Is automatic cognate detection good enough for phylogenetic inference?

Taraka Rama[1,2], Johann-Mattis List[3], Johannes Wahle[1] & Gerhard Jäger[1]

[1]Tübingen University, [2]Oslo University & [3]MPI Jena

Jena, CESC 2017

September 13, 2017

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

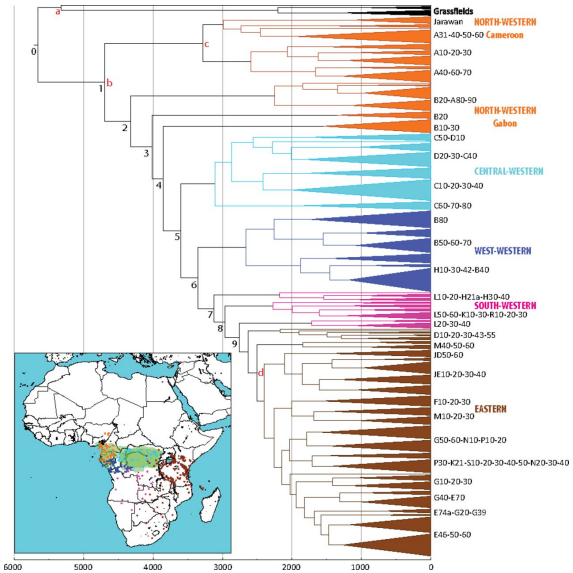
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

DFG

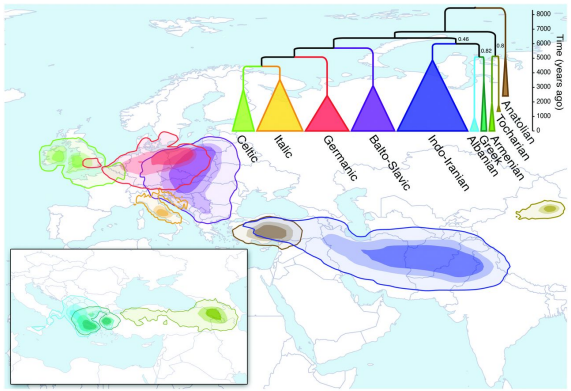
erc
European Research Council
Established by the European Commission

Computational historical linguistics

- massive progress within past 15 years
- automated language classification
- inferring time depth and homeland of language families
- automatic reconstruction of proto-languages
- discovery of statistical patterns in language change
- ...



(Grollemund et al, 2015)



(Bouckaert et al, 2012)

Computational historical linguistics

Manual cognate judgements

- most work depends on manually coded cognate judgments on Swadesh lists
  - labor intensive
  - subjective
  - not fully replicable
  - induces bias in favor of well-studied language families

Computational historical linguistics

Manual cognate judgements

- most work depends on manually coded cognate judgments on Swadesh lists
  - labor intensive
  - subjective
  - not fully replicable
  - induces bias in favor of well-studied language families

language	*iso*	*gloss*	*gloss_id*	*transcr.*	*cogn._class*
ELFDALIAN	qov	woman	962	ˈkɛ̀lɪŋg	woman:Ag
DUTCH	nld	woman	962	vrɑu	woman:B
GERMAN	deu	woman	962	fraŭ	woman:B
DANISH	dan	woman	962	ˈg̊ʰvenə	woman:D
DANISH_FJOLDE		woman	962	kvinʲ	woman:D
GUTNISH_LAU		woman	962	ˈkvɪnːˌfolk	woman:D
LATIN	lat	woman	962	ˈmulier	woman:E
LATIN	lat	woman	962	ˈfeːmina	woman:G
ENGLISH	eng	woman	962	wʊmən	woman:H
GERMAN	deu	woman	962	vaĭp	woman:H
DANISH	dan	woman	962	ˈd̥ɛːmə	woman:K

