From alignment of etymological data to phylogenetic inference via population genetics

Javad Nouri[1] **Jukka Sirén**[3] **Jukka Corander**[2,4] **Roman Yangarber**[1]
[1]Department of Computer Science
[2]Department of Mathematics and Statistics
[3]Department of Biosciences
University of Helsinki, Finland
[4]Department of Biostatistics,
University of Oslo, Norway
first.last@helsinki.fi

Abstract

This paper presents a method for linking models for aligning linguistic etymological data with models for phylogenetic inference from population genetics. We begin with a large database of genetically related words—sets of cognates—from languages in a language family. We process the cognate sets to obtain a complete alignment of the data. We use the alignments as input to a model developed for phylogenetic reconstruction in population genetics. This is achieved via a natural novel projection of the linguistic data onto genetic primitives. As a result, we induce phylogenies based on aligned linguistic data. We place the method in the context of those reported in the literature, and illustrate its operation on data from the Uralic language family, which results in family trees that are very close to the "true" (expected) phylogenies.

1 Introduction

Recently, mathematical theory of statistical physics has been shown to unite stochastic *models* of evolution in seemingly diverse fields, such as population genetics, ecology and linguistics (Blythe and McKane, 2007; Blythe, 2009; Baxter et al., 2009; Vázquez et al., 2010). However, statistical *inference* about language evolution under such models is complicated by the practically intractable form of likelihoods for even a moderate set of languages. This calls for novel ways to probabilistic evaluation of any particular phylogenetic model and for learning the most plausible genealogies from data. In the context of population genetics, such an approach is introduced in (Sirén et al., 2011; Sirén et al., 2013) by combining diffusion-based approximations of conditional distributions with adaptive Monte Carlo methods. In contrast to coalescent-based likelihoods, this approach enables analysis of much larger data collections, as the sufficient statistics from the data correspond under these models to the empirical allele frequencies of each population, rather than genetic characteristics of single individuals. This property makes these models attractive from the perspective of evolutionary linguistics.

The field of evolutionary linguistics, or computational etymology, addresses a range of problems, including: automatic identification of sets of cognates—genetically related words; finding genetic relations across languages in a language family; finding patterns of recurrent sound correspondence among groups of languages; reconstruction of proto-forms in ancestral (usually unobserved) languages; etc. These problems are interdependent. When approached by traditional methods, work proceeds in cycles, through iterative refinement via the *comparative method*. In our work, we take sets of cognate words as given, and focus on the problems of genetic relations and patterns of correspondence. The problem of reconstruction is also addressed, indirectly.

Based on automatically derived *pairwise* correspondences among the languages in a given corpus of cognate sets[1]—we aim to determine the overall structure of the language family. To find the correspondences, we try to find the best alignment of the complete data at the level of individual sounds—or, equivalently, symbols, since we assume that our data is phonetically transcribed.

An important aspect of our approach is that we aim to use all available data—to avoid subjective

[1]The creators of the input dataset posit that the elements of a cognate set derive from a common origin—a word in the ancestral proto-language.

Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, pages 27–37,
Berlin, Germany, August 11, 2016.

bias, which would be inherent in selecting some subset of available data, as is sometimes done with short 50- to 200-word lists. We learn *patterns of correspondence* directly from the data, in explicit form. We let only the data determine what rules are inherent in it; i.e., we look for correspondences that are inherently encoded in a given dataset—rather than relying on externally supplied (and possibly biased) assumptions or "priors." The models assume no *a priori* knowledge or "universal" principles—e.g., no preference for aligning a symbol with itself, aligning a vowel with a vowel rather than a consonant, etc.

The main idea of the approach we are exploring here—summarized in Figure 1—is to create a bridge between the two domains: on the linguistic side, alignment of etymological data, and on the population-genetics side, phylogenetic inference. The two domains operate on different kinds of objects: in linguistics we have languages, words and sounds, whereas in genetics we have populations, individuals, and their DNA sequences, and although there are apparent similarities, it is not obvious how these can be combined. In Section 4 we formalize the problem of alignment and present some details about the alignment models we use—step B in the figure. Section 6 describes our population-genetics model for phylogenetic inference (step D). Section 5 shows how we can "glue" these two together, by means of a cross-domain projection—mapping information obtained from linguistic alignments into a form usable in population genetics (step C). In Section 7 we present some results from the combined approach, which involves building pairwise distance matrices and constructing phylogenetic trees (steps E–F). The resulting trees are compared to manually-constructed gold standards, to get an estimate of the quality of the inference pipeline.

Building phylogenetic trees by applying models from population genetics to an *alignment* of a language family has not been attempted previously, to our knowledge. In section 2 we review several approaches to etymological alignment from the last decade, and describe the data we use in our experiments, in Section 3. We conclude with a discussion and current work, in Section 8.

