicostatistical method to the study of this family. He took a version of the list of basic concepts first proposed by Swadesh (1955) and attempted to complete it as much as possible using several sources, to obtain a nearly uniform lexical inventory. The method consists of comparing the word pairs from the lexical inventory one by one, determining whether they are cognates or not,

and then calculating the percentage of cognates, which is taken as a measure of the degree of relation between the languages. Constenla Umaña did this with the aid of matrices of phonetic correspondences established in Constenla Umaña (1981). Despite predating Henikoff and Henikoff (1992) by more than a decade, the main ideas regarding the matrices of phonetic correspondences and the BLOSUM matrices are very similar. From there, he established quantitative criteria according to which some languages are considered members of the same subgroup. His 1985 classification is represented in Figure 1.

Figure 1. Phylogenetic tree for the Chibchan family according to Constenla Umaña (1985a). Some of the names are changed so that they match the ones used by Constenla in his English language publications. Source: produced by the authors based on Constenla Umaña (1985a).

It is clear that one of the weaknesses of this method is that the decision as to whether the terms used in two languages for the same concept are cognates is a binary choice: either they are cognates or they are not. The identification of cognates is grounded on extensive research on sound correspondences within the family, which is, as other scholars have pointed out (List, Walworth, Greenhill, Tresoldi & Forkel, 2018), a highly time-consuming task. Moreover, their binary nature leads to significant impacts on the classification if human errors are committed, and controversial cognates are far from unheard of, even in well-studied language families (see, for instance, Mallory and Adams, 2006, regarding the Indo-European family). Another issue with the method is that it is based exclusively on lexical substitution, while all other types of change are ignored. For example, since phonological change is not considered, except as a part of the identification of the cognates themselves, the degree of similarity between the cognates is not taken into account in the traditional lexicostatistical method.

Constenla Umaña (1989, 1990, 1995, 2005, 2008) refined his classification in various publications over the following years, combining the lexicostatistical method with the comparative method, so that the progressively more detailed subgroupings were not only based on cognates, but also on phonological and morphosyntactic data. His last classification, published in Spanish in Constenla Umaña (2008) and in English in Constenla Umaña (2012), is represented in Figure 2.

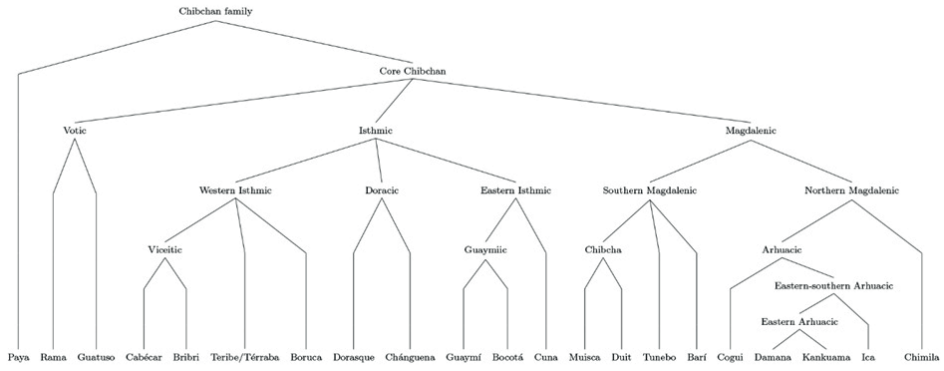

Figure 2. Phylogenetic tree for the Chibchan family according to Constenla Umaña (2008, 2012). Source: produced by the authors based on Constenla Umaña (2008, 2012).

This must be considered the "gold standard" to which the results of our algorithm should be compared, since no other classification with nearly this level of detail and refinement currently exists, and it is not challenged by any other recent classification (cf. Pache, 2018, p. 17-20). One drawback is that it cannot be used for quantitative comparisons, because Constenla Umaña incorporates the phonological and morphosyntactic aspects in a qualitative way, in many cases giving them preference over quantifiable lexical data.

It is also relevant to mention that Constenla Umaña (2012, p. 419) finds a greater phonological diversification among the Central American (Votic and Isthmic) languages than among the South American (Magdalenic) ones and therefore assumes southern Central America to be the homeland for Proto-Core-Chibchan. The qualitative analysis of Wichmann, Müller and Velupillai (2010) agrees with this homeland hypothesis.

Bioinformatic methods in historical linguistics

Results of bioinformatic methods like the one used in the present study applied to genealogical classifications of Chibchan languages have never been published. However, several scholars have proposed adaptions for its use in historical

linguistics in general. Covington (1996) modified the algorithm proposed by Needleman and Wunsch (1970), penalizing intersyllabic holes with progressively higher values. This assumption is based on the fact that syllable boundaries may also correspond to morpheme boundaries, and it is not very likely that phonemic insertions or deletions extend beyond those boundaries. For example, the Latin word *do* and the Greek word *didomi* should be aligned:

```
--do--
didomi
```

where the loss is morphemic and not phonemic. In comparison, an alignment of the type

```
d--o--
didomi
```

would be considered analogous by Needleman & Wunsch (1970).