- The goal of automated cognate detection is to do this automatically

Datasets

Dataset	Words	Conc.	Lang.	Families	Cog.	Div.
ABVD (Greenhill et al., 2008)	12414	210	100	Austronesian	3558	0.27
Sino-Tibetean (Peiros, 2004)	8694	110	81	Sino-Tibetean	1128	0.13
IELex (Dunn, 2012)	11479	208	52	Indo-European	2459	0.20

Expert Trees

- Expert trees were obtained from Glottolog (Hammarström et al., 2015)

Glottolog

Glottolog provides a comprehensive catalogue of the world's languages, language families and dialects. [...] The languoids are organized via a genealogical classification (the Glottolog tree) that is based on available historical-comparative research[...]. (http://glottolog.org/)

Automated Cognate Detection

LexStat

- algorithm first propose in List (2012) and then further enhanced in List (2014), List et al. (2016) and List et al. (2017)
- the algorithm is generally based on the alignment-based workflow for cognate detections
- implemented as part of LingPy (lingpy.org,List and Forkel (2016))

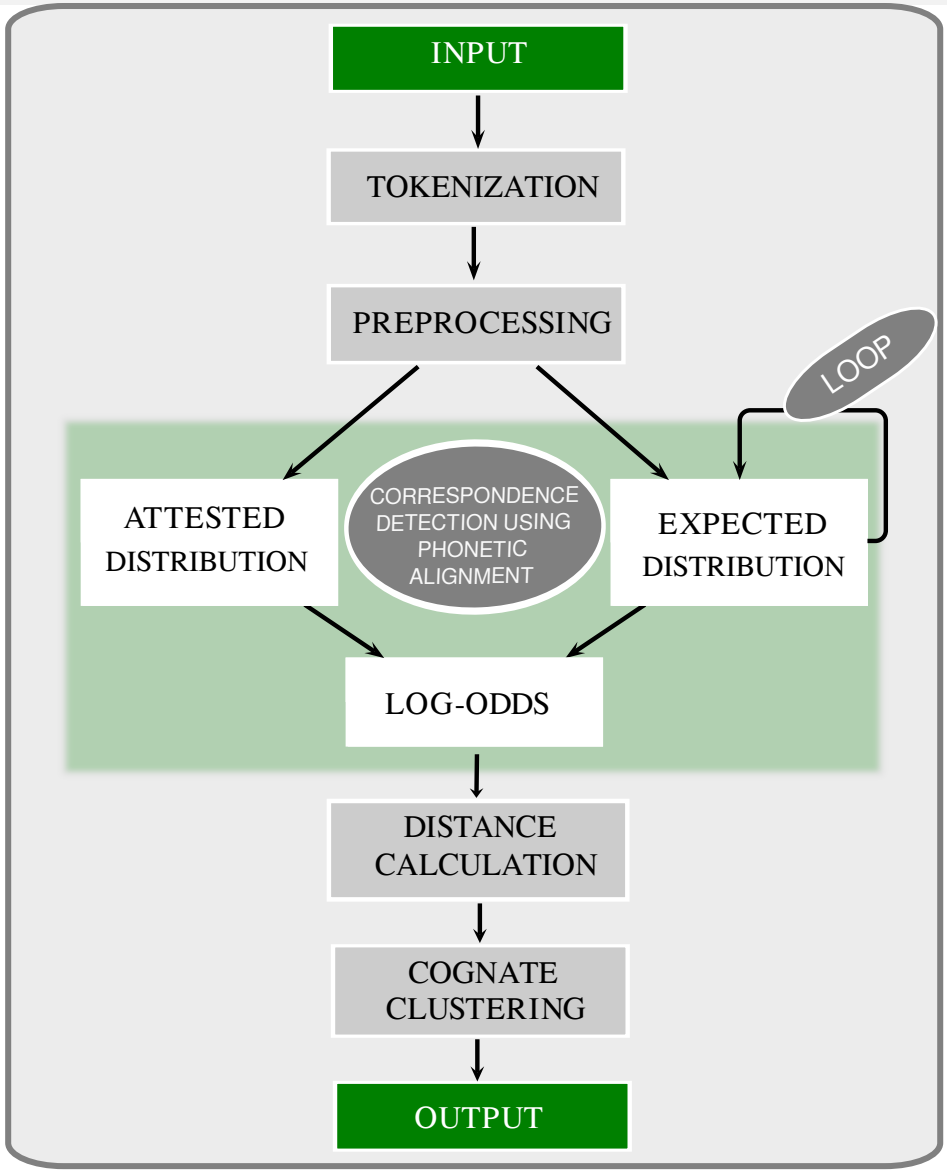
SVM

- Classification based approach (Jäger and Sofroniev, 2016; Jäger et al., 2017)