2 Related Work

The last 15 years have seen a surge in interest in computational modeling of language relationships, change and evolution. We have been developing a family of models for this task, called the *Etymon* models, (Wettig et al., 2011; Wettig et al., 2012; Nouri and Yangarber, 2016), etc.[2]

Methods introduced in (Kondrak, 2002), inspired by alignment in machine translation, learn one-to-one sound correspondences between words in pairs of languages. Kondrak (2003), and Wettig et al. (2011) find more complex—many-to-many—sound correspondences. These methods focus on alignment. They model the context of the sound changes in a limited way, while it is known that most evolutionary changes are conditioned on the context of the evolving sound. Bouchard-Côté et al. (2007) propose MCMC-based methods to model context, and operate on more than one pair of languages at a time.[3]

The Etymon models, similarly to other work, operate at the phonetic level only, leaving semantic judgements to the creators of the input databases. Some prior work has attempted to approach semantics by computational means as well. We do not address the problem of discovering cognates; this problem is attempted, e.g., in, (Kondrak, 2004; Kessler, 2001; Steiner et al., 2011) and semi-automatically in (Bouchard-Côté et al., 2007). Our Etymon models begin with a set of etymological data (or more than one such set) for a language family as given, and treat the given cognate set as a fundamental unit of input. We use the principle of *recurrent sound correspondence*, as in much of the literature, including (Kondrak, 2002; Kondrak, 2003), and others.

One approach to evaluating our alignment models, is to try to infer relationships among entire languages within the family. Construction of phylogenies is studied extensively, e.g., by (Nakhleh et al., 2005; Ringe et al., 2002; Barbançon et al., 2009). This work differs from ours in that it operates on manually pre-compiled sets of *characters*. Each character is a distinctive feature of languages, which takes on different values among different languages within the family. All Etymon models operate at the level of sounds within words and cognate sets.

There is extensive work on alignment in the machine-translation (MT) literature, with some

[2]Please see *http://etymon.cs.helsinki.fi/* for the publicly available software packages.

[3]The running time did not scale well when the number of languages was above three; (Bouchard-Côté et al., 2009) describe improved models to align multiple languages.

Figure 1: Outline of the components in the inference pipeline

methods from MT alignment projected onto alignment in etymology. The intuition is that sentences that are translation of each other in MT correspond to cognate words in etymology, and words in MT correspond to sounds in etymology. The notion of regularity of sound change in etymology, which is what our models try to capture, is loosely similar to contextually conditioned correspondence of translation words across languages. For example, (Kondrak, 2002) employs MT alignment from (Melamed, 1997; Melamed, 2000). One might employ the IBM models for MT alignment, (Brown et al., 1993), or the HMM model, (Vogel et al., 1996). Among the MT-related models, (Bodrumlu et al., 2009) is similar to ours in that it is based on MDL, the Minimum Description Length principle. There are important differences between our alignment problem vs. alignment in MT. Evolutionary sound correspondence is conditioned by local context, whereas in MT correspondences may depend on much wider context. There is no analogue to the notion of phonetic *features* in MT. Phonetic correspondences in etymological data—which apply throughout the language—have no analogue in semantic shift processes in a such way as to be captured by MT alignment models. Neither are phonetic features used in the aforementioned work from the area of automatic transliteration, e.g., (Zelenko, 2009).

Our work on the Etymon models is closely related to a series of generative models in (Bouchard-Côté et al., 2007) through (Hall and Klein, 2011), in the following respects.

In (Wettig et al., 2011) some context is modeled in the form of coding pairs of symbols, as in (Kondrak, 2003). Bouchard-Côté et al. (2007) and Hall and Klein (2011) handle context by conditioning the symbol being generated upon the symbols immediately preceding and following. Wettig et al. (2012) and Nouri and Yangarber (2016) use much broader context by building decision trees, so that non-relevant context information does not grow model complexity.

In (Wettig et al., 2011) sounds / symbols are treated as atomic—not analyzed in terms of their phonetic makeup. (Bouchard-Côté et al., 2007) recognize "natural classes" in defining the context of a sound change, though not in generating the symbols themselves; (Bouchard-Côté et al., 2009) encode as a prior which sounds are "close" to each other. In (Wettig et al., 2012) and later Etymon models, we code each sound in terms of the individual phonetic features that make up the sound.

Etymon models are based on the information-theoretic MDL principle, e.g., (Grünwald, 2007)—like (Wettig et al., 2011) and unlike (Bouchard-Côté et al., 2007; Hall and Klein, 2011). MDL brings important theoretical benefits, since models chosen in this way are guided by data with no free parameters or hand-picked "priors." The data analyst chooses the model class and structure, and the coding scheme, i.e., a *decodable* way to encode both model and data. This determines the learning strategy—we optimize the cost function, which is the code length determined by these choices.

Objective function: For the objective function to optimize during alignment, we use the prequential code-length (Dawid, 1984), as explained in (Wettig et al., 2011). Normalized Maximum Likelihood (NML) as presented in (Wettig et al., 2012; Nouri and Yangarber, 2016) could be used as an alternative to prequential coding. Although NML reduces the code length, and brings other advantages, it did not have a significant effect on the quality of the alignments required in the experiments presented here.

Some of our work on modeling language change and evolution, (Nouri and Yangarber, 2016) shows that alignment may not be a necessary goal for obtaining efficient compression; in case of models that circumvent alignment, it is less clear how they can be combined with population-genetics models.

Additional prior work related to the population-genetics models is referenced throughout the paper and in Section 6.