One criticism that can be made to Covington (1996) is the fact that the sequences, once aligned, are measured using the distance given by Levenshtein (1966) (who was a computational scientist and not a linguist), such that the distance between two sequences of n symbols $A = a_1 a_2 ... a_n$ and $B = b_1 b_2 ... b_n$ is given by

$$d(A, B) = \sum_{k=1}^{n} S(a_k, b_k),$$

Where:

$$S(a_k, b_k) = \begin{cases} 0 & \text{if } a_k = b_k \\ 1 & \text{if } a_k \neq b_k \end{cases}$$

It is a fact that inside a computer two symbols are either equal or not, but in phonology and phonetics there are much more subtle contrasts. As a response to this criticism, List (2010) introduced the idea of extracting the phonetic inventory of a complete language family and group the phones together in sets based on certain similarities, treating each of these classes as one single symbol for alignment and distance purposes. This means that the Spanish words *banda* and *panda* would not only be perfectly aligned, but their distance would even be zero, because /p/ and /b/ would be considered the same symbol.

Methods

In this work we intend to measure the distance between every possible pair of 17 languages in the Chibchan family. This includes all the languages, both living and extinct, for which it was possible to compile sufficiently extensive and reliable word lists. For the calculation of these distances, we used the list of words for 100 basic concepts published by Swadesh (1971) and filled it out as completely as possible using a number of sources. Five of the terms were excluded for different reasons: either because they are not featured in the sources for most of the languages, because they do not refer to a native concept or because of frequent colexification with another term in the list.[3] For this reason, the final list used for this study consists of 95 words. All the transcriptions are phonological, since some of the sources do not provide allophonic details.

The sources specific to each language (with alternative names in parentheses) are the following: Holt (1999) and República de Honduras (2018) for Paya (Pech); Centro de Investigación y Documentación de la Costa Atlántica (1987) for Rama; Constenla Umaña (1998) for Guatuso (Malecu); Margery Peña (1989) for Cabécar; Krohn (2022) for Bribri; Quesada (2000) and Constenla Umaña (2007) for Térraba (Teribe, Naso); Quesada Pacheco (1999, 2019) for Boruca (Brunka); Quesada Pacheco (2018) for Guaymí (Ngäbere); Margery Peña & Arias Rodríguez (2005) for Bocotá (Buglere); Forster (2011) and Orán and Wagua (2011) for Cuna (Guna); Gómez Aldana (2020) for Muisca (Chibcha); Headland (1997) for Tunebo (Uwa, Uw Cuwa); Mogollón Pérez (2000) for Barí; Ortiz Ricuarte (2000) for Cogui (Kogi, Cágaba); Trillos Amaya (2000) for Damana (Guamaca, Malayo); Landaburu (2000) for Ica (Arhuaco); and Meléndez Lozano (2000) for Chimila (Ette Taara).[4]

The majority of the terms not included in these sources, especially for the Magdalenic languages, were found in Constenla Umaña (1985a), Constenla Umaña (2005) and Huber and Reed (1992). In this way, we were able to complete the list for most languages, although some ended up being incomplete.

It is important to point out that the use of Swadesh lists has been significantly criticized since their publication (e.g., Bergsland & Vogt, 1962; Eska & Ringe, 2004; McMahon & McMahon, 2006). The criticism regards both the assembly

[3] The meanings 'lie', 'sit' and 'stand' are not featured in most Chibchan dictionaries and word lists, 'dog' refers to a domestic animal imported from Europe, and 'bark' is typically colexified with 'skin', which is also included in the list.

[4] From now on, for the sake of simplicity and clarity, we will only use the language names employed by Constenla Umaña (2012). This does not mean that we oppose other name variants that also are used in literature, such as endonyms.

of the lists themselves as well as the fact that Swadesh (1955) and subsequent scholars applying the lexicostatistical method assumed a rate of lexical change that is more predictable than what the evidence suggests. That said, many authors (e.g., Lohr, 2000; Kessler, 2001; Peust, 2015; Zhang & Gong, 2016) also support the use of lexicostatistical methods inspired by the works of Swadesh.

Although there is much literature regarding the most accurate ways to compare languages diachronically, little involves a majority consensus. In our case, we chose to use Swadesh lists, firstly, because of the availability in the literature of this lexical set for the Chibchan languages. Furthermore, our work focuses on phonological similarities, and a lexical base is needed for such a comparison. If the methodology yields satisfactory results, it means that Swadesh lists can indeed be used as a base for genealogical classifications, although they might not constitute the optimal corpus.

