String alignment, evolution, phylogenetics, and how languages evolve

Rahul Siddharthan

The Institute of Mathematical Sciences, Chennai, India

Bits & Strings: iCEL Workshop on Computational Epigraphy, 28 March 2024

DNA evolution

- Single nucleotide substitution
- Insertion/deletion
- Segmental duplication
- Recombination
- etc...

DNA evolution

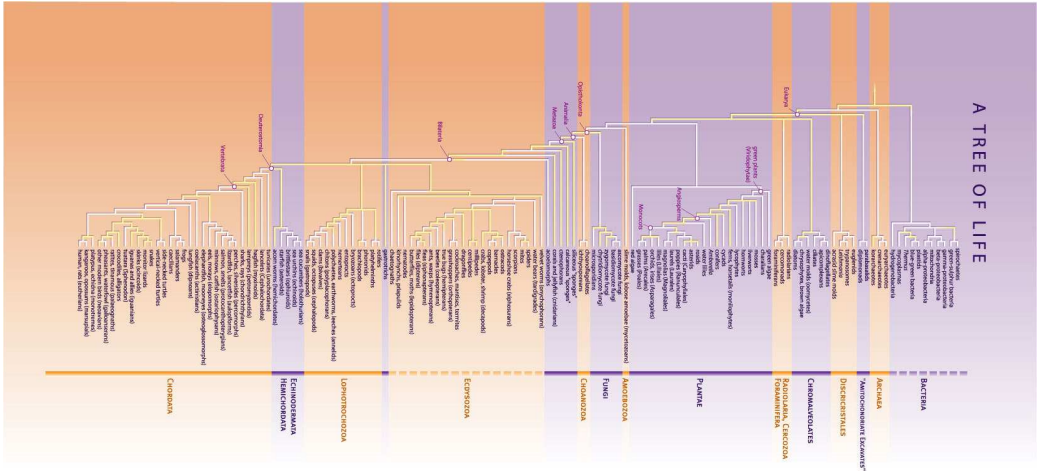
- Single nucleotide substitution
- Insertion/deletion
- Segmental duplication
- Recombination
- etc...

For phylogenetics

- Most models consider only single-nucleotide mutation
- Start with a multiple sequence alignment, throw out all indels
- More sophisticated models exist, but it gets complicated...

Tree of life

From Science magazine's special issue, 13 June 2003



Sequence alignment

Task: given two sequences, match the bases that are likely to be derived from a common ancestor.
Or, given a "scoring function" for alignments, find the alignment that optimises the scoring function.
Exact methods exist, as well as approximate but faster methods.

Sequence alignment

Example of alignment:

```
ACAATGCAGTGACCCAGCGT----ACGTTAAGATCATG
ACAGTG---TGTCCCAGCCTACACACGT-AAGTTCATG
```

Given a scoring function, the task is to find the alignment that optimises that function.
Vertically aligned bases can be scored, eg, with "log-odds" score

$$s(b_1, b_2) = \log \frac{p(b_1, b_2)}{p(b_1)p(b_2)}$$

For protein alignments there are standard substitution matrices used.
We score the gaps with a "gap penalty": for a gap of length ℓ, this is either a "linear penalty"

$$g = \ell d$$

that is, proportional to the gap, or an "affine penalty"

$$g = d + (\ell - 1)e$$

where the penalty for the first gap, d, is larger than the penalty for increasing the gap, e.

Global alignment: Needleman-Wunsch algorithm

An example of dynamic programming, building up large alignments from shorter ones, O(mn) where m, n are sequence lengths

For two sequences x, y of lengths ℓ_1 and ℓ_2, construct a matrix F_{ij}, in which each element is calculated from an earlier element (with smaller i and/or j).

