Multiple Sequence Alignment in Historical Linguistics

A Sound Class Based Approach

Johann-Mattis List (Heinrich Heine University Düsseldorf) [1]

1 Sequences

Many structures we are dealing with - be it in daily life or in science, can be modeled as sequences. The music we listen to is a sequence of sound waves, the movies we watch are sequences of pictures, and the meals we cook can be seen as sequences of instructions drawn from a recipe book. Formally, a sequence can be defined as follows (cf. Böckenbauer and Bongartz 2003, 30f):

Definition 1.1 Given an *alphabet* (a non-empty finite set, whose elements are called *characters*), a *sequence* is an ordered list of characters drawn from the alphabet. The elements of sequences are called *segments*.

2 Alignment Analyses

Alignment analyses are the most common way to compare sequences. The main idea behind alignment analyses is to show, which segments of two or mores sequences correspond to each other. Formally, an alignment can be defined as follows (cf. Kruskal 1983, Gusfield 1997, p. 216,Durbin et al. 2002, pp. 12-22,Böckenbauer and Bongartz 2003, p. 79):

Definition 2.1 An *alignment* of two sequences *s* and *t* is a two-row matrix in which both sequences are aranged in such a way that all matching and mismatching segments occur in the same column, while empty cells, resulting from empty matches, are filled with gap symbols.

2.1 Pairwise Sequence Alignment

The general algorithm for pairwise sequence alignments roughly consists of the following steps (cf. Durbin et al. 2002, pp. 12-22, Kruskal 1983):

- Construct a matrix in which all segments of two sequences are confronted with each other and with gap characters.
- Calculate the score of all subsequences recursively by filling the matrix from left to right and from top to bottom.
- Employ a scoring function in each recursion step, which evaluates, whether the characters in each cell should be matched with themselves or with gap characters.
- Retrieve the alignment by applying a traceback function which reconstructs the ‘path of choices’ (Durbin et al. 2002) which led to the final value.

2.2 Multiple Sequence Alignment

The most common technique of multiple sequence alignment is based on a *progressive* heuristic (cf. ibid., pp. 145-148, Thompson, Higgins, and Gibson 1994), which was first proposed by Feng and Doolittle (1987). The heuristic consists of the following steps:

[1] Contact: `listm@phil.uni-duesseldorf.de`

- Align all sequences pairwise and store the scores in a matrix.
- Construct a guide tree from the matrix.
- Align all sequences along the guide tree, going from its leaves to its root.

3 Sequence Comparison in Historical Linguistics

Sequence comparison in historical linguistics, i.e. the comparison of words and morphemes which have evolved from the same ancestor forms, is traditionally carried out manually. Yet the technique by which systematic correspondences between words from different languages are retrieved are not much different from the techniques which are applied in automatic sequence analyses in evolutionary biology. Figure 1 illustrates, how the systematic comparison of the words German *Zahn* [ʦ^{h}aːn], English *tooth* [tʊːθ], Italien *dente* [dɛnte], and French *dent* [dɑ̃] leads to a reconstruction of the proto stages of the languages.