- A pair of words is classified as cognate or not based on a feature vector describing this pair
- Implementation: http://www.evolaemp.uni-tuebingen.de/svmcc/

LexStat

- scoring functions for alignments are computed *individually* for each language pair, modeling regular sound correspondences in classical linguistics
- scores for both global and local alignment analyses are combined and agglomerated
- alignment algorithm is sensitive for morpheme boundaries if they are annotated (*secondary alignment*, List (2014))
- sequences are represented as *multi-tiered structures* which allows to handle *prosodic context*
- agglomerative clustering procedure has been replaced by a community detection algorithm (Infomap, Rosvall and Bergstrom (2008))

LexStat



LexStat Algorithm (List 2014)

LexStat: Cognate Set Partitioning

A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

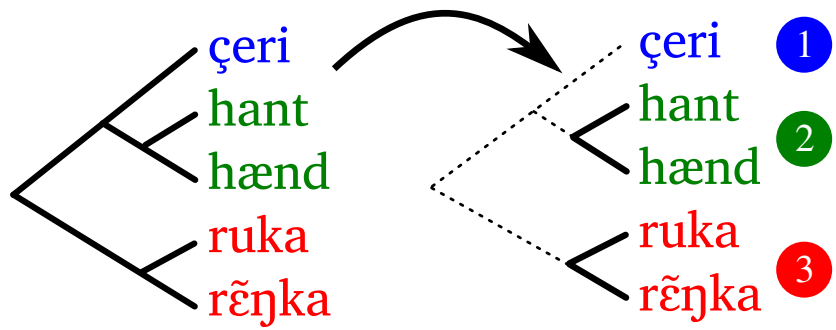
List et al. (2017)

LexStat: Cognate Set Partitioning

A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

B



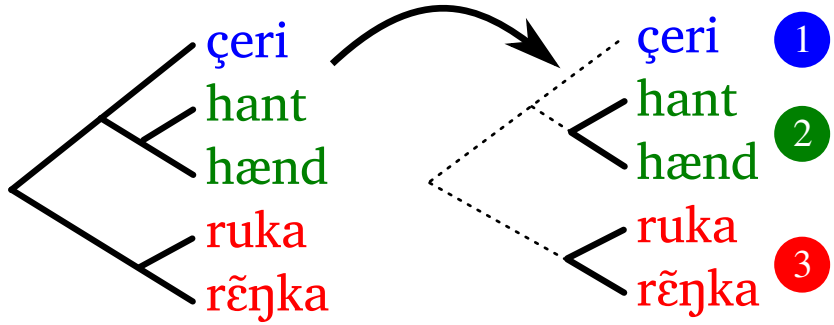
List et al. (2017)

LexStat: Cognate Set Partitioning

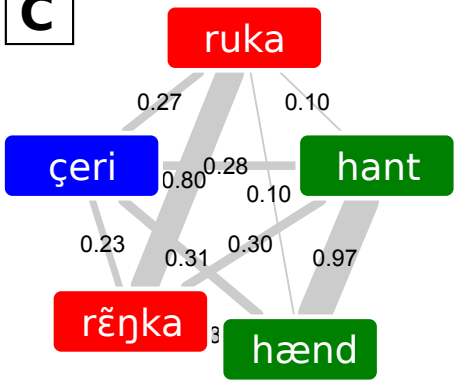
A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