Each element F_{ij} contains two components:

- The score of the best alignment of the first i bases in sequence 1 and the first j bases in sequence 2. This is

$$F_{ij} = \max \begin{cases} F_{i-1,j-1} + s(x_i, y_j) \\ F_{i-1,j} - d \\ F_{i,j-1} - d \end{cases}$$

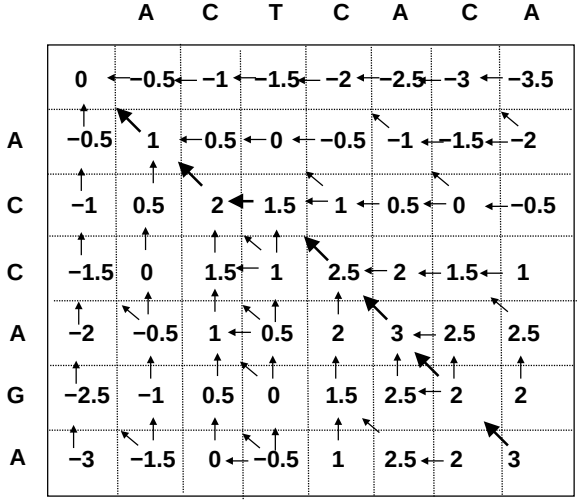
  (for simplicity, we are not considering affine gap penalties, but the algorithm can be extended to include those.)

- A pointer to the previously calculated element of the F matrix from where this element was calculated.

F_{00} is defined to be zero, and "overhanging" bases have the same gap penalty d.
Fill out the entire F matrix, then "backtrace" from the bottom-right element to get the answer.

$$F_{ij} = \max(F_{i-1,j-1} + s(x_i, y_j), F_{i-1,j} - d, F_{i,j-1} - d) \quad + \text{ pointer}$$

Let us align "ACTCACA" and "ACCAGA" with
$s(b_1, b_2) = 1$ if $b_1 = b_2, -1$ otherwise; $d = 0.5$. Here is the F-matrix:



Traceback: bold arrows give alignment
ACTCAGA

Local alignment: Smith-Waterman

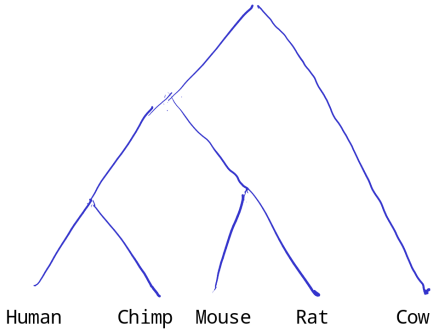
Often we don't want to align the entire sequence, but want to find subsequences of both sequences that align well. This can be done by a simple modification of the above Needleman-Wunsch algorithm:

- Create F-matrix as before
- But now do not let an F_{ij} become negative: if it is calculated as negative, set it to zero without pointers
- Also, keep track of largest F_{ij} element calculated thus far
- After calculating F-matrix, traceback not from bottom-right from largest element, until you hit a zero

For gapless alignments, each element F_{ij} is just calculated from $F_{i-1,j-1}$ and the traceback goes diagonally too. But there are more efficient algorithms for this case.

Phylogenetics

```
Human: AACATGGCACCTGAC
Chimp: AACACGGCACCTAAC
Mouse: AACGCGGAACCCAAC
Rat:   AACGCGGTACCCAAG
Cow:   ATCGCGAAACGGAAC
```



Phylogenetic distance

Given two sequences,

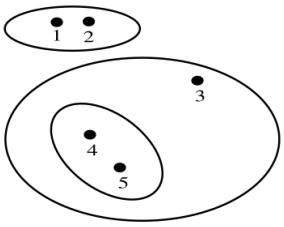
- Crude distance measure: the number of mismatches between them
- Better: estimate the *evolutionary time* when they diverged, assuming a constant mutation rate

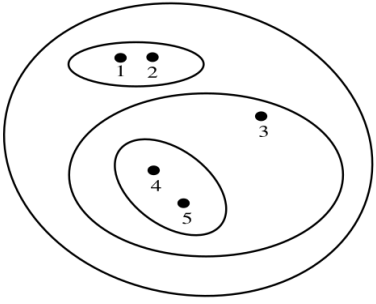
Mutation rates and evolutionary time

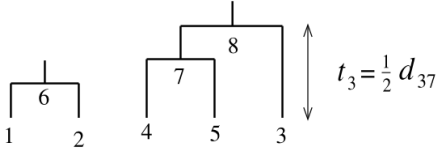
Given an instantaneous rate matrix $R_{\beta\alpha}$ = rate of mutation from nucleotide α to nucleotide β,

