Phylogenetic Analysis of Ancient Lexical Lists

Erin Becker, Associate Director, The Carpentries & Niek Veldhuis, Professor of Assyriology, UC Berkeley, Department of Near Eastern Studies

Introduction (Lexical Texts)

Lexical lists:

- Lists of Sumerian words, documenting a dead language
- Written in cuneiform on clay tablets
- Used in scribal education
- Often thematic (lists of trees, wooden objects, animals, foods)

Where and When?

- Ancient Mesopotamia (present-day Iraq); ca 1900-1200 BCE

Lists of trees, wooden objects, animals, meat cuts, professions, metal objects, etc. were transmitted over many centuries. Like genomes, these lists continuously changed by adding new entries, omitting entries, or by changing the order of entries. Lists are organized in *sections*; each section includes related words (for instance: parts of a chariot). The order of sections also changes over time.

Scholars will assess the relationship between two versions of a lexical text by looking at **presence/absence** of entries, **order** of entries within a section, and **order of sections** within the entire text.

Lexical lists and other cuneiform texts are published in transliteration and translation with lemmatization on ORACC (Open Richly Annotated Cuneiform Corpus; http://oracc.org).

Cuneiform:	
Transliteration	{ŋeš}taškarin
Translation	boxwood
Lemmatization	taškarin[boxwood]N

Four representations of a Sumerian word

Data Acquisition and Formatting

Lemmatized data, made available by ORACC in JSON format, is used in our analyses. Data consist of ca. 135 exemplars of the list of trees and wooden objects, ranging from 2 to 750 lines in length. The data is arranged in a DTM (Document-Term Matrix) for comparison of the documents on **presence/absence of entries**.

Doc/lemma	ŋeškin[birch]N	ŋešnu[sandalwood]N	ŋeštin[vine]N	taškarin[boxwood]N
P459216	0	1	1	0
P253866	0	0	0	0
P332930	0	1	1	1

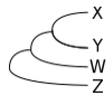
For the analysis of **section order** each item in each text is assigned to a **section.** Sections are defined through the best preserved text: the Nippur version. Consecutive lemmas that share a sequence of least three characters (k-mer with k=3) belong to the same section

pana[bow]N
epana[quiver]N
gagpana[arrow]N
gagsisa[arrow]N
gagsieš[arrow]N
==========
ešad[trap]N

The section "bow" is defined by a sequence of lemmas that share a sequence of at least 3 characters either in the Sumerian Citation Form or in the English guide word. Square brackets and part of speech are ignored. The word for "trap" is the beginning of a new section

Introduction (Phylogenetics)

Phylogenetics is a field of evolutionary biology which uses patterns of similarity and difference among organisms to make hypotheses about relatedness. Hypothesized relationships are displayed as trees where more closely related organisms are separated by fewer branching points and/or shorter total branch length.



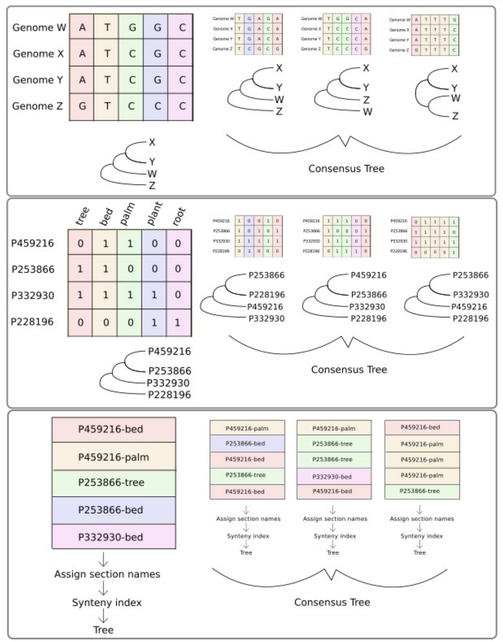
Species X is most closely related to Species Y. Both are next most closely related to Species W and most distantly related to Species Z.

Patterns in DNA sequences, protein sequences, gene order, morphological characters, or others can be used to make phylogenetic hypotheses.