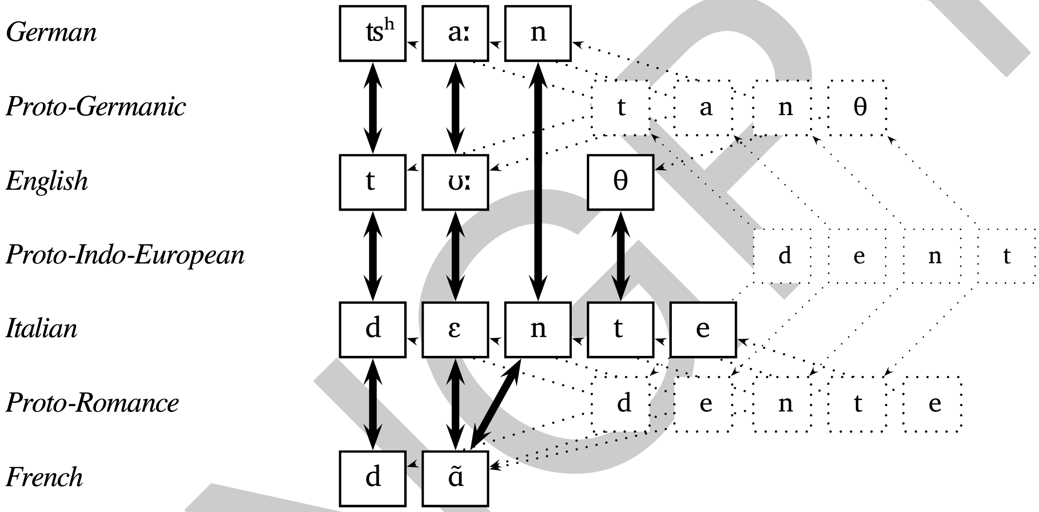


Figure 1: An Evolutionary Scenario Drawn from Sequence Comparison in Historical Linguistics

3.1 Sequence Similarity

Two incompatible views regarding the similarity of sequences can be distinguished in linguistics. One may call them the *synchronic* and the *diachronic* view. The differences between them can be summarized as follows:

- Sequences are judged to be *synchronically* similar if the segments of the sequences are phonetically similar ('phenotypic resemblence', Lass 1997, p. 130).
- Sequences are judged to be *diachronically* similar if the segments of the sequences correspond *systematically* ('genotypic resemblence', ibid.).

Tables 1 and 2 illustrate the differences between the two viewpoints (cf. Hock 1991, p. 557 for the Greek-Malay example).

3.2 Sound Classes

Neither the synchronic nor the diachronic view on sequence similarity is suitable for automatic approaches to sequence alignment in historical linguistics. The synchronic notion is not suitable, since

Greek	mati	‘eye’	≈	Malay	mata	‘eye’
Greek	θεɔs	‘god’	≈	Spanish	diɔs	‘god’

Table 1: Synchronic Similarity

German	ʦʰaːn	‘tooth’	≈	English	tʊːθ	‘tooth’
Spanish	eʧo	‘fact’	≈	French	fɛ	‘fact’

Table 2: Diachronic Similarity

the processes of sound change do not simply follow the similarity metric suggested by the synchronic framework. The diachronic notion, on the other hand, does state sequence similarity in absolute terms of *cognacy*: There is no distinction between degrees of similarity, but only an absolute and language-dependent statement regarding corresponding and non-corresponding segments and sequences. For automatic approaches, however, we need a metric which incorporates both the phonetic ‘shape’ of the sequences being compared, and the likelihood of certain sounds to correspond systematically with other sounds in genetically related languages. Here, the concept of *sound classes* which was originally derived by Aron Dolgopolsky (cf. Dolgopolsky 1986) comes into play. It is based on the following observations:

- ‘[Even] the most divergent languages show examples of phonetic change which are remarkably similar’ (Arlotto 1972, p. 77).
- Sounds which often occur in correspondence relations in genetically related languages can be clustered into *classes* (cf. Dolgopolsky 1986, Burlak and Starostin 2005, 272f).
- In contrast to the pure notion of synchronic and diachronic similarity, *sound classes* incorporate phonetic detail **and** systematic correspondence patterns within a probabilistic framework.

So far, sound classes have been only applied as a stochastic device to determine genetic language relationship (cf. Dolgopolsky 1986, Baxter and Manaster Ramer 2000, Mortarino 2009, Turchin, Peiros, and Gell-Mann 2010). In alignment analyses, however, sound classes have not been applied so far. The LingPy algorithm currently supports two sound class models (and offers an easy way to define new models based on simple text files as input): (1) The original sound class model proposed by Dolgopolsky, in which, based on an empirical basis, sounds are divided into ten types, distinguished ‘in such a way that phonetic correspondences inside a “type” are more regular than those between different types’ Dolgopolsky (1986, p. 35). (2) The sound class model proposed by the ASJP project (see `http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm`), which is based on the ASJP code (Brown et al., 2008), a transcription system, which reduces the full range of the IPA alphabet to 41 symbols. An automatic approach for the calculation of the frequency of sound correspondences among the languages of the world (Holman, Brown, and Wichmann, 2011) is employed to determine transition probabilities among the 41 sound classes.