B



C



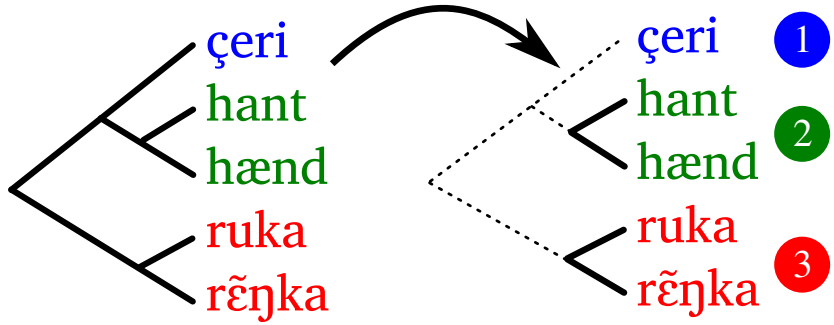
List et al. (2017)

LexStat: Cognate Set Partitioning

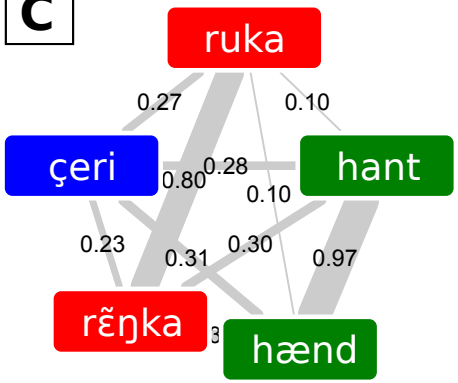
A

		GREEK	GERMAN	ENGLISH	RUSSIAN	POLISH
GREEK	çeri	0.00	0.72	0.69	0.73	0.77
GERMAN	hant	0.72	0.00	0.03	0.91	0.70
ENGLISH	hænd	0.69	0.03	0.00	0.91	0.68
RUSSIAN	ruka	0.72	0.91	0.91	0.00	0.20
POLISH	rɛ̃ŋka	0.77	0.70	0.68	0.20	0.00