3 Data

As we mentioned, we aim to use large-scale etymological databases, rather than small, manually-selected sets of characters of the languages. For

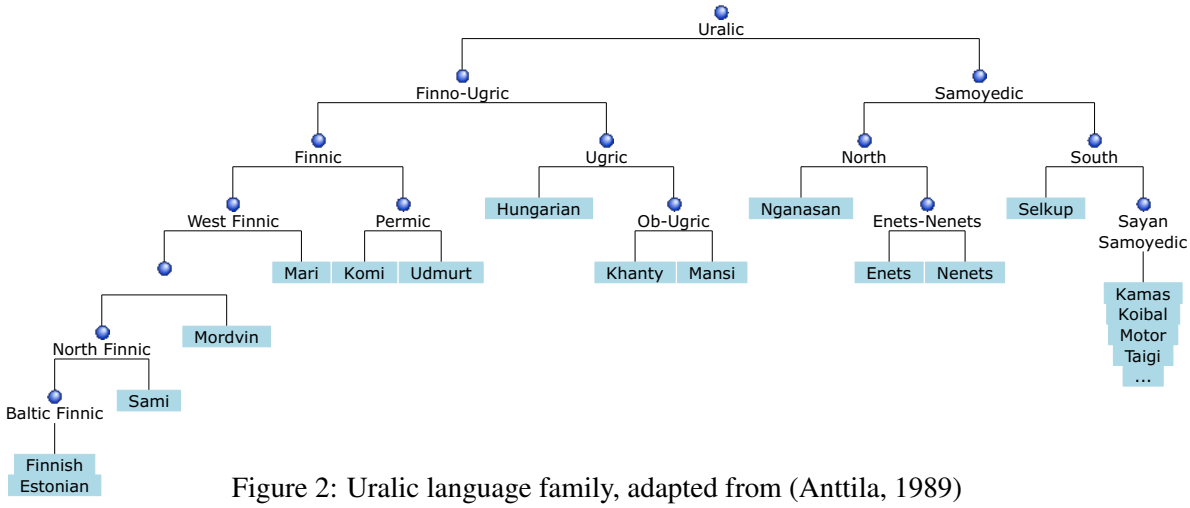

Figure 2: Uralic language family, adapted from (Anttila, 1989)

$*Proto$	k	a	r	.	n	e	$š$		v	e	n	e	$š$
	|	|	|	|	|	|	|		|	|	|	|	|
$Finnish$	k	$ā$	r	.	n	e	.		v	e	n	e	.
	|	|	|	|	|	|	|		|	|	|	|	|
$Mordvin$	k	.	r	e	n	.	$č$		v	e	n	.	$č$

Figure 3: Sample alignments for Finnish and Mordvin: *kaarne/krenč* 'raven', *vene/venč* 'boat', with unobserved, hypothesized proto word-forms

the Uralic language family, we use the StarLing Uralic database, (Starostin, 2005), based on (Rédei, 1991) and expanded. The database contains 2586 Uralic cognate sets. Whereas much of the prior work is based on small manually preselected subsets of the data—so-called "Swadesh lists" of 100 (or 40, 50, etc.) words—we use complete large data sets. In this paper, we focus on a sub-tree of Uralic, viz., the Finno-Ugric subfamily—i.e., excluding the remaining Samoyedic sub-tree of Uralic—which contains most of the extant Uralic data. Our experiments use the 10 "principal" Finno-Ugric languages.[4]

One arrangement of the Uralic languages accepted by some linguists is shown in Figure 2, adapted from Encyclopedia Britannica and (Anttila, 1989). Note, that this is the subject of some debate in modern scholarship, and alternative phylogenies have some acceptance among linguists. Figure 2 shows the phylogeny most widely accepted today. Other theories, e.g., posit a "Volgaic" branch, which groups together Mari with Mordvin languages, where this phylogeny posits Mari on an independent branch, an offshoot from the "West Finnic" subgroup, see, e.g., (Anttila, 1989). We use this phylogeny as a gold-standard in our experiments.

In our experiments we need a measure of distance between phylogenies proposed by different approaches. For comparison, we can treat the phylogenies as *unrooted*, leaf-labeled (URLL) trees. One distance measure for URLL trees is introduced in (Robinson and Foulds, 1981). Based on this particular distance measure, the distance between the gold standard tree and the tree with a Volgaic branch would be 0.14, (see discussion in Section 7).

4 Pairwise Alignment

We use our Etymon models, described in (Wettig et al., 2011; Wettig et al., 2012), for aligning the etymological data.e summarize the main features of these models in this section. We begin with pairwise alignment: aligning words from two languages at a time. For each word pair, the task of alignment means finding exactly which symbols correspond. The simplest form of such alignment at the symbol level is a 1-1 pair $(\sigma : \tau) \in \Sigma \times T$, a single symbol σ from the *source alphabet* Σ with a symbol τ from the *target alphabet* T. We denote the sizes of the alphabets by $|\Sigma|$ and $|T|$.

To model *insertions* and *deletions*, we augment both alphabets with a special empty symbol—denoted by a dot—and write the augmented alphabets as $\Sigma^{\cdot}$ and $T^{\cdot}$. We can then align word pairs,

[4]The 10 Finno-Ugric languages used in the experiments are: est=Estonian, fin=Finnish, khn=Khanty, kom=Komi, man=Mansi, mar=Mari, mrd=Mordvin, saa=Saami, udm=Udmurt, unk/ugr=Hungarian. The StarLing database also contains data on dialects for the 7 languages excluding {fin, est, unk}; in the figures, the suffix after the code identifies the *principal* dialect—having the largest number of entries in StarLing. Some of these dialects are quite far apart; in other experiments we also use the second-largest dialects, giving 17 languages in total.

such as *vene—venč* ("boat" in Finnish and Mordvin), in many ways, including, e.g., as in Figure 3, where the alignment on the right contains symbol pairs: $(v:v), (e:e), (n:n), (e:.), (.:\check{c})$. Note that, since the Proto language is not observed, the alignment model might actually prefer to align (e:č) in these examples, especially if this pattern appears several times (which it does)—since there is no *a priori* penalty for vowel-consonant alignment, as mentioned in the Introduction.