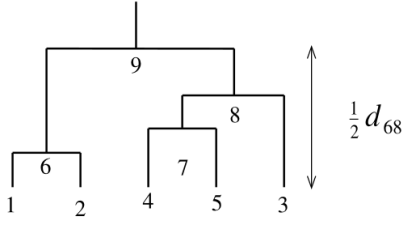
- Calculate the *transition matrix* $T_{\beta\alpha}(t)$ = probability of seeing β at time $T + t$ given α at time T
- Given two sequences, estimate their joint *likelihood* given this matrix
- Estimate their divergence time t by *maximizing this likelihood*

UPGMA: from distance matrix to rooted tree









Other methods

- Neighbour-joining (produces unrooted tree)
- Parsimony
- Maximum likelihood
- Bayesian methods

Given instantaneous rate matrix $R_{\beta\alpha}$, suppose the number of nucleotides α at a given locus in a population of size N is N_α with $\sum_\alpha N_\alpha = N$.
Then

$$\frac{d}{dt}N_\beta = -\sum_{\alpha\neq\beta} R_{\alpha\beta}N_\beta + \sum_{\alpha\neq\beta} R_{\beta\alpha}N_\alpha$$

Defining the diagonal elements of R as $R_{\beta\beta} \equiv -\sum_{\alpha\neq\beta} R_{\alpha\beta}$, this can be written as

$$\frac{d}{dt}N_\beta = \sum_{\alpha} R_{\beta\alpha}N_\alpha$$

which has the solution

$$N_\beta(t) = \sum_\alpha \left(e^{Rt}\right)_{\beta\alpha} N_\alpha(0)$$

Dividing by N and taking the initial population fraction of α to be 1,

$$T_{\beta\alpha}(t) = \left(e^{RT}\right)_{\beta\alpha}$$

.
Different choices of R determine different transition matrices T.

Let the "stationary distribution" at a locus (in the long-time limit) be $\overrightarrow{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$. Then

For time-reversible T

$T_{\alpha\beta}(t)\pi_\beta = T_{\beta\alpha}(t)\pi_\alpha$
("detailed balance")

Let the "stationary distribution" at a locus (in the long-time limit) be $\overrightarrow{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$. Then

For time-reversible T

$T_{\alpha\beta}(t)\pi_\beta = T_{\beta\alpha}(t)\pi_\alpha$
("detailed balance")

Multiplicativity

$\sum_\beta T_{\beta\alpha}(t_1) T_{\gamma\beta}(t_2) = T_{\gamma\alpha}(t_1 + t_2)$

Let the "stationary distribution" at a locus (in the long-time limit) be $\overrightarrow{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$. Then

For time-reversible T

$T_{\alpha\beta}(t)\pi_\beta = T_{\beta\alpha}(t)\pi_\alpha$
("detailed balance")

Multiplicativity

$\sum_\beta T_{\beta\alpha}(t_1) T_{\gamma\beta}(t_2) = T_{\gamma\alpha}(t_1 + t_2)$

"Pulley principle" (Felsenstein)

The root of a rooted tree is not uniquely determined, only the sum of branch-lengths to its children is determined.

Choices of R

- Jukes-Cantor (1969): all off-diagonal elements of R are equal, $R_{\beta\alpha} = \mu$ for $\beta \neq \alpha$, $R_{\beta\beta} = -3\mu$
- Kimura (1980): different transition and transversion rates; $R_{\beta\alpha} = \mu$ for A$\leftrightarrow$G, C$\leftrightarrow$T, $= \nu$ for other $\beta \neq \alpha$, $R_{\beta\beta} = -2\nu - \mu$
- Both Jukes-Cantor and Kimura result in uniform stationary $\vec{\pi} = (0.25, 0.25, 0.25, 0.25)$

Models yielding non-uniform $\vec{\pi}$

Felsenstein 1981 (F81) model

$$R = \mu \begin{pmatrix} \pi_A - 1 & \pi_A & \pi_A & \pi_A \\ \pi_C & \pi_C - 1 & \pi_C & \pi_C \\ \pi_G & \pi_G & \pi_G - 1 & \pi_G \\ \pi_T & \pi_T & \pi_T & \pi_T - 1 \end{pmatrix}$$

$$T_{\alpha\beta}(t) = e^{-\mu t}\delta_{\alpha\beta} + (1 - e^{-\mu t})\pi_\alpha.$$