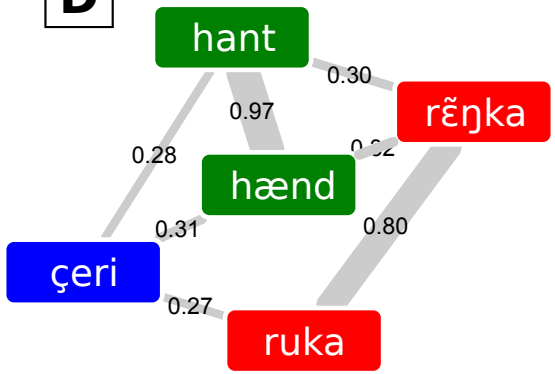
B



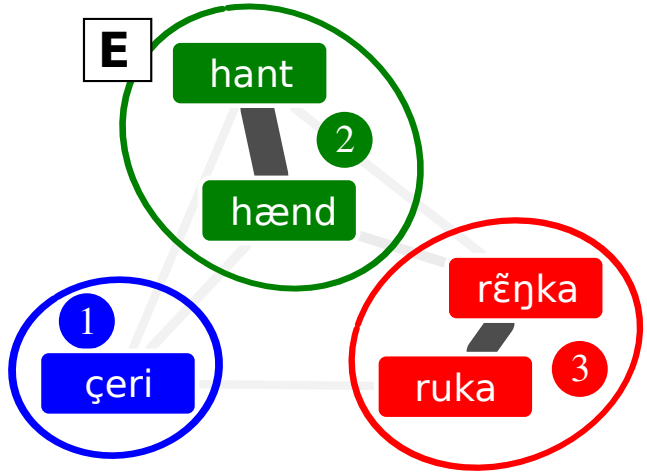
C



D

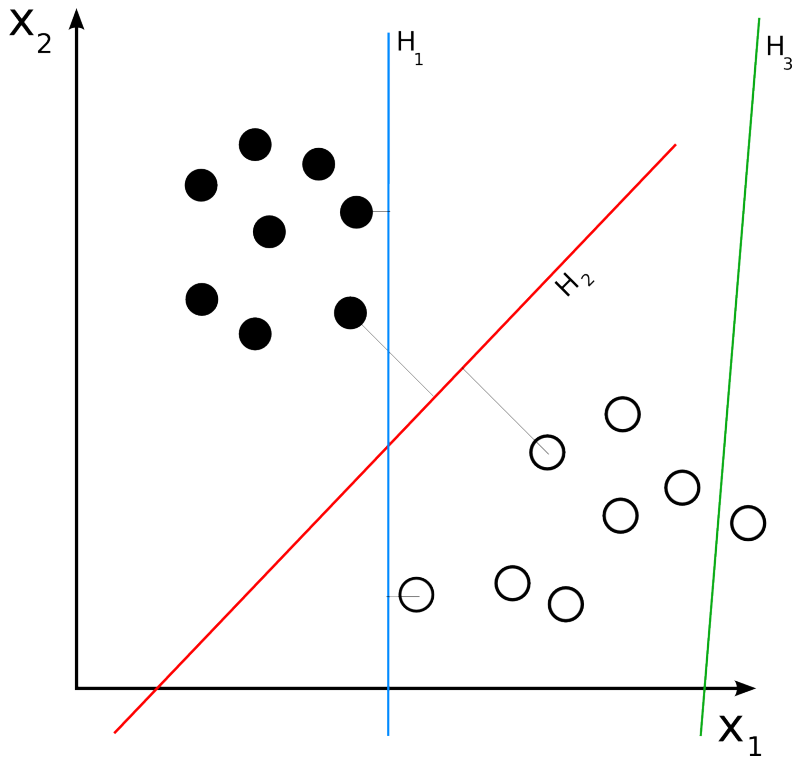


E



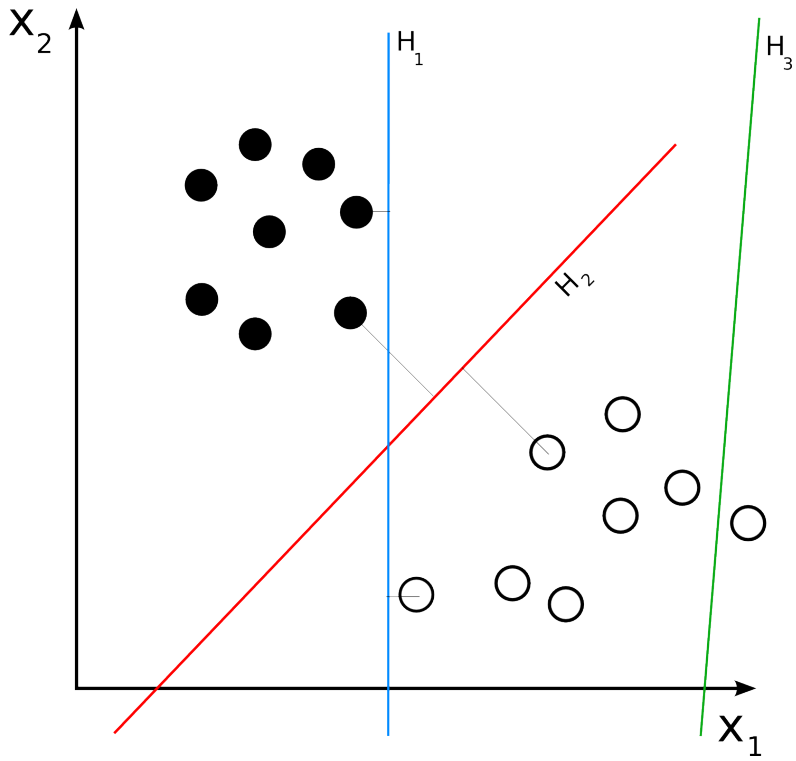
List et al. (2017)