We also assembled an integrated phonological inventory for all the 17 Chibchan languages, consisting of 10 vowel qualities (that combine with distinctive nasality and quantity) and 28 consonants. On this basis, we determined the set of binary features needed to distinguish all the phonemes. For the vowels of the Chibchan languages, as shown by Krohn (2021), five features are necessary for specifying the quality: [high], [low], [tense], [back] and [round]. Additionally, two features describing secondary distinctive properties must be used: [nasal] and [long]. A matrix with the feature values for all the vowel phonemes is found in Appendix I. In order to contrast all the consonant phonemes identified in the Chibchan languages, we used the following nine features: [sonorant], [continuant], [delayed release], [nasal], [voice], [lateral] [labial], [coronal] and [dorsal]. This selection is a subset of the system proposed by Hayes (2009, pp. 95-97) and includes just enough features to ensure that each Pan-Chibchan phoneme is distinguished from any other phoneme by at least one feature (which explains the exclusion of, for example, [approximant], [front] and [back]), with only a few exceptions.[5] A matrix with the feature values for all the consonant phonemes found in the Chibchan languages is included in Appendix II.

[5] The first exception one is the distinction between /ɾ/ and /r/ in Guatuso, which is the only Chibchan language with a contrast between these two phonemes. The features [tap] and [trill] proposed by Hayes (2009) were not included because they would produce artificially large distances between rhotics and similar phonemes, and any rhotic is still distinguished from /l/ by the feature [lateral]. Moreover, the phoneme transcribed /ɹ/ by some authors is considered equivalent to /ɽ/, which we assigned the same features as the other alveolar rhotics, because it generally does not seem to be lateral. Another exception concerns the aspirated stops in Teribe, which, according to Hayes' (2009) system, are distinguished from the non-aspirated ones by the feature [+spread glottis]; however, this feature was not

This use of distinctive features is of course a simplistic and exploratory strategy, since the original purpose of the features is to express phonological contrasts and natural classes at a language-specific level, not within groups of languages, given that the latter is not a single phonological system. Nevertheless, we needed a way to quantify more fine-grained distances between phonemes and the adoption of distinctive features is a logical approach. Another apparent issue is that changes in a feature such as [sonorant], which is a major class feature, seemingly should be assigned a higher weight than alterations of more specific features. However, this is largely made up for by the fact that a change in the value of [sonorant] almost always entails a change in another feature.[6] It is important to emphasize that this feature system is just an exploratory approach to the quantification of similarities between phonemes, a topic that needs a lot more attention in the future.

Given two phonemes f_1 and f_2, we defined a distance d_f such that if f_1 and f_2 are from different classes (one vowel and one consonant), then $d_f(f_1, f_2) = 1$, and if both are from the same class, then:

$$d_f(f_1, f_2) = \frac{\#\text{different features}}{\#\text{features of the category}}.$$

This incorporation of features in the calculation of distances between aligned words is, as far as we are aware, completely novel and we believe it to provide more precise measurements than any other, more simplified, calculations of distances between phonemes, even in comparison to List's (2010) class-based approach. The optimal set of features and the weight assigned to each specific feature is, of course, a topic that is far broader than the scope of the present paper.

We aligned the words from the Swadesh lists using the basic algorithm of Needleman and Wunsch (1970), where the reward for a match is $n = 1$, the penalization for a non-match is $m = 1 - d_f(f_1, f_2)$ and the penalization for a hole is $k = 0$. Then, we calculated a non-normalized distance between the two alignments given by $A = a_1...a_n$ and $B = b_1...b_n$, where a_k and b_k are phonemes, or holes, represented by Ø. The non-normalized distance is defined by

$$\widetilde{D}(A, B) = \sum_{k=1}^{n} d_f(a_k, b_k),$$

where we define $d_f(Ø, f) = 1$, f being any phoneme (since holes cannot match in the alignments).

included, since it would significantly increase the degree of similarity between all the other phonemes, as they would all share the same value for this feature.