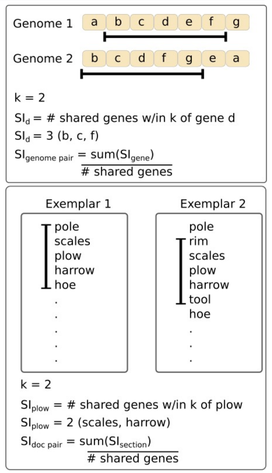
Shared Issues

Lexical lists and genomic data share characteristics making similar approaches appropriate.

- Texts with ordered information
- Broken/fragmented texts
- Texts copied with changes
- Incomplete collection of exemplars

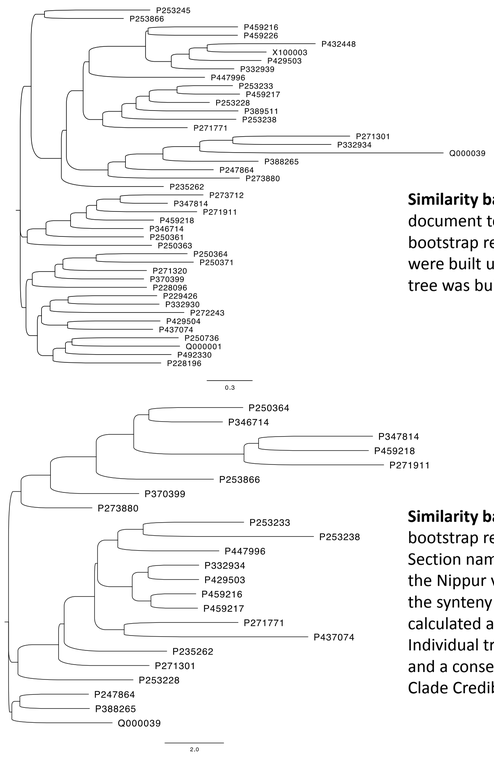


Bootstrapping is a method of assessing how sensitive results are to changes in the data. A large number of replicates are produced by sampling with replacement and the analysis is carried out on each replicate. If the results are robust, most replicates will be similar.



Synteny is a measure of similarity in order of items across exemplars. In phylogenetics, synteny is the conservation of gene order across genomes. In our analysis, it is the conservation of section order across exemplars.

Analysis



Similarity based on entry presence/absence. A document term matrix was constructed and 1000 bootstrap replicates were taken. Individual trees were built using Neighbor Joining and a consensus tree was built using Maximum Clade Credibility.

Similarity based on section order. One thousand bootstrap replicates were taken of the corpus. Section names were assigned based on sections in the Nippur version (Q000039). For each replicate, the synteny index for each document pair was calculated and used to construct a distance matrix. Individual trees were built using Neighbor Joining and a consensus tree was built using Maximum Clade Credibility.

Results and Future Work

Entry presence/absence turns out to be a rather bad predictor for dependency among tablets. Two duplicating tablets from Old Babylonian Isin (P459216 and P459217) end up in different branches of the tree, presumably because they are broken at different places. The tree does correctly identify four tablets that have little to do with the rest: P250736, Q000001, P492330, and P228196 (at the bottom of the tree)

The "section order" tree closely aligns P346714 (from Ur) with P250364 (unknown provenance). P250364 certainly does not come from Ur but may well descend from the Ur version. The Old Babylonian Isin texts P459216 and P459217 align with two exemplars from Ugarit (P332934 and P429503), which are several centuries later. The Ugarit version is unlikely to derive directly from Isin – but Isin may well have been a station on the way.

Future work:

- order of entries within a section
- applying the method on other groups of lexical texts

Contact

Erin Becker
Associate Director, The Carpentries
Email: ebecker@carpentries.org
Website: https://github.com/ErinBecker/digital-humanities-phylogenetics

Niek Veldhuis
Professor of Assyriology, Dept. of Near Eastern Studies, UC Berkeley
Email: veldhuis@berkeley.edu
Website: http://oracc.org/dcclt