SVM



via Wikimedia Commons

SVM



via Wikimedia Commons

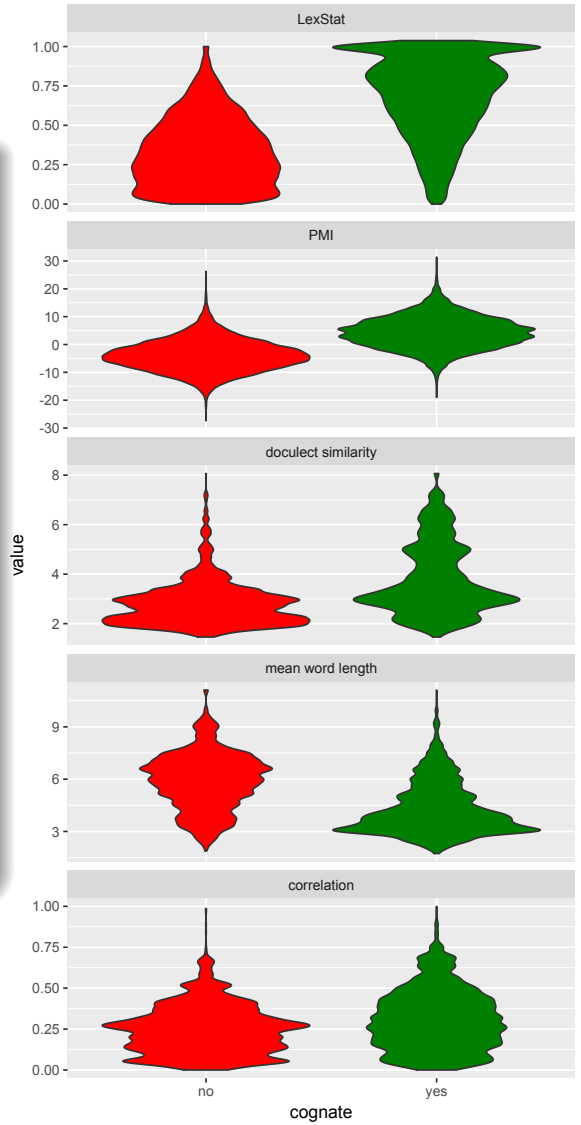
Model selection

- each synonymous word pair is a data point
- cognate (yes/no) as dependent variable
- Feature selection
  - seven features from (Jäger and Sofroniev, 2016) + LexStat similarity as candidate features
  - feature selection via cross-validation on training data

SVM

Model selection

- five informative features
  - LexStat similarity
  - PMI similarity
  - doculect similarity
  - measures of concept stability
    - mean word length
    - correlation between string similarity and doculect similarity
- linear kernel



Performance of the ACD Methods

- All scores are B-Cubed scores (Bagga and Baldwin, 1998)

dataset	*Precision*		*Recall*		*F-score*	
	LexStat	*SVM*	*LexStat*	*SVM*	*LexStat*	*SVM*
Indo-European	0.896	0.877	0.750	0.770	0.817	0.820
Austronesian	0.791	0.781	0.801	0.855	0.796	0.817
Sino-Tibetean	0.928	0.848	0.301	0.409	0.455	0.552

Observations

- current methods for automatic cognate detection are fairly accurate
- they are objective and replicable
- they make it much easier to obtain large datasets for cross-linguistic historical language comparison

Question: Are they good enough for phylogenetic inference?

Mr. Bayes

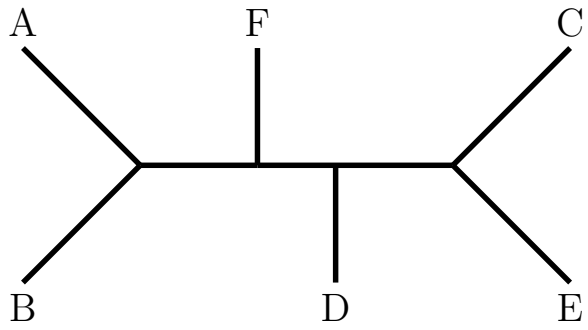
- An (ideal) phylogenetic model consists of
  1. Phylogenetic tree + branch length
  2. (Explicit) model of character evolution along the branches of the tree
- Mr. Bayes (Ronquist et al., 2012) provides a framework for such phylogenetic analysis
- Uses Markov Chain Monte Carlo methods to estimate a posterior of the parameters of the model