If we align all languages simultaneously, rather than pairwise, there may be additional information in *other* languages (which there is), that may help the model disfavor (e:č). N-way alignment will be revisited in the conclusion.

According to the MDL Principle, the aim is to code these aligned word pairs as compactly as possible. To construct such a code, we "transmit" the aligned data by listing the "events"—the observed symbol pairs $(\sigma : \tau)$. Since the code needs to be uniquely decodable, after each word pair we transmit a special event $(\# : \#)$ to mark the word boundaries. The code length (or cost) for the *complete*, aligned data is our objective function that the algorithm optimizes. Lower code-length means that the algorithm has found a way of aligning the data that is more compact, i.e., it has discovered more regularity in the data.

Using prequential coding, or Bayesian Marginal Likelihood, the total cost of coding the aligned data is given by:

$$L(D) = \tag{1}$$
$$-\sum_{e\in E}\log\Gamma\big(C(e)+\alpha(e)\big)+\sum_{e\in E}\log\Gamma\big(\alpha(e)\big)$$
$$+\log\Gamma\left[\sum_{e\in E}\big(C(e)+\alpha(e)\big)\right]-\log\Gamma\left[\sum_{e\in E}\alpha(e)\right]$$

where $E = \Sigma^{\cdot}\times T^{\cdot}\cup\{(\# : \#)\}$ is the event space, $C(e)$ stores the number of times event e occurs in the complete alignment, and $\alpha(e) = 1$ are the uniform Dirichlet priors.

Learning the model from the observed data now means iteratively re-aligning word pairs, and updating the matrix C, which stores the counts of all observed alignment events. The sparser C becomes, the lower the code-length will be.

Summary of the Algorithm: We start with an initial *random* alignment for each pair of words in the corpus. We then alternate between two steps: **A.** update the count matrix and compute the code length, and **B.** re-align all word pairs in the corpus, using dynamic-programming re-alignment. During the dynamic-programming step, for each word pair we find the best alignment, i.e., the one with the lowest cost given the alignments for rest of the words. The algorithm is described in detail in (Wettig et al., 2011).

The algorithm is similar to Expectation-Maximization (EM), but is in fact greedy. The iterative steps monotonically decrease the cost function, and thus compress the data. We continue until we reach convergence. To avoid local optima, we use Simulated Annealing.

5 Projection

To be able to apply phylogenetic reconstruction methods from population genetics we need to define appropriate analogues for the notions of *population*, *individual*, *locus*, and *allele*, which are the essential inputs to the population genetics models, described in the next section.

It is natural to identify population with *language*, and individuals with *words* in the language. Next, suppose that the proto-language L^* (the root of the family tree) had been fully observed, as in Figure 3. Then, for any leaf language L_i, we could align L_i to L^* (pairwise). We could then fix the set of sounds of L^* as the set of "*loci*" (sites) in the "DNA" of the individuals. We treat each sound s of L^* as a locus, in the sense that from the complete alignment from L_i to L^* we can observe the distribution of sounds (from L_i's alphabet) that were aligned to s. Thus, the *alleles* are the various sounds (in L_i's alphabet) which appear aligned to s in the words in L_i. Each L_i will have its distinctive distribution of alleles at each locus. Thus, in the Mordvin examples in Figure 3, at the "locus" labeled *e* in the Proto-language, we would observe the "allele" *e* once, and the allele *dot* twice.

However, in general, we have no access to L^*, and we proceed indirectly as follows. Suppose, for instance, $\{L_i\}$ are the 10 languages from the Finno-Ugric sub-family of Uralic. We designate each L_i, in turn, as a *reference* language—in place of the unobserved L^*. The reference L_i "donates" its sounds as the loci, to be aligned to each of the remaining 9 (*target*) languages. As before (with L^*), at each site, a target population L_j has a distinctive distribution over the *alleles*—symbols drawn from the **universal** phonetic alphabet, which is simply the union of the individual al-

phabets. In this way, each reference language L_i induces one dataset D^{L_i} of allele distributions in the remaining 9 populations, giving a total of 10 input datasets. These datasets are processed by the population genetics model introduced below.

Although "sacrificing" the reference language in this way skews the dataset, we compensate for this by averaging the estimated pairwise distances over *all* 10 datasets $\{D^{L_i}\}$. When we calculate the distances of languages based on a single reference, there will be a higher level of variance in the estimates and as a consequence Neighbor-Join and similar algorithms can easily lead to incorrect trees. When we instead calculate the average distance for any pair of languages (L_i, L_j) over the 8 remaining references, the variance in the estimates stabilizes (because the mean distance estimate will be much less variable) and consequently the NeighborJoin algorithm shows more accurate performance. To verify empirically these basic statistical arguments—that using the mean distances is more stable than any single estimate—we ran simulations with artificial data sets (Figure 4). In the simulation we perturb the pairwise distances with Normal noise, using mean 0 and σ as shown on the X-axis. The upper curve is the (average) URLL distance from trees built on single estimates to the gold-standard tree in Figure 2; the lower curve is the URLL distance from the tree based on the mean of the estimates to the gold-standard tree. The figures confirm the higher stability of the mean (of 8 estimates in A, 15 estimates in B), as compared to any single estimate, which is according to the expectations. In addition, there may be a small effect caused by the fact that some reference language can produce slightly better results than another, but the main effect should be the one explained above.

