HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
Faculty of Science
Department of Computer Science
ALGODAN Center of Excellence

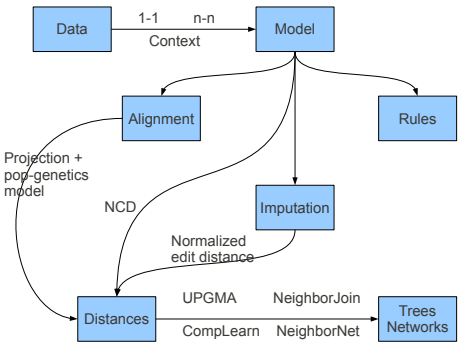
Suvi Hiltunen
Javad Nouri
Hannes Wettig
Roman Yangarber

with Jukka Corander
Jukka P. Sirén
Department of Mathematics and Statistics

Analysis of Etymological Data via MDL

We develop MDL-based models for studying etymological data. The data consists of *cognate sets*: sets of genetically related words—words deriving from a common (unobserved) ancestor in the proto-language—in different (observed) languages within a language family. One goal is to find the best possible *alignment* of all the words in the data. The alignment must respect the *Principle of Regular Sound Correspondence*: sound changes that occur as a given language evolves are not random, but apply deterministically throughout the language, typically conditioned on the features and the context of the sound. Thus, a complementary goal is to discover the rules of sound change that best describe the data.

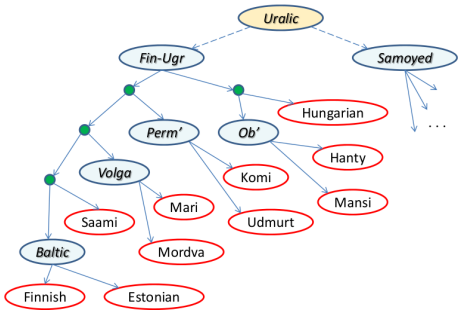
OUTLINE



DATA

We have several databases of *cognate sets* from different language families, including the Uralic family. The databases may conflict with regard to inclusion of specific words in a cognate set.

ID	EST	FIN	KHN	KOM	MAN	MAR	MRD	SAA	UDM	UGR
91	-	-	-	ćepeĺ ?	-	cəwešte ?	-	-	čepĺt ?	csip ?
92	-	-	-	-	-	čeve ?	-	-	-	csëpp ?
93	-	-	-	tovta ?	śúś ?	-	-	-	-	-
94	-	-	-	śuź ?	-	-	-	cisku ?	-	sas ?
95	-	-	-	ǯoǯ	-	-	-	-	ǯǰǯol	-
96	-	-	čäčə	ćuž	šošəɣ	šača	šačo	-	-	-
97	ammak	hama	čäma	-	šomməɣ	-	-	-	-	-
98	-	-	-	-	-	-	-	cuoʒ'ʒâ	-	-
99	-	-	čuš	-	šuš	-	šašto	-	-	-
100	-	-	čoŋɣ	-	šaŋk	čaŋɣe	čavo	-	-	-
101	-	-	-	-	-	šapka ?	-	-	-	sápad ?
102	hape ?	hapan ?	-	-	-	šapə̄	čapamo ?	-	-	savanyú
103	-	-	čäkən	ǯagal	šə̄kət	-	-	-	ǯokal	čäk
104	händ	häntä	čeŋč	-	šjs	-	-	-	-	-



THE OBJECTIVE

We begin with pairwise alignment—one language pair at a time.
According to the Minimum Description Length (MDL) principle, we can compress the data effectively if we can discover *regularity* in the data. This regularity is the laws of sound change that we seek.
Thus, the objective function that we optimize is the MDL codelength; using Bayesian marginal likelihood, or *prequential* coding:

$$L(D) = -\sum_{e\in E}\log\Gamma\big(c(e)+\alpha(e)\big) + \sum_{e\in E}\log\Gamma\big(\alpha(e)\big)$$

$$+\log\Gamma\Big[\sum_{e\in E}\big(c(e)+\alpha(e)\big)\Big] - \log\Gamma\Big[\sum_{e\in E}\alpha(e)\Big]$$