6 For instance, the difference between /n/ and /d/ is not only the feature [sonorant], but also [nasal].

As a concrete example, we will show how the distance between the words /unta:s/ and /uɬi:xa/, respectively from Rama and Malecu, both meaning 'sand'.[7] The Needleman–Wunsch algorithm aligns them in the following way:

u n t a: s _

u _ ɬ i: x a

The two aligned words can be expressed as $w_1 = (p_1^{(1)}, \ldots, p_k^{(1)})$ and $w_2 = (p_1^{(2)}, \ldots, p_k^{(2)})$, where k in this specific case is 6, as each word consists of 5 phonemes plus one hole. Given the two words w_1 and w_2, and considering $k = \text{len}(w_1, w_2)$, we define the distance between the words as

$$d_w(w_1, w_2) = \sum_{i=1}^{k} \frac{d_p(p_i^{(1)}, p_i^{(2)})}{k}.$$

In the example, we have d_p(u, u) = 0, d_p(n, _) = 1, d_p(t, ɬ) = 0.222, d_p(a:, i:) = 0.429, d_p(s, x) = 0.222 and d (_, a) = 1, thus

$$d_w(\text{unta:s,uɬi:xa}) = \frac{0 + 1 + 0.222 + 0.429 + 0.222 + 1}{6} = 0.479$$

The value 0.479 is relatively high and more typical for words that are not cognates. This way of calculating distances has a clear drawback: short words that are very different (i.e., probably not cognates) have a low weight, whereas longer words that are cognates, but that have mutated significantly, could have an exaggerated weight. We therefore calculated a normalized distance D given by the formula:

$$D(A, B) = \frac{\widetilde{D}(A, B)}{n},$$

where n is the number of characters in each word after the alignment. This ensures that $0 \leq D(A, B) \leq 1$ for any aligned word pair (A, B), where $D(A, B) = 0$ only if $A = B$.

After obtaining these distances between every word pair in any two languages, we calculated the average of the distances, in order to obtain what we will call the divergence value between the languages. If we name the two languages L_1 and L_2, then the divergence value is denoted by $\Delta(L_1, L_2)$. This value behaves in a very similar way to a metric in the mathematical sense; that is, if L_1 and L_2 are two languages, the following is fulfilled:

- $\Delta(L_1, L_2) \geq 0$

[7] In order to illustrate both the alignment and the calculation of the distance in the clearest way possible, in this example we use two words that are apparently not cognates, since they therefore present greater differences between each other.

- $\Delta (L_1, L_2) \geq \Delta (L_2, L_1)$
- $\Delta (L_1, L_2) = 0$ iff $L_1 = L_2$

We then used an algorithm similar to the one given by Huffman (1952) to produce the phylogenetic tree. This algorithm looks for the two closest languages, groups them together and considers that group one sole language. From there, it defines the divergence value between this group and another language *L* as the lowest divergence value between *L* and any language in the given group. It keeps track of the order in which these groups were formed, and this is used to construct a binary tree. It must be clarified that, even though the use of trees in general is common practice in historical linguistics for language classification, our method bears a closer resemblance to the original proposal of Huffman (1952), which employs exclusively binary trees, and is thus somewhat more sensitive to diachronic changes.

Results and discussion

The classification generated by the algorithm explained above is shown in Figure 3. The number on each node is the divergence value between the two branches, which would be 0 for two identical languages and 1 for two maximally different languages or branches. It is important to bear in mind that these distances do not only refer to cognates, but to the mean divergence value of all the words in the Swadesh list, independent of their origin.

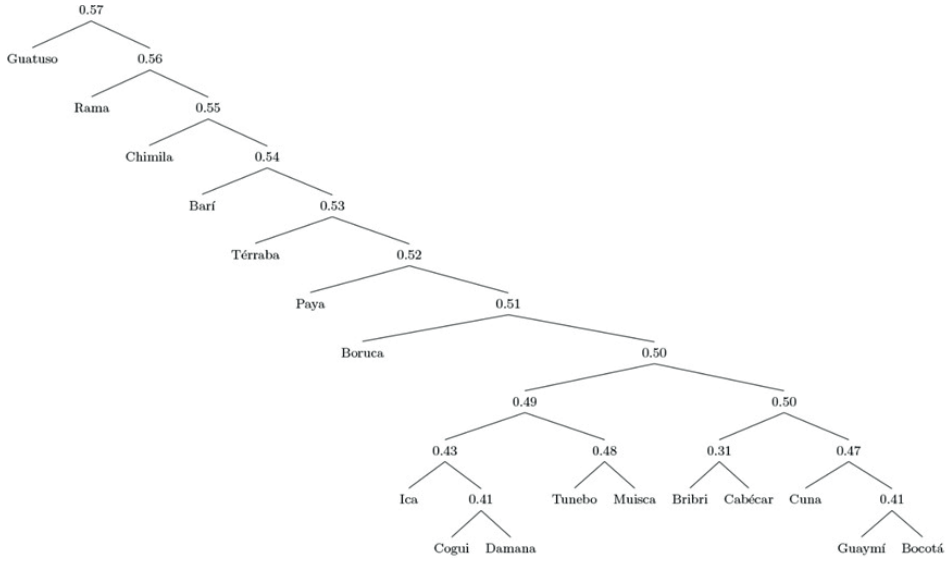

Figure 3. Phylogenetic tree generated by our algorithm for the 17 Chibchan languages included in this study. The number on each node is the divergence value between the two branches.