6 Population genetics model

With this definition of population, individual, locus, and allele, we proceed to the method for building the phylogenetic tree based on each complete aligned data set. Below we introduce expressions for conditional distributions that jointly determine a hierarchical probability model for the count data derived from the alignment. The model reflects the degree of relatedness among the languages through a tree topology and the corresponding branch length parameters.[5] We consider modeling the relatedness of K languages by a rooted bifurcating tree topology T representing the order of divergence from a common ancestral language. The leaves of the topology T correspond to the K modern (observed) languages, whereas the inner nodes correspond to ancestral (unobserved) languages. The length of each branch c of T is a parameter to be inferred from the output of the alignment algorithm using the introduced two-part coding approach. Our Beta-Dirichlet model describes stochastic changes in the alignment patterns of loci by separating the shared alleles S among two or more languages from those that are present in a single language only (private alleles P). From the perspective of genetics, the latter correspond to novel mutations that arise over time in any particular population and are not observed elsewhere. For a locus, the conditional distribution of alleles for a node c of T, either observed or ancestral, is determined by the relative frequencies ψ_{Sc} and ψ_{Pc} of values in S and P, respectively. Here $\psi_{Sc} = (\psi_{Sc1}, \ldots, \psi_{Scr})$ is a vector of relative frequencies for the r alleles in S and ψ_{Pc} is a scalar of the total relative frequency of alleles in P, so that $\psi_{Pc} + \sum_{j=1}^{r} \psi_{Scj} = 1$. By definition, ψ_{Pc_a} equals zero for the root node c_a.

For each node c except the root, the conditional distribution of the relative frequency of the values in the private set ψ_{Pc} given the relative frequency $\psi_{Ppa(c)}$ in the parent node $pa(c)$ is defined as the Beta distribution:

$$\psi_{Pc} \mid \psi_{Ppa(c)} \sim Beta(\phi_{Pc}\mu_{Pc}, \phi_{Pc}(1-\mu_{Pc})) \quad (2)$$

where μ_{Pc} corresponds to the mean of the distribution and ϕ_{Pc} determines the variance, given by

$$Var(\psi_{Pc}) = \frac{\mu_{Pc}(1-\mu_{Pc})}{\phi_{Pc}+1}$$

The relative frequencies of the shared features ψ_{Sc} have the conditional distribution:

$$(1-\psi_{Pc})^{-1}\psi_{Sc} \mid \psi_{Pc}, \psi_{Ppa(c)}, \psi_{Spa(c)} \sim \\ \sim Dirichlet(\phi_{Sc}\mu_{Sc1}, \ldots, \phi_{Sc}\mu_{Scr}) \quad (3)$$

where again μ_{Scj} and ϕ_{Sc} control the first two central moments of the distribution.

[5] The underlying theory relies on concepts from theoretical population genetics, (Ewens, 2004; Blythe and McKane, 2007); the reader may refer also to (Sirén et al., 2011; Sirén et al., 2013), for a detailed account of the model structure.

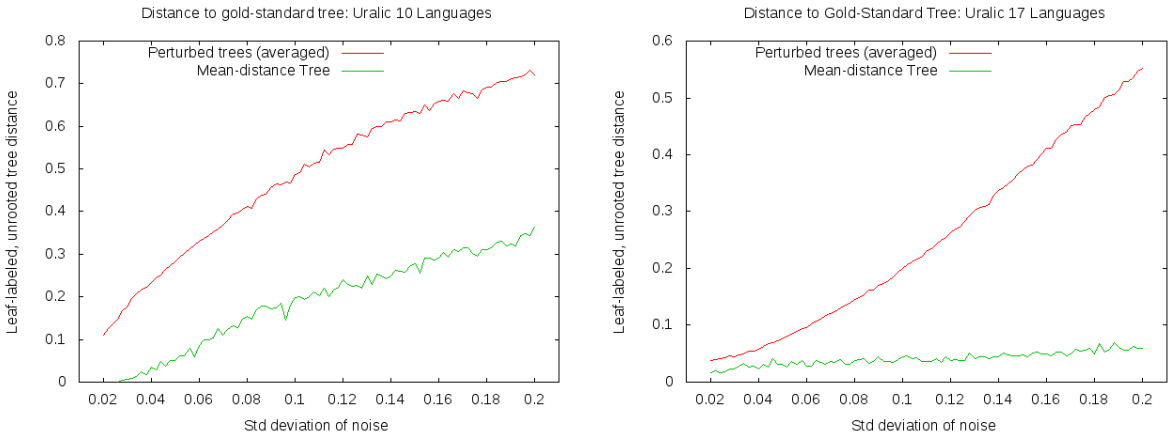

Figure 4: Stability of phylogeny based on sample means of pairwise distances vs. individual samples: (A) for 10 Uralic languages; (B) for 17 Uralic languages