We use units with $\mu = 1$ and define $q = e^{-t}$, then

$$T_{\alpha\beta}(q) = q\delta_{\alpha\beta} + (1 - q)\pi_\alpha.$$

Models yielding non-uniform $\vec{\pi}$

Hasegawa, Kishino, Yano 1985 (HKY85) model

$$R = \mu \begin{pmatrix} * & \pi_A & \kappa\pi_A & \pi_A \\ \pi_C & * & \pi_C & \kappa\pi_C \\ \kappa\pi_G & \pi_G & * & \pi_G \\ \pi_T & \kappa\pi_T & \pi_T & * \end{pmatrix}$$

where the diagonal elements are such that columns sum to zero. We choose time units such that $\mu = 1$ and, as before, define proximities as $q = e^{-t}$ (however, these no longer have the simple interpretation of probability of conservation as in the F81 model).

Defining $\pi_R = \pi_A + \pi_G$; $\pi_Y = \pi_C + \pi_T$; $e_2 = e^{-t} \equiv q$; $e_3 = q^{\kappa\pi_R + \pi_Y}$; and $e_4 = q^{\kappa\pi_Y + \pi_R}$

Transition probabilities for HKY85

Identity for purines ($\beta = \alpha$ for $\alpha = A, G$):

$$T_{\beta\alpha} = \pi_\alpha + \pi_\alpha \frac{\pi_Y}{\pi_R} e_2 + \left(1 - \frac{\pi_\alpha}{\pi_R}\right) e_3$$

Identity for pyrimidines ($\beta = \alpha$ for $\alpha = C, T$):

$$T_{\beta\alpha} = \pi_\alpha + \pi_\alpha \frac{\pi_R}{\pi_Y} e_2 + \left(1 - \frac{\pi_\alpha}{\pi_Y}\right) e_4$$

Transition, purine ($\beta = A, \alpha = G$ or vice versa):

$$T_{\beta\alpha} = \pi_\beta \left(1 + \frac{\pi_Y}{\pi_R} e_2 - \frac{1}{\pi_R} e_3\right)$$

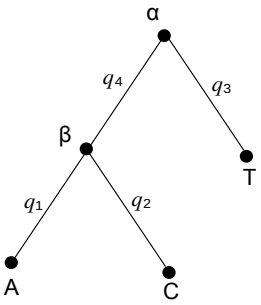
Transition, pyrimidine ($\beta = C, \alpha = T$ or vice versa)::

$$T_{\beta\alpha} = \pi_\beta \left(1 + \frac{\pi_R}{\pi_Y} e_2 - \frac{1}{\pi_Y} e_4\right)$$

Transversion (all other cases):

$$T_{\beta\alpha} = \pi_\beta (1 - e_2)$$

Likelihood calculation



The likelihood, at a particular locus x, of seeing the nucleotides $\{S^k\}$ at the leaves (k denotes the leaf or species label) is a product over the tree of all transition matrices along edges, summed over all unknown ancestors. In this case S^k = A, C, T for k = 1, 2, 3, and

$$P(\{S^k\}|\{q\}) = \sum_{\alpha} \pi_\alpha T_{T\alpha}(q_3) \sum_{\beta} T_{\beta\alpha}(q_4) T_{A\beta}(q_1) T_{C\beta}(q_2)$$

Likelihood calculation: Felsenstein algorithm

(assuming binary tree, can be generalized)

- Define $L_i(\alpha)$ = likelihood of leaves below node i given that the nucleotide at node i is α
- If i is a leaf node with nucleotide x, set $L_i(\alpha) = \delta_{\alpha x}$
- Otherwise, let the two children of i be j and k with proximities q_j and q_k. Then $L_i(\alpha) = \sum_\beta T_{\beta\alpha}(q_j) L_j(\beta) \sum_\gamma T_{\gamma\alpha}(q_k) L_k(\gamma)$. (For non-binary trees this step can be generalized to a product over all children, with the appropriate number of sums.)
- Termination: let the root node be $2n-1$, then the likelihood of the leaves is $\sum_\alpha L_{2n-1}(\alpha)\pi_\alpha$.

Complexity: $O(NL\Sigma^2)$ where N = number of species, L = number of sites, Σ = size of alphabet

The intermediate calculations should be cached.