What is of interest is to compare this tree with the one produced by Constenla Umaña (1985a) (displayed previously in Figure 1), based on a simple cognate count, and the one published in Constenla Umaña (2008, 2012) (displayed previously in Figure 2), considered here as the gold standard, since it also incorporates phonological and morphosyntactic data, and is much more elaborate than any other classification of the Chibchan languages.

As can be seen, there are many affinities between our analysis and the latest classification by Constenla Umaña. Among the most striking ones is the fact that Cuna, Guaymí and Bocotá, all of which appeared unclassified within the Chibchan family in Constenla Umaña (1985a), turn up with the same internal divisions as in Constenla Umaña's (2008, 2012) Eastern Isthmic subgrouping. Our algorithm also connects these to the Viceitic languages (Bribri and Cabécar), which in turn appear as the most closely related language pair in the whole Chibchan family, with a divergence value of only 0.31.

Another important fact is that the Arhuacic languages (Ica, Cogui and Damana) are grouped together in our analysis. This was proposed already in Constenla Umaña (1985a), but in our classification the Arhuacic languages are also linked on a higher level to two of the other Magdalenic languages (Tunebo and Muisca). This connection was not found in the work of Constenla Umaña (1985a), where Tunebo was not even considered a part of the Chibchan family, only of the Paya-Chibchan microphylum.

On the other hand, our algorithm did not manage to link Chimila and Barí to the other Magdalenic languages, placing them together on a much higher level in the tree. These two languages are situated outside the Chibchan family in Constenla Umaña (1985a), so it is clear that their vocabularies must have evolved in quite a different manner than the other languages classified as Magdalenic by Constenla Umaña (2008, 2012).

The link between Guatuso and Rama on the very top of our tree is also worth pointing out. According to Constenla Umaña (2008, 2012), these two languages do indeed form a group of their own, separated from the Isthmic and the Magdalenic groups. This Votic group was not explicitly identified by Constenla Umaña (1985a).

The appearance of most of the Central American languages near the top of the tree and most of the South American languages near the bottom makes sense since the Chibchan family is thought to have originated in Central America

and these languages have been shown to exhibit a higher degree of phonological variation than the South American ones, as we pointed out earlier.

There are, nonetheless, also some important differences between our classification and the one developed by Constenla Umaña (2008, 2012). A significant dissimilarity is the position of Paya, which the algorithm places in between the two Isthmic languages Teribe and Boruca, whereas Constenla Umaña (2008, 2012) asserts it to be the most distinct of all the Chibchan languages, being the only one not belonging to the core of the family.

A second discrepancy involves the just mentioned Teribe and Boruca, which in our tree appear separated from the other Western Isthmic languages. In the case of Teribe, this is not surprising, however, since this language tends to turn up isolated within the Core-Chibchan grouping in lexically based analyses (Portilla, 2014), which is the case, for instance, in Constenla Umaña (1985a); the inclusion of this language in the Western Isthmic group in Constenla Umaña (2008, 2012) is, rather, based on grammatical features.

A significant observation is that the complex subgroupings at the bottom of the tree are much more solid than the order of the languages located along the main trunk. Minor tweaks on the distinctive features generate some variation in the arrangement in the upper part of the tree, while the lower subgroupings stay the same. This can be interpreted in such a way that the single languages placed along the trunk are still partially uncategorized.

Conclusions

Compared to Constenla Umaña's (1985a) analysis, which is based on the number of shared cognates between the languages, our algorithm performs significantly better for the subgroupings if the refined classification presented in Constenla Umaña (2008, 2012) is used as a reference. For every language, our classification is either closer to Constenla Umaña's latest proposal or equally imprecise as the one presented by Constenla Umaña (1985a).

This means that manual identification of cognates seems to be an unnecessary step in genealogical classification of languages, since our method does not make explicit use of that dichotomic concept. Instead, a more fine-grained analysis of phonological similarities between word lists, without considering whether the words have the same origin or not, appears to yield a more precise subdivision, although it also provides some erroneous results. The omission of manual identifying of cognates does not mean that the concept of cognates is irrelevant for the algorithm; on the contrary, it is at the core of the algorithm

because, statistically, cognates will on average yield much lower divergence values than non-cognates.

This method also comes with the benefit of being much less time-consuming and more objective. Our algorithm is of course not sufficient to create exact genealogical classifications of languages but might be a useful tool for generating a starting point for analyses that incorporate additional factors at a later stage.