We choose parameters of the two distributions as

$$\mu_{Pc} = 1 - e^{-m_c\tau_c}(1 - \psi_{Ppa(c)}) \quad (4)$$

$$\mu_{Scj} = \frac{\psi_{Spa(c)j}}{1 - \psi_{Ppa(c)}} \quad (5)$$

$$\phi_{Pc} = \frac{\mu_{Pc}}{\frac{(1-e^{-(m_c+1)\tau_c})}{(m_c+1)} - (1 - \mu_{Pc})(1 - e^{-\tau_c})} - 1 \quad (6)$$

$$\phi_{Sc} = \frac{(m_c + 1)(1 - \mu_{Pc})e^{-\tau_c}}{1 - e^{-(m_c+1)\tau_c}} \quad (7)$$

to yield the same expectation and covariance structure as obtained under the Wright-Fisher infinite alleles model (Sirén et al., 2013; Ewens, 2004). The parameter τ_c represents the relative time between a node and its ancestral language and m_c is an effective mutation parameter in the branch connecting c and $pa(c)$. For the relative frequencies ψ_{Sc_a} in the root node c_a, a uniform distribution is assumed in the model. Assuming conditional independence of all loci for which count data is derived in the alignment, a product multinomial distribution is obtained for the feature counts conditionally on the unknown relative frequency parameters, such that

$$p(\mathbf{x}|\psi) = \prod_{l=1}^{L}\prod_{c=1}^{K} p(\mathbf{x}_l^{(c)}|\psi_{lPc}, \psi_{lSc}), \quad (8)$$

where $p(\mathbf{x}_l^{(c)}|\psi_{lPc}, \psi_{lSc})$ is the joint multinomial probability of the feature counts $\mathbf{x}_l^{(c)}$ for the locus l in language c, where the relative frequencies are now indexed. Notice that the remaining parameters in 2 and 3 are set to be constant over the loci, thus representing the average tendency over the loci.

In our fully Bayesian probabilistic formulation, prior distributions are assigned to all the unknown parameters. Similar to (Sirén et al., 2013), here we have used uniform distributions on the interval $(0, 1)$ for the time parameters τ and exponential distributions with mean 1 for the relative mutation parameters m. As in Bayesian phylogenetics in general, the tree topologies are assigned a uniform prior distribution. These choices have been made to specify vaguely informative prior distributions which should not have any considerable effect on the resulting posterior inferences.

Using the implementation from (Sirén et al., 2013), the Adaptive Metropolis (AM) algorithm, (Haario et al., 2001) can be applied to generate samples from the conditional posterior distribution of τ, m and ψ, given a topology T and the partition of the features to sets P and S. In our MCMC simulations we used 100000 iterations in total, out of which the initial sequence of 20000 iterations was discarded as burn-in and the chain was thinned by including every 8th iteration in the final sample. This resulted in posterior samples of size 10000 values.

Here, the AM algorithm is first used to generate the posterior samples separately for each pair of languages in a given alignment, which allows us to compute the distance between the two languages as the sum of relative times τ since the divergence from a common ancestral language. Then, we construct the tree topology corresponding to the particular alignment by finding the unrooted binary tree using the neighbour joining al-

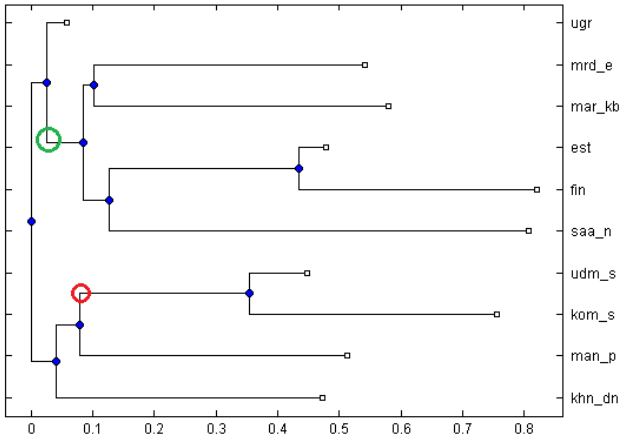

Figure 5: Phylogenetic (unrooted) tree computed by NeighborJoin, using pairwise distances averaged over 10 Uralic datasets.

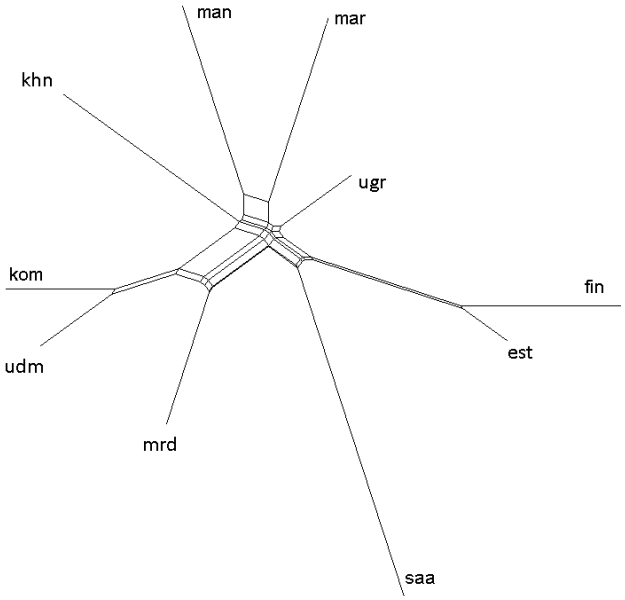

Figure 6: Phylogenetic network computed by NeighborNet, using same datasets.