Inferring a phylogenetic tree from sequence data

Two tasks

- Infer *topology* of tree
- Infer *branch lengths* of tree

Approach

- Markov Chain Monte Carlo:
  - Sample space of tree topologies
  - For each topology, calculate branch lengths as per maximum likelihood
  - Use likelihood to accept/reject (Metropolis algorithm)

Bayesian phylogenetics

- The *likelihood* of a tree can be related to the *posterior probability* via Bayes' theorem:

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}$$

  where the denominator is a constant.
- If the *prior* $P(T)$ is a constant then sampling from a likelihood is equivalent to sampling from a posterior.
- But we can use a prior to favour probable trees, disfavour improbable trees, rule out impossible trees.

Linguistic phylogenetics

- Nucleotides $\rightarrow$ words
- Homologous nucleotides $\rightarrow$ cognate words
- Likelihoods $\rightarrow$?

Literature examples

NATURE | VOL 426 | 27 NOVEMBER 2003

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is "the most intensively studied, yet still most recalcitrant, problem of historical linguistics"[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the 'Kurgan expansion' and the 'Anatolian farming' hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.

Source: *Language*, Vol. 91, No. 1 (MARCH 2015), pp. 194-244

ANCESTRY-CONSTRAINED PHYLOGENETIC ANALYSIS SUPPORTS THE INDO-EUROPEAN STEPPE HYPOTHESIS

WILL CHANG
University of California, Berkeley

CHUNDRA CATHCART
University of California, Berkeley

DAVID HALL
University of California, Berkeley

ANDREW GARRETT
University of California, Berkeley

Discussion of Indo-European origins and dispersal focuses on two hypotheses. Qualitative evidence from reconstructed vocabulary and correlations with archaeological data suggest that Indo-European languages originated in the Pontic-Caspian steppe and spread together with cultural innovations associated with pastoralism, beginning c. 6500–5500 BP. An alternative hypothesis, according to which Indo-European languages spread with the diffusion of farming from Anatolia, beginning c. 9500–8000 BP, is supported by statistical phylogenetic and phylogeographic analyses of lexical traits. The time and place of the Indo-European ancestor language therefore remain disputed. Here we present a phylogenetic analysis in which ancestry constraints permit more accurate inference of rates of change, based on observed changes between ancient or medieval languages and their modern descendants, and we show that the result strongly supports the steppe hypothesis. Positing ancestry constraints also reveals that homoplasy is common in lexical traits, contrary to the assumptions of previous work. We show that lexical traits undergo recurrent evolution due to recurring patterns of semantic and morphological change.*

Literature examples

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Kolipakam V, Jordan FM, Dunn M, Greenhill SJ, Bouckaert R, Gray RD, Verkerk A. 2018 A Bayesian phylogenetic study of the Dravidian language family. *R. Soc. open sci.* **5**: 171504.
http://dx.doi.org/10.1098/rsos.171504

A Bayesian phylogenetic study of the Dravidian language family

Vishnupriya Kolipakam[1,2], Fiona M. Jordan[2,3,4], Michael Dunn[2,5], Simon J. Greenhill[4,6], Remco Bouckaert[4,7], Russell D. Gray[4] and Annemarie Verkerk[2,4]

[1]Wildlife Institute of India, Post Box 18, Chandrabani, Dehradun 248001, India
[2]Evolutionary Processes in Language and Culture, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

Opinion TRENDS in Ecology and Evolution Vol.20 No.3 March 2005

A phylogenetic approach to cultural evolution

Ruth Mace and Clare J. Holden

Department of Anthropology, University College London, Gower Street, London, UK, WC1E 6BT

There has been a rapid increase in the use of phylogenetic methods to study the evolution of languages and culture. Languages fit a tree model of evolution well, at least in their basic vocabulary, challenging the view that blending, or admixture among neighbouring groups, was predominant in cultural history. Here, we argue that we can use language trees to test hypotheses about not only cultural history and diversification, but also bio-cultural adaptation. Phylogenetic comparative methods take account of the non-independence of cultures (Galton's problem), which can cause spurious statistical associations in comparative analyses. Advances in phylogenetic methods offer new possibilities for the analysis of cultural evolution, including estimating the rate of evolution and the direction of coevolutionary change of traits on the tree. They also enable phylogenetic uncertainty to be incorporated into the analyses, so that one does not have to treat phylogenetic trees as if they were known without error.

experimentation are limited, but also because humans show such a remarkable range of cross-cultural variation.

