Multiple Sequence Alignment in Historical Linguistics[1]

1 Introduction

1.1 Sequences

Sets

- Sets are **unordered** lists of **unique** objects.
- Sets are **compared** by comparing the **objects** of different sets.

Sequences

- Sequences are **ordered** lists of **non-unique** objects.
- Sequences are **compared** by comparing both the **objects** (segments) and the **structure** of different sequences.

1.2 Alignments

- In alignment analyses, the corresponding **segments** of two or more **sequences** are ordered in such a way that they are **set against each other**. Segments which do not correspond to any other segments are marked by **gaps** (-). In this way, both, the structure and the segments of two or more sequences can be compared.

2 Automatic Alignment Analyses

2.1 Pairwise Sequence Alignment

- Create a **matrix** which confronts **all segments** of two **sequences**, either with **each other**, or with **gaps**.
- Seek the **path** through the matrix which is of the **lowest cost** (or the **highest score**).
- Calculate the cost (or the score) **cumulatively** by scoring the **matching** of segments with segments and with gaps by means of a specific **scoring function**.

2.2 Multiple Sequence Alignment

Guide-Tree Heuristics

- Due to computational restrictions, multiple sequence alignment (MSA) is based on **heuristics**.
- Heuristics based on **guide-trees** are the most common ones used in computational biology.
- Based on **pairwise alignment scores**, a **guide-tree** is reconstructed, and the sequences are **stepwise** added to the MSA along it (Feng & Dolittle 1987).

Profiles

- The guide-tree heuristic can be enhanced by the application of **profiles**.
- A profile consists of the **relative frequency** of all segments of an MSA in all its positions, thus, a profile represents an MSA as a **sequence of vectors**.
- Aligning profiles to profiles instead of aligning two representative sequences of two given MSA yields better results, since more information can be taken into account.

3 Alignments in Historical Linguistics

3.1 Similarity

Synchronic Similarity

- Sounds in different languages are judged to be similar, if they show **resemblences** regarding **the way they are produced or perceived**.

Diachronic Similarity

- Sounds in different languages are judged to be **similar**, if they **go back to a common ancestor**.

3.2 Sound Classes

Correspondence Classes In sound class approaches, sounds are "divided into several types and thereby distinguished in such a way that phonetic correspondences inside a 'type' are more regular than those between different 'types'." (Dolgopolsky 1986, 35).

Diachronic Similarity Similarity is not based on **synchronic resemblances** of sounds but on **class-membership**: two sounds, how dissimilar they may be from a synchronic perspective, may still belong to the same class. Class membership indicates that the **probability that sounds occur in a correspondence relationship** in genetically related languages is considerably high.

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

4 Lingpy

- LingPy (`www.lingulist.de/lingpy`) is a suite of open source Python modules for sequence comparison, and distance analyses in quantitative historical linguistics. The library allows to carry out both pairwise and multiple alignments of strings encoded in IPA or X-Sampa, using different methods and algorithms, such as global (Needleman & Wunsch 1970) and local (Smith & Waterman 1981) pairwise alignments, multiple alignments based on guide trees (Feng & Doolittle 1987), profiles (Thompson *et al.* 1994), or iteration (Barton & Sternberg 1987).

4.1 Main Ideas

Alignment of Sound Class Sequences In contrast to previous approaches, which base the alignment on the sequences as **they are given from the input**, within the sound class approach, the input strings are first **converted to sound classes** before they are aligned.

Transitions Between Sound Classes In contrast to previous sound class approaches (cf. e.g. Turchin et al. 2010), which do not allow for **transitions between sound classes**, this approach is based on a **specific scoring function**, which defines (diachronic) similarity among different sound classes.

4.2 Working Principle

- **Input** IPA-encoded strings.
- **Tokenize** the IPA-encoded strings.
- **Convert** the strings to sound-class strings.
- **Align** the sound-class strings.
- **Output** IPA-encoded, aligned strings.

4.3 Scoring

Directionality of Sound Changes One crucial characteristic of certain well-known sound changes is their **directionality**, i.e. if certain sounds change, this change will go into a certain direction and the **reverse change can rarely be attested**.

Directionality and Sound Correspondences While the nature of certain sound changes **may be directional**, sound correspondences do not directly reflect this directionality, and neither do scoring functions for sequence alignments, since these are **not directional *per definitionem***, since the distance or similarity between two segments is always the same, regardless from which segment we start to compare.

Reflecting Directionality in Undirected Networks In this approach, the directionality of certain sound changes is accounted for by creating a **non-metric scoring function**. While in a metric scoring function the distance between two segments *A* and *B* would depend on the distance of *A* and *B* to a third segment *C* in such a way that, according to the **triangle inequality** the distance from *A* to *B* could **not exceed the sum of the distances from *A* to *C* and from *B* to *C***, this does not hold for the probability of those sound correspondences, which occur as a product of directional sound change.

5 Performance of the Method

Please refer to `http://lingulist.de/` for the examples given in the slides.

References

Barton, Geoffrey J., & Michael J. E. Sternberg. 1987. A strategy for the rapid multiple alignment of protein sequences : Confidence levels from tertiary structure comparisons. *Journal of Molecular Biology* 198.327 – 337.

Dolgopolsky, A. B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In *Typology Relationship and Time*, ed. by Vitaly V. Shevoroshkin & T. L. Markey, 27–50. Karoma Publisher, Inc. Originally published in Russian as "Gipoteza drevnejščego rodstva jazykov Severnoj Evrazii (problemy fonetičeskich sootvetstvij)" in 1964.

Feng, D. F., & R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* 25.351–360.

Needleman, Saul B., & Christan D. Wunsch. 1970. A gene method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48.443–453.

Smith, T. F., & M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 1.195–197.

Thompson, J. D., D. G. Higgins, & 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. *Nucleic Acids Research* 22.4673–4680.