In general, the proposed method has some weaknesses that need to be addressed. First, because it takes only phonological information into account, the precision might suffer in cases of lexical borrowing, which must be sorted out manually. Second, although the method can estimate divergences quantitatively up to a certain point (e.g., the last common ancestor of Bribri and Guaymí is prior to the last common ancestor of Cuna and Guaymí), this does not provide a clear-cut chronology without the use of additional tools. Third, it might well be the case that the 100-terms Swadesh lists are far from an optimal lexical base for such a classification; perhaps a different set of terms, either bigger or smaller, would produce even more precise results. Fourth, it is not yet clear what is the optimal set of distinctive features and how much weight should be assigned to each of them. Finally, and maybe most importantly, a genealogical classification of languages cannot be made solely on the basis of phonological data. Nevertheless, these weaknesses do not invalidate the method, and many improvements can be introduced in the future. It generally provides a new tool for historical linguistics, although somewhat imperfect at the moment, like any other quantitative method that exists.

References

Bergsland, K. & Vogt, H. (1962). "On the validity of glottochronology". *Current Anthropology, 3*(2), pp. 115-153.

Centro de Investigación y Documentación de la Costa Atlántica. (1987). *Diccionario elemental rama*. Retrieved from http://www.turkulka.net/docs/RMA001R035I001.pdf

Constenla Umaña, A. (1981). *Comparative Chibchan Phonology* (Doctoral thesis). University of Pennsylvania, USA.

Constenla Umaña, A. (1985a). "Clasificación lexicoestadística de las lenguas de la familia chibcha". *Estudios de Lingüística Chibcha, 4*, pp. 155-197.

Constenla Umaña, A. (1985b). "Las lenguas dorasque y chánguena y sus relaciones genealógicas". *Revista de Filología y Lingüística de la Universidad de Costa Rica, 11*(2), pp. 81-91.

Constenla Umaña, A. (1989). "La subagrupación de las lenguas chibchas: algunos nuevos indicios comparativos y léxico-estadísticos". *Estudios de Lingüística Chibcha, 8*, pp. 17-72.

Constenla Umaña, A. (1990, August). *En torno a la subagrupación de las lenguas chibchas* (Talk). Las sociedades no imperiales en los países visitados por Cristóbal Colón durante sus cuatro viajes al nuevo mundo, Panama City.

Constenla Umaña, A. (1995). "Sobre el estudio de las lenguas chibchenses y su contribución al conocimiento del pasado de sus hablantes". *Boletín del Museo del Oro, 38-39*, pp. 13-55.

Constenla Umaña, A. (1998). *Gramática de la lengua guatusa*. Heredia: Editorial de la Universidad Nacional.

Constenla Umaña, A. (2005). "¿Existe relación genealógica entre las lenguas misumalpas y las chibchenses?" *Estudios de Lingüística Chibcha, 24*, pp. 7-85.

Constenla Umaña, A. (2007). *La lengua de Térraba*. San José: Editorial de la Universidad de Costa Rica.

Constenla Umaña, A. (2008). "Estado actual de la subclasificación de las lenguas chibchenses y de la reconstrucción fonológica y gramatical del protochibchense". *Estudios de Lingüística Chibcha, 27*, pp. 117-135.

Constenla Umaña, A. (2012). "Chibchan languages". In L. Campbell & V. Grondona (Eds.), *The indigenous languages of South America: a comprehensive guide* (pp. 391-439). Berlin/Boston: De Gruyter Mouton.

Covington, M. A. (1996). "An algorithm to align words for historical comparison". *Computational Linguistics, 22*(4), pp. 481-496.

Crick, F., Barnett, L., Brenner, S. & Watts-Tobin, R. (1961). "General nature of the genetic code for proteins". *Nature, 192*(4809), pp. 1227-1232.

Dixon, R. M. W. (1997). *The rise and fall of languages*. Cambridge: Cambridge University Press.

Eska, J. F. & Ringe, D. (2004). "Recent work in computational linguistic phylogeny". *Language, 80*, pp. 569-582.

Forster, K. D. (2011). *Paya kuna: an introductory grammar*. Dallas: SIL International. https://www.sil.org/system/files/reapdata/52/88/95/52889502165272309191140937197112082328/LCDD_14_Paya_Kuna.pdf

Gabb, W. (1875). "On the Indian tribes and languages of Costa Rica". *Proceedings of the American Philosophical Society, 14*(95), pp. 438-602. https://www.jstor.org/stable/981937#metadata_info_tab_contents

Gómez Aldana, D. F. (2020, September 5). *Diccionario muisca-español*. http://muysca.cubun.org/Categor%C3%ADa:Diccionario

Hayes, B. (2009). *Introductory Phonology*. Chichester: Wiley-Blackwell.