	D(T,G)	Count	% of Total
	0.000000	1	0.0000
	0.142857	14	0.0007
	0.285714	126	0.0062
	0.428571	1018	0.0502
	0.571429	8114	0.4003
	0.714286	60444	2.9819
	0.857143	363112	17.9135
	1.000000	1594196	78.6471
Total		2027025	

Table 1: URLL tree distances from gold standard.

gorithm, (Felsenstein, 2004). Finally, a summary tree for all languages is obtained by combining the information over all considered alignments. As the described procedure is used separately for each sample obtained from the posterior distribution of the pairwise distances, it results in a measure of statistical uncertainty associated with the topology by counting the relative number of times the obtained tree has a certain topology. Conditional on any topology constructed in this manner, one can obtain posterior inferences for its branch lengths directly from the posterior samples by including the fraction of samples leading to the particular topology.

The software suite implementing this model has been made available to the public.[6]

7 Experiments

In this section we present some results from using the combined pipeline approach, summarized in Figure 1, applied to the Uralic data.

Since we have 10 input datasets that each contribute different pairwise distances, we average these distances over all 10 datasets (for each language pair (a, b), averaging over the 8 datasets where neither a nor b is reference). A topology obtained using this method is shown in Figure 5. Recall, that this tree is *unrooted*,[7] and identifying the node circled in green with the *Finno-Ugric* node in "gold-standard" Uralic trees yields a strong resemblance to the "true" topology. The main deviation in the derived topology is at the node circled in red, corresponding to Permic (ancestor of Komi and Udmurt), which "should" be in the other subtree relative to the Finno-Ugric root. This resulting tree has a URLL distance of 0.28 from the gold-standard tree we introduced in Section 3. To get an intuitive sense of the quality of this result, we observe that the number of unrooted leaf-labeled trees with n nodes is $(2n-3)!!$, (see, e.g., (Ford, 2010)), which is over 2 million for 10 nodes. These trees and their distance from the gold-standard are summarized in Table 1. In the table, $D(T, G)$ denotes the distance of a selected tree to the gold standard. It is easy to check that the expected distance for a randomly selected URLL 10-leaf tree from is over 0.963, with a standard deviation of 0.17. The chance of picking a tree with distance 0.28 or less at random is under 7×10^{-5}.

[6]URL: *http://www.helsinki.fi/bsg/*. Compatibility between the etymological and the population-genetic suites will be maintained also in future releases.

[7]NeighborJoin selects the root via a heuristic, which only tries to minimize the length of the longest root-to-leaf path.

For a deeper investigation of the relations among the languages, we generate a phylogenetic network in SplitsTree4, (Huson and Bryant, 2006), (Figure 6), from the posterior expectations of the pair-wise distances using the Neighbor-Net method, (Bryant and Moulton, 2004). As described in the original article, (Bryant and Moulton, 2004), the sizes of the boxes in the center of the network represent uncertainty about the phylogenetic position of the adjacent leaf nodes. For instance, there is negligible uncertainty about the position of the common ancestor of Finnish and Estonian. In contrast, the greatest uncertainty is related to the position of Permic, which is the only branch in the tree in Figure 5 that deviates from the gold-standard structure. The relevance of the introduced alignment method is highlighted by the fact that our reconstruction of the language relatedness in terms of trees yields results highly congruent with gold-standards .

8 Discussion and current work

Using recent advances from population genetics, we have obtained a promising approach to fully probabilistic inference about language genealogies based on unsupervised etymological alignment. According to our knowledge, this work represent a first attempt to do such inference and it will be of considerable interest to investigate further the properties of this model family in the linguistics context. The essential elements that enable the use of a powerful population-genetics modeling approach are: a. the mapping of sounds to genetic loci which allow the use of a distribution to represent the evidence in the data; b. use of each language in turn as a *reference* language in the pair-wise alignment, instead of an (unobserved) proto-language. Since the model-based distances are averaged over a set of reference languages, the resulting distance estimates are considerably more stable than the individual estimates, as demonstrated in our numerical experiments; c. the novel diffusion approximation-based population-genetics models offer an enormous computational advantage over standard coalescent likelihood-based models. Moreover, the latter models would be considerably more difficult to adapt to the linguistic setting, since they are by definition individual-based, in contrast to the models used here, which enable a direct modeling of languages as a whole by frequencies of the mapped sounds.

Current work includes using context of sounds in aligning the word pairs, and applications to etymological data sets from other language families, and extension for modeling of *internal* nodes in the tree. One direction is using Turkic data (from StarLing), where some of the ancestral languages *are* observed, and examining how accurately the model identifies these languages with internal nodes of the phylogeny. We are also extending the presented model to work with more than 1-1 symbol alignment, using, e.g., 2-2 alignments found in (Kondrak, 2003; Wettig et al., 2012). Finally, using methods for direct N-way alignment—e.g., as suggested in (Steiner et al., 2011)—we may be able to obtain useful estimates of the sounds in the hidden Proto-language, and how they align to sounds in the observed languages. This would in a sense provide the "true" sites, and allow us to circumvent the need for averaging over distances obtained from alignment to reference languages, potentially improving the overall accuracy.