Mr. Bayes

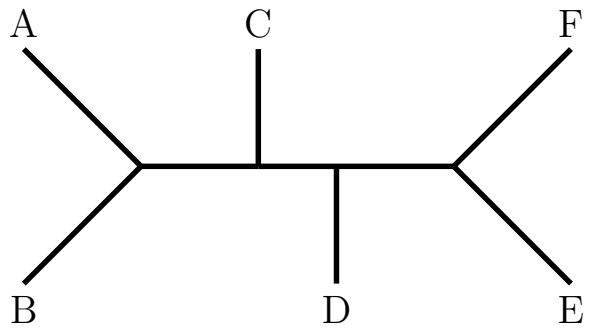
- An (ideal) phylogenetic model consists of
  1. **Phylogenetic tree** + branch length
  2. (Explicit) model of character evolution along the branches of the tree
- Mr. Bayes (Ronquist et al., 2012) provides a framework for such phylogenetic analysis
- Uses Markov Chain Monte Carlo methods to estimate a posterior of the parameters of the model

Generalized Quartet Distance

- number of quartets not having the same topology in both trees divided by the number of quartets in the expert tree (Pompei et al., 2011)

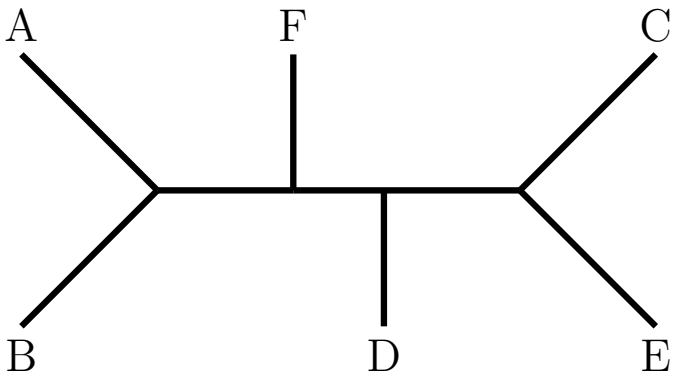


Expert Tree

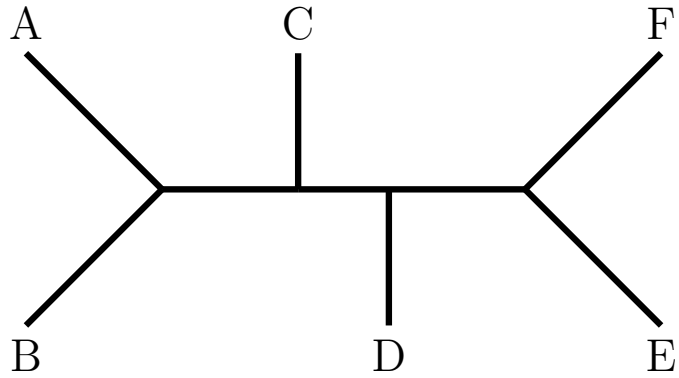


Inferred Tree

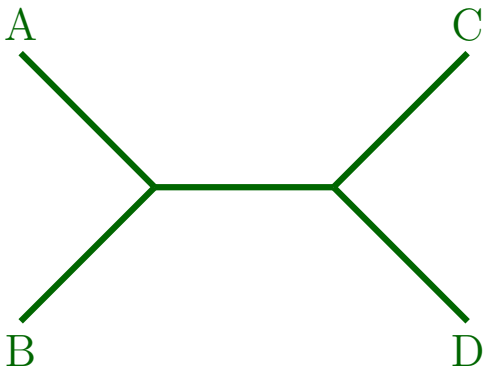
Generalized Quartet Distance