Headland, E. R. (1997). *Uw cuwa (tunebo) – español, español – uw cuwa (tunebo), con una gramática uw cuwa (tunebo)*. Bogotá: Instituto Lingüístico de Verano. Retrieved from https://www.webonary.org/tunebo/files/DiccBilTunebo-Esp_37538.pdf

Henikoff, S. and J. G. Henikoff. (1992). "Amino acid substitution matrices from protein blocks". *Proceedings of the National Academy of Sciences, 89*(22), pp. 10915-10919.

Herzog, W. (1886). "Über die Verwandtschaftsbeziehungen der costaricensischen Indianer-Sprachen mit denen von Central- und Süd-Amerika". *Archiv für Anthropologie, 16*, pp. 623-627.

Holt, D. (1999). *Pech (Paya)*. Munich: Lincom Europa.

Huber, R. Q. & R. B. Reed. (1992). *Vocabulario comparativo. Palabras selectas de lenguas indígenas de Colombia*. Bogotá: Instituto Lingüístico de Verano.

Huffman, D. A. (1952). "A method for the construction of minimum-redundancy codes". *Proceedings of the IRE, 40*(9), pp. 1098-1101.

Kessler, B. (2001). *The Significance of Word Lists: Statistical Tests for Investigating Historical Connections between Languages*. Stanford, CA: CSLI Publications.

Krohn, H. S. (2021). "Vowel systems of the Chibchan languages". *Forma y Función, 34*(2). https://doi.org/10.15446/fyf.v34n2.88423

Krohn, H. S. (2022). *Diccionario bribri–español español–bribri*. Retrieved from http://www.haakonkrohn.com/bribri

Landaburu, J. (2000). "La lengua ika". In M. S. González de Pérez & M. L. Rodríguez de Montes (Eds.), *Lenguas indígenas de Colombia: una visión descriptiva* (pp. 733-748). Bogotá: Instituto Caro y Cuervo.

Levenshtein, V. I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics-Doklady, 10*(8), pp. 707-710.

List, J. M. (2010). Phonetic alignment based on sound classes. In M. Slavkovik (Ed.), *Proceedings of the 15th Student Session of the European Summer School for Logic, Language and Information* (pp. 192-202). Copenhague.

List, J. M. (2014). *Sequence comparison in historical linguistics* (Doctoral thesis). Düsseldorf University, Germany.

List, J. M., Walworth, M., Greenhill, S. J., Tresoldi, T. & Forkel, R. (2018). "Sequence comparison in computational historical linguistics". *Journal of Language Evolution, 3*(2), pp. 130-144. https://doi.org/10.1093/jole/lzy006

Lohr, M. (2000). "New approaches to lexicostatistics and glottochronology". In C. Renfrew, A. McMahon & L. A. Trask (Eds.), *Time Depth in Historical Linguis-*

tics (Vol. 1, pp. 209-223). Cambridge: McDonald Institute for Archaeological Research.

Mallory, J. P. & Adams, D. Q. (2006). *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford: Oxford University Press.

Margery Peña, E. (1989). *Diccionario cabécar-español español-cabécar*. San José: Editorial de la Universidad de Costa Rica.

Margery Peña, E. & Arias Rodríguez, M. (2005). "Vocabulario español-bocotá". *Estudios de Lingüística Chibcha, 24*, pp. 87-121.

McMahon, A. & McMahon, R. (2006). "Why linguists don't do dates: evidence from Indo-European and Australian languages". In P. Forster & C. Renfrew (Eds.), *Phylogenetic Methods and the Prehistory of Languages* (pp. 153-160). Cambridge: McDonald Institute for Archaeological Research.

Meléndez Lozano, M. A. (2000). "Reseña bibliográfica del chimila". In M. S. González de Pérez & M. L. Rodríguez de Montes (Eds.), *Lenguas indígenas de Colombia: una visión descriptiva* (pp. 789-792). Bogotá: Instituto Caro y Cuervo.

Mogollón Pérez, M. C. (2000). "Fonología de la lengua barí". In M. S. González de Pérez & M. L. Rodríguez de Montes (Eds.), *Lenguas indígenas de Colombia: una visión descriptiva* (pp. 219-227). Bogotá: Instituto Caro y Cuervo.

Müller, F. (1882). *Grundriss der Sprachwissenschaft* (vol. 2). Vienna: Alfred Hölder.

Needleman, S. B. & Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology, 48*(3), pp. 443-453.

Orán, R. B. & Wagua, A. (2011). *Gayamar sabga. Diccionario escolar gunagaya-español*. Panama City: Equipo EBI Guna, AECID, MEDUCA and MEF.

Ortiz Ricuarte, C. (2000). "La lengua kogui: fonología y morfosintaxis nominal". In M. S. González de Pérez & M. L. Rodríguez de Montes (Eds.), *Lenguas indígenas de Colombia: una visión descriptiva* (pp. 757-780). Bogotá: Instituto Caro y Cuervo.