Acknowledgments

This research was supported in part by the FinUgRevita and Uralink Projects of the Academy of Finland, and by the National Centre of Excellence "ALGODAN: Algorithmic Data Analysis" of the Academy of Finland. We thank Teemu Roos for his assistance, and Hannes Wettig, who contributed to building the original alignment models.

References

Raimo Anttila. 1989. *Historical and comparative linguistics*. John Benjamins.

François G. Barbançon, Tandy Warnow, Don Ringe, Steven N. Evans, and Luay Nakhleh. 2009. An experimental study comparing linguistic phylogenetic reconstruction methods. In *Proceedings of the Conference on Languages and Genes*, UC Santa Barbara. Cambridge University Press.

Gareth J. Baxter, Richard A. Blythe, William Croft, and Alan J. McKane. 2009. Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change*, 21(2):257–296.

Richard A. Blythe and Alan J. McKane. 2007. Stochastic models of evolution in genetics, ecology and linguistics. *Journal of Statistical Mechanics: Theory and Experiment*, 2007:P07018.

Richard A. Blythe. 2009. Generic modes of consensus formation in stochastic language dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P02059.

Tugba Bodrumlu, Kevin Knight, and Sujith Ravi. 2009. A new objective function for word alignment. In *Proc. NAACL Workshop on Integer Linear Programming for NLP*, pages 169–174, Copenhagen, Denmark.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL:2007)*, pages 887–896, Prague, Czech Republic.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL09)*.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

David Bryant and Vincent Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, 21(2):255–265.

A.P. Dawid. 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292.

Warren J. Ewens. 2004. *Mathematical population genetics: theoretical introduction*, volume 1. Springer Verlag.

Joseph Felsenstein. 2004. *Inferring phylogenies*. Sunderland, Massachusetts: Sinauer Associates.

Daniel J. Ford. 2010. Encodings of cladograms and labeled trees. *Electronic Journal of Combinatorics*, 17:1556–1558.

Peter Grünwald. 2007. *The Minimum Description Length Principle*. MIT Press.

Heikki Haario, Eero Saksman, and Johanna Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

David Hall and Dan Klein. 2011. Large-scale cognate recovery. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Daniel H. Huson and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–267.

Brett Kessler. 2001. *The Significance of Word Lists: Statistical Tests for Investigating Historical Connections Between Languages*. The University of Chicago Press, Stanford, CA.

Grzegorz Kondrak. 2002. Determining recurrent sound correspondences by inducing translation models. In *Proceedings of COLING 2002: 19^{th} International Conference on Computational Linguistics*, pages 488–494, Taipei.

Grzegorz Kondrak. 2003. Identifying complex sound correspondences in bilingual wordlists. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing (CICLing-2003)*, pages 432–443, Mexico City. Springer-Verlag Lecture Notes in Computer Science, No. 2588.

Grzegorz Kondrak. 2004. Combining evidence in cognate identification. In *Proceedings of the Seventeenth Canadian Conference on Artificial Intelligence (Canadian AI 2004)*, pages 44–59, London, Ontario. Lecture Notes in Computer Science 3060, Springer-Verlag.

I. Dan Melamed. 1997. Automatic discovery of noncompositional compounds in parallel data. In *The Second Conference on Empirical Methods in Natural Language Processing*, pages 97–108, Hissar, Bulgaria.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Luay Nakhleh, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language (Journal of the Linguistic Society of America)*, 81(2):382–420.

Javad Nouri and Roman Yangarber. 2016. Modeling language evolution with codes that utilize context and phonetic features. In *Proceedings of CoNLL: Conference on Computational Natural Language Learning, at ACL-2016*, Berlin, Germany, August. Association for Computational Linguistics.

Károly Rédei. 1991. *Uralisches etymologisches Wörterbuch*. Harrassowitz, Wiesbaden.

Don Ringe, Tandy Warnow, and A. Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

D.F. Robinson and L.R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1–2):131–147.

Jukka Sirén, Pekka Marttinen, and Jukka Corander. 2011. Reconstructing population histories from single nucleotide polymorphism data. *Molecular Biology and Evolution*, 28(1):673–683.

Jukka Sirén, William P. Hanage, and Jukka Corander. 2013. Inference on population histories by approximating infinite alleles diffusion. *Journal of Molecular Biology and Evolution*, 30(2):457–468.

Sergei A. Starostin. 2005. Tower of Babel: StarLing etymological databases. http://newstar.rinet.ru/.

Lydia Steiner, Peter F Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.

Fabian Vázquez, Xavier Castelló, and Maxi San Miguel. 2010. Agent based models of language competition: Macroscopic descriptions and order-disorder transitions. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P04007.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of 16^{th} Conference on Computational Linguistics (COLING 96)*, pages 169–174, Copenhagen, Denmark.

Hannes Wettig, Suvi Hiltunen, and Roman Yangarber. 2011. MDL-based Models for Alignment of Etymological Data. In *Proceedings of RANLP: the 8^{th} Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgaria.

Hannes Wettig, Kirill Reshetnikov, and Roman Yangarber. 2012. Using context and phonetic features in models of etymological sound change. In *Proc. EACL Workshop on Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 37–44, Avignon, France.

Dmitry Zelenko. 2009. Combining MDL transliteration training with discriminative modeling. In *Proceedings of the ACL-IJCNLP*, Singapore.