4 A New Method for Multiple Sequence Alignment

4.1 Main Ideas

The new method for multiple sequence alignment in historical linguistics proposed here draws heavily on the traditional framework of progressive multiple sequence alignment in evolutionary biology as it is presented in the popular CLUSTAL W package (Thompson, Higgins, and Gibson, 1994). Apart from this, there are four major modifications, which have been made in order to suite the specific needs of historical linguistics:

- Phonetic sequences are internally represented as *sound classes*.
- *Scoring functions* define specific transition probabilities among sound classes.
- *Position-specific scoring* is not only based on the traditional methods used in evolutionary biology (cf. Thompson, Higgins, and Gibson 1994), but also on prosodic context.
- Alignments are automatically searched for *swapped sites*.

4.1.1 Position-Specific Scoring

The procedure for position-specific scoring based on prosodic context is a very simple adaptation of some general findings of linguistic research on sound change:

- It is assumed that sound change occurs more frequently in prosodically *weak* positions of phonetic sequences (cf. Geisler 1992).
- Given the sonority structure of a phonetic sequence, one can, apart from the *initial* and *final* positions, distinguish positions of *ascending*, *maximum* and *descending* sonority.
- These positions can be ordered in a *hierarchy of strength* (initial > ascending > descending > maximum > final).
- Based on the relative strength of all sites in a phonetic sequence, substitution scores and gap penalties are modified by scaling factors, favoring changes in weaker positions and aggravating them in stronger positions.

4.1.2 Detection of Swapped Sites

The automatic detection of swapped sites is based on the very simple fact that multiple alignment analyses which - in contrast to pairwise alignment analyses - are not sensitive to transpositions usually align swaps in a linear way, resulting in complementary structures, which can easily be detected, once an alignment analysis has been carried out. Once these structures can be determined, a specific scoring procedure for swaps is applied in order to confirm that the complementary structures might really point to swapped sites.

4.2 Working Procedure

The working procedure of the algorithm is illustrated in Figure 2.

4.3 Implementation

The algorithm is implemented as part of the LingPy library (List 2011, see `http://lingulist.de/lingpy/`). LingPy is a suite of open source Python modules for sequence comparison, distance analyses, data operations and visualization methods in quantitative historical linguistics.

5 Performance of the Method

5.1 Evaluation

In biological analyses, the performance of alignment algorithms is traditionally tested by comparing manually edited alilgnments (*reference alignments*) with those produced by the respective algorithms (*test alignments*). There exist different evaluation measures for determining the goodness of an algorithms. For this study, the following four scores shall be used:

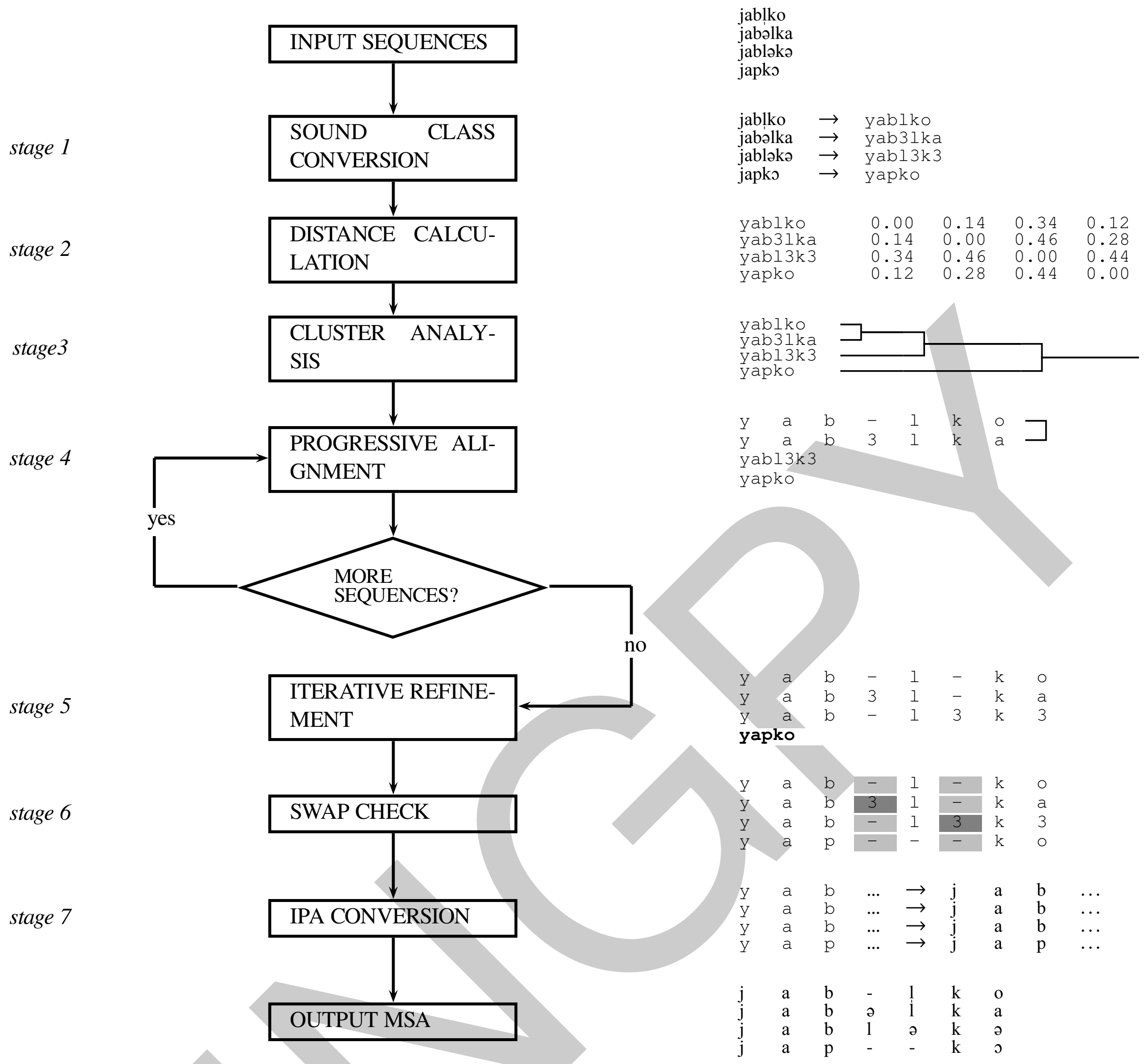


Figure 2: The Working Procedure of the LingPy Algorithm

- The *percentage of identical columns* score (PIC) calculates how many columns match in the reference and the test alignment.
- The *percentage of identical rows* score (PIR) calculates how many rows match in the reference and the test alignment.
- The *sum-of-pairs* score (SP) calculates the size of the intersection of aligned pairs of residues in the reference and the test alignment divided by the size of aligned pairs of resdidues in the reference alignment (cf. Thompson, Plewniak, and Poch 1999).
- The *modified rand index* (MRI) checks 'whether the same elements are together in the [test] alignment and the [reference] alignment' (Prokić, Wieling, and Nerbonne 2009, p. 21).

In order to evaluate the algorithm, the BulDial Gold Standard (ibid.) has been used as a reference alignment. It consists of 152 manually edited multiple alignments for a total of 192 taxa of Bulgarian dialects with a total of ca. 30,000 sequences. Two different models of the algorithm were tested: (1) The DOLGO model, consisting of 11 sound classes (vowels are treated as a separate class compared to Dolgopolsky's original proposal), a scoring function which was based on a strong CV distinction and simple matching otherwise, and a default gap penalty of -6. (2) The ASJP model, based on the ASJP code as sound class model and a scoring function derived from the work on the frequencies of sound correspondences of (Holman, Brown, and Wichmann 2011). Apart from this, the results of an earlier approach to multiple sequence alignment based on the ALPHAMALIG algorithm (Prokić, Wieling, and Nerbonne 2009) were kindly provided by the authors.