Pache, M. J. (2018). *Contributions to Chibchan Historical Linguistics* (Doctoral thesis). Leiden University, The Netherlands.

Peust, C. (2015). *Towards establishing a new basic vocabulary list (Swadesh list)* (Unpublished manuscript). http://www.peust.de/peustBasicVocabularyList.pdf

Portilla, M. (2014). "La posición del naso (térraba-teriba) dentro de la ramaístmica de la familia chibcha". *Estudios de Lingüística Chibcha, 33*, pp. 241-264.

Quesada, J. D. (2000). *A grammar of Teribe*. Munich: Lincom Europa.

Quesada Pacheco, M. Á. (1999). *Diccionario boruca - español español - boruca*. San José: Editorial de la Universidad de Costa Rica.

Quesada Pacheco, M. Á. (2018). *Diccionario guaymí (ngäbere) - español y español-guaymí (ngäbere)*. Munich: Lincom Europa.

Quesada Pacheco, M. A. (2019). *Gramática boruca*. Múnich: Lincom.

República de Honduras. (2018). *Diccionario bilingüe escolar pesh - español, español - pesh*. Tegucigalpa: Secretaría de Educación.

Steiner, L., Cysouw, M. & Stadler, P. (2011). "A pipeline for computational historical linguistics". *Language Dynamics and Change, 1*(1), pp. 89-127. https://doi.org/10.1163/221058211X570358

Swadesh, M. (1955). "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics, 21*(2), pp. 121-137.

Swadesh, M. (1971). *The origin and diversification of language*. Piscataway, New Jersey: Transaction Publishers.

Trillos Amaya, M. (2000). "Categorías gramaticales del ette taara: lengua de los chimilas". In M. S. González de Pérez & M. L. Rodríguez de Montes (Eds.), *Lenguas indígenas de Colombia: una visión descriptiva* (pp. 749-756). Bogotá: Instituto Caro y Cuervo.

Uhle, M. (1890). "Aminberwandtschaften und Wanderungen der Tschitscha". *Proceedings of the International Congress of Americanists, 7*, pp. 466-489.

Watson, J. D. & Crick, F. H. (1953). "A structure for deoxyribose nucleic acid". *Nature, 171*(4356), pp. 737-738.

Wichmann, S., Müller, A. & Velupillai, V. (2010). "Homelands of the world's language families: A quantitative approach". *Diachronica, 27*(2), pp. 247-276.

Zhang, M. & Gong, T. (2016). "How many is enough? Statistical Principles for Lexicostatistics". *Frontiers in Psychology, 7*, 1916. https://doi.org/10.3389/fpsyg.2016.01916

Appendix I. Features for the vowel phonemes found in Chibchan languages.

	i	ɯ	u	ɪ	ʊ	e	ɤ	o	a	ɒ
high	+	+	+	+	+	–	–	–	–	–
low	–	–	–	–	–	–	–	–	+	+
lax	–	–	–	+	+	–	–	–	–	–
back	–	+	+	–	+	–	+	+	+	+
round	–	–	+	–	+	–	–	+	–	+

The matrix shows only the different vowel qualities; the feature [+nasal] is added to nasal vowels and [+long] is added to long vowels, while all other vowels present

the negative value for these features. All the features follow the analysis presented in Krohn (2021). This means that the symbols /ɨ/ and /ə/, used by some authors, correspond to /ɯ/ and ɤ/, respectively. Moreover, the phonemes transcribed as /ɛ/ and /ɔ/ by some authors are considered as /e/ and /o/, respectively, since no Chibchan language presents a contrast between high and low mid vowels. The symbol /ɒ/ represents the Guaymí phoneme that is usually transcribed /ɔ/, since it can be considered to be phonetically [+low].

Appendix II. Features for the consonant phonemes found in Chibchan languages.

	p	b	t	d	k	g	ʔ	ɸ	β	s	z	ɬ	ʃ	ʒ	x	ɣ	h	ʦ	ʧ	ʤ	m	n	ɲ	ŋ	l	ɾ / r / ɽ
son	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+	+
voice	–	+	–	+	–	+	–	–	+	–	+	–	–	+	–	+	–	–	–	+	+	+	+	+	+	+
nasal	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	–	–
cont	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	–	–	–	–	–	–	–	+	+
del rel	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	–	–	–	–	–	–
lateral	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	+	–
labial	+	+	–	–	–	–	–	+	+	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–
coronal	–	–	+	+	–	–	–	–	–	+	+	+	+	+	–	–	–	+	+	+	–	+	+	–	+	+
dorsal	–	–	–	–	+	+	–	–	–	–	–	–	+	+	+	+	–	–	+	+	–	–	+	+	–	–