Cultures as species

We define culture broadly, as behavioural traditions that are transmitted by social learning. At the population level, humans structure themselves into cultures or ethno-linguistic groups, which we define here as a group of people who speak the same language. Many parallels have been drawn between cultural and biological evolution, both at the level of parallels between genes and cultural traits (or variants), and at the level of species and cultures [5]. Culture evolves in the sense that occasional errors arise in cultural transmission (equivalent to mutations in biological evolution), leading to change through time [6,7].

For the purposes of phylogenetic analysis, languages and cultures are treated as being analogous to species (Table 1), although there has been a vigorous debate about how far we can treat cultures as discrete, bounded units,

Literature examples

RESEARCH ARTICLE

BEASTling: A software tool for linguistic phylogenetics using BEAST 2

Luke Maurits[1]*, **Robert Forkel**[2], **Gereon A. Kaiping**[3], **Quentin D. Atkinson**[1,2]

1 School of Psychology, University of Auckland, Auckland, New Zealand, **2** Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany, **3** Leiden University Centre for Linguistics, Leiden University, Leiden, the Netherlands

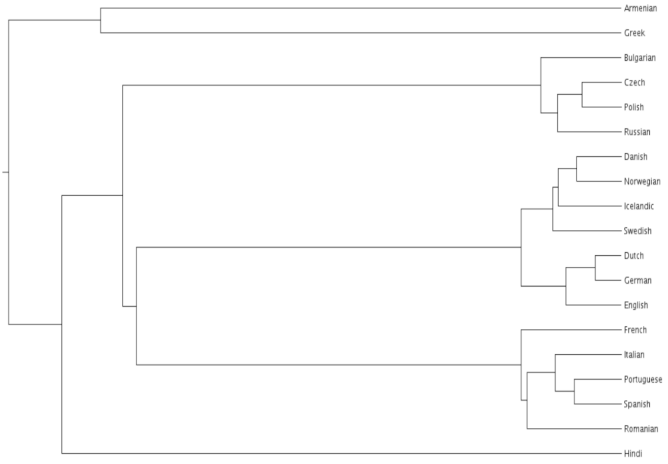
* luke@maurits.id.au

Abstract

We present a new open source software tool called BEASTling, designed to simplify the preparation of Bayesian phylogenetic analyses of linguistic data using the BEAST 2 platform. BEASTling transforms comparatively short and human-readable configuration files into the XML files used by BEAST to specify analyses. By taking advantage of Creative Commons-licensed data from the Glottolog language catalog, BEASTling allows the user to conveniently filter datasets using names for recognised language families, to impose monophyly constraints so that inferred language trees are backward compatible with Glottolog

Beastling sample output

Language_ID	Feature_ID	IPA	Value
Dutch	all	ɑlə	5
Swedish	all	alː	5
English	all	ɔːl	5
German	all	al	5
Spanish	all	entero	2
Danish	all	æˈl	5
Bulgarian	all	vsiɡki	3
Greek	all	kaθe	7
Romanian	all	tot	6
Russian	all	fsʲe	3
Polish	all	fʃistʃi	3
Norwegian	all	ɑlə	5
French	all	tu	6
Czech	all	fʃɪxnʲɪ	3
Icelandic	all	aɬlir	5
Italian	all	intɛro	2
Hindi	all	sab	4
Portuguese	all	todu	6
Spanish	all	todo	6
Romanian	all	ɨntreg	2
French	all	œʃje	2
Italian	all	tutto	6
Hindi	all	saːraː	1
Dutch	ashes	ɑs	10
Swedish	ashes	asːka	10
English	ashes	æʃ	10
German	ashes	aʃə	10
Spanish	ashes	θeniθa	12
Danish	ashes	asg	10
Bulgarian	ashes	pɛpel	8
Greek	ashes	staxti	14
Armenian	ashes	mɔxir	9
Romanian	ashes	ʧenuʃə	12
Russian	ashes	zela	11
Polish	ashes	pɔpʲuw	8



Caveats

- Phylogenetic trees in biology are sensitive to choice of sequence, similar concerns may apply to linguistic trees
- "Selection pressures" vary on different parts of the genome, and likely for different words in different languages
- Horizontal transfer is common among languages
- Algorithms should never be used as a black box!

Thank you