5.2 Results

Table 3 lists the scores of the different algorithms compared to the reference alignment. As the scores show, the ASJP model of LingPy performs best throughout all evaluations.

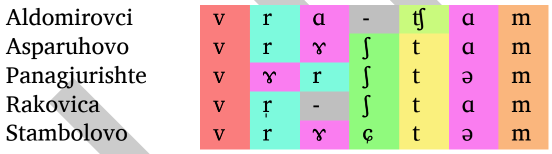
	ASJP	**DOLGO**	**ALPHA**
Perfect Alignments	132 (87%)	123 (81%)	103 (69%)
Perc. of Ident. Col.	0.9313	0.8952	0.8409
Per. of Ident. Rows	0.9531	0.9043	0.7632
Sum of Pairs	0.9901	0.9855	0.9825
Modified Rand Index	0.9902	0.9844	0.9824

Table 3: Results of the Test on the BulDial Gold Standard

Given these results, some advantages of the LingPy algorithm can be summarized:

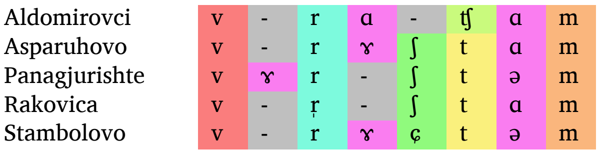
- The method for swap detection identifies 19 of 21 swapped sites in ASJP, 13 of them are aligned properly.
- The scoring function in ASJP is based on a large empirical basis, allowing fine distinctions along with the extended model of sound classes.
- The application of prosodic profiles enhances both the calculation of the guide tree and the alignment process.

Figures 3 and 4 illustrate differences between ASJP and ALPHA, where LingPy correctly identifies the swapped site. Figures 5 and 6 show differences in the alignments of ASJP and DOLGO, resulting from the different scoring functions of the two models.



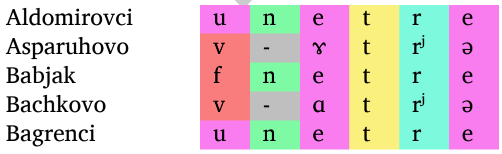
Aldomirovci	v	r	ɑ	-	ʧ	ɑ	m
Asparuhovo	v	r	ɤ	ʃ	t	ɑ	m
Panagjurishte	v	ɤ	r	ʃ	t	ə	m
Rakovica	v	r̥	-	ʃ	t	ɑ	m
Stambolovo	v	r	ɤ	ç	t	ə	m

Figure 3: MSA 21: LingPy-ASJP



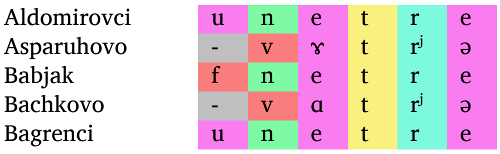
Aldomirovci	v	-	r	ɑ	-	ʧ	ɑ	m
Asparuhovo	v	-	r	ɤ	ʃ	t	ɑ	m
Panagjurishte	v	ɤ	r	-	ʃ	t	ə	m
Rakovica	v	-	r̥	-	ʃ	t	ɑ	m
Stambolovo	v	-	r	ɤ	ç	t	ə	m

Figure 4: MSA 21: ALPHAMALIG



Aldomirovci	u	n	e	t	r	e
Asparuhovo	v	-	ɤ	t	rʲ	ə
Babjak	f	n	e	t	r	e
Bachkovo	v	-	ɑ	t	rʲ	ə
Bagrenci	u	n	e	t	r	e

Figure 5: MSA 27: LingPy-ASJP



Aldomirovci	u	n	e	t	r	e
Asparuhovo	-	v	ɤ	t	rʲ	ə
Babjak	f	n	e	t	r	e
Bachkovo	-	v	ɑ	t	rʲ	ə
Bagrenci	u	n	e	t	r	e

Figure 6: MSA 27: LingPy-DOLGO

Acknowledgments

The research leading to this study was funded by the German Federal Ministery of Education and Research (BMBF). I am deeply indebted to Jelena Prokić, Martijn Wieling, and John Nerbonne for providing me with the gold standard dataset of Bulgarian dialects, the source code for the calculation of the CDE and MRI scores, and the results of their analysis using the ALPHAMALIG algorithm. I would also like to thank Hans Geisler for his helpful, critical, and inspiring support.