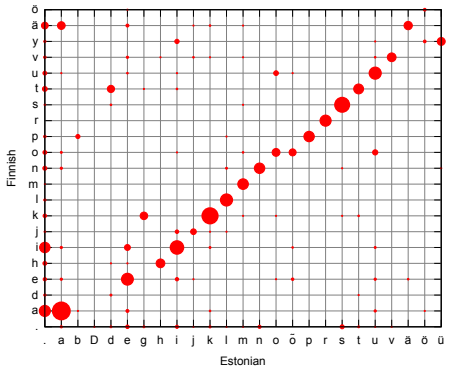
Using Normalised Maximum Likelihood (NML) gives somewhat better compression overall.

BASELINE (1-1) ALIGNMENT MODEL

For a given word-pair, many alignments are possible: Finnish and Hanty words meaning *year*:

v	*u*	*o*	*s*	*i*		*v*	*u*	*o*	*s*	*i*	
|	|	|	|	|		|	|	|	|	|	etc...
a	*l*	.	.	.		.	*a*	.	*l*	.	

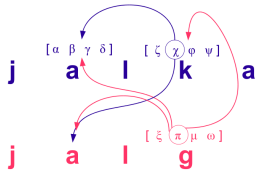
(The symbol "." indicates deletion or insertion.)
Search algorithm: begin with a random alignment, and iteratively realign one word pair at a time via Dynamic Programming, using the currently best alignment of the remainder of the data.



The algorithm converges to a (locally optimal) alignment of the complete data. The area of the circle is proportional to the probability mass of each 1-1 symbol alignment.

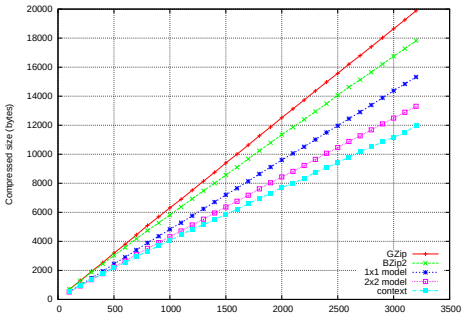
CONTEXT MODELS

We code each sound σ as a vector of phonetic features, **and** coding is conditioned on (features of) sounds in the context of σ—the model can query the history that has been coded so far.



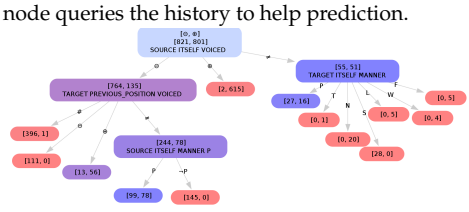
COMPRESSION RATES

The test of the model "goodness" is compression power: the cost of the complete (aligned) data:



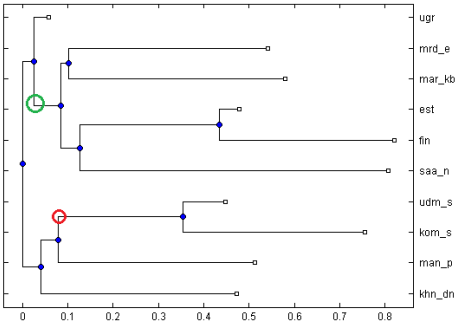
RULES AS DECISION TREES

The model learns one tree for coding each feature of a sound, minimizing the tree cost. Each node queries the history to help prediction.



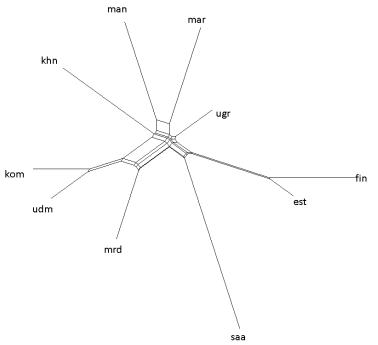
RECONSTRUCTING PHYLOGENIES

We obtain pairwise language distances in several ways from the alignment models, and induce trees using, e.g., UPGMA, NeighborJoin:



PHYLOGENETIC NETWORKS

NeighborNet (SplitsTree) helps identify the uncertainty in the phylogenetic reconstructions:



Applying to other language families: Turkic