References

Arlotto, A. (1972). *Introduction to historical linguistics*. Boston: Mifflin.

Baxter, W. H. and A. Manaster Ramer (2000). “Beyond lumping and splitting: Probabilistic issues in historical linguistics”. In: *Time depth in historical linguistics*. Ed. by C. Renfrew, A. McMahon, and L. Trask. 2 vols. Papers in the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research, pp. 167–188.

Brown, C. H. et al. (2008). “Automated classification of the world’s languages. A description of the method and preliminary results”. In: *Sprachtypologie und Universalienforschung* 61.4, pp. 285–308.

Burlak, S. A. and S. A. Starostin (2005). *Sravnitel’no-istoričeskoe jazykoznanie (Comparative-historical linguistics)*. Moskva: Akademia.

Böckenbauer, H.-J. and D. Bongartz (2003). *Algorithmische Grundlagen der Bioinformatik*. German. Stuttgart, Leipzig, and Wiesbaden: Teubner.

Dolgopolsky, A. B. (1986). “A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia”. In: *Typology, Relationship and Time. A collection of papers on language change and relationship by Soviet linguists*. Ed. and trans. from Russian by V. V. Shevoroshkin. Ann Arbor: Karoma Publisher, pp. 27–50; Dolgopolsky, A. B. (1964). “Gipoteza drevnejshego rodstva jazykovyx semej Severnoj Evrazii s verojatnostej tochky zrenija (A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia)”. In: *Voprosy Jazykoznanija* 2, pp. 53–63.

Durbin, R. et al. (2002). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. 7th ed. Cambridge: Cambridge University Press.

Feng, D. F. and R. F. Doolittle (1987). “Progressive sequence alignment as a prerequisite to correct phylogenetic trees”. In: *Journal of Molecular Evolution* 25.4, pp. 351–360. PMID: `3118049`.

Geisler, H. (1992). *Akzent und Lautwandel in der Romania*. Romanica Monacensia 38. Tübingen: Narr.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences*. Cambridge: Cambridge University Press.

Hock, H. H. (1991). *Principles of historical linguistics*. 2nd ed. Berlin: Mouton de Gruyter.

Holman, E. W., C. H. Brown, and S. Wichmann (2011). *Sound correspondences in the world’s languages*. URL: `http://wwwstaff.eva.mpg.de/~wichmann/wwcPaper23.pdf`.

Kruskal, J. B. (1983). “An overview of sequence comparison. Time warps, string edits, and macromolecules”. In: *SIAM Review* 25.2, pp. 201–237. JSTOR: `2030214`.

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.

List, J.-M. (2011). *LingPy. A Python library for quantitative historical linguistics*. Version 1.0. URL: `http://lingulist.de/lingpy`.

Mortarino, C. (2009). “An improved statistical test for historical linguistics”. In: *Statistical Methods and Applications* 18.2, pp. 193–204.

Prokić, J., M. Wieling, and J. Nerbonne (2009). “Multiple sequence alignments in linguistics”. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (Athens, Greece, Mar. 30, 2009). Stroudsburg, PA: Association for Computational Linguistics, pp. 18–25. ACM Digital Library: `1642052`.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). “CLUSTAL W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. In: *Nucleic Acids Research* 22.22, 4673–4680. PMID: `7984417`.

Thompson, J. D., F. Plewniak, and O. Poch (1999). “A comprehensive comparison of multiple sequence alignment programs”. In: *Nucleic Acids Research* 27.13, pp. 2682–2690. PMID: `10373585`.

Turchin, P., I. Peiros, and M. Gell-Mann (2010). “Analyzing genetic connections between languages by matching consonant classes”. In: *Journal of Language Relationship* 3, pp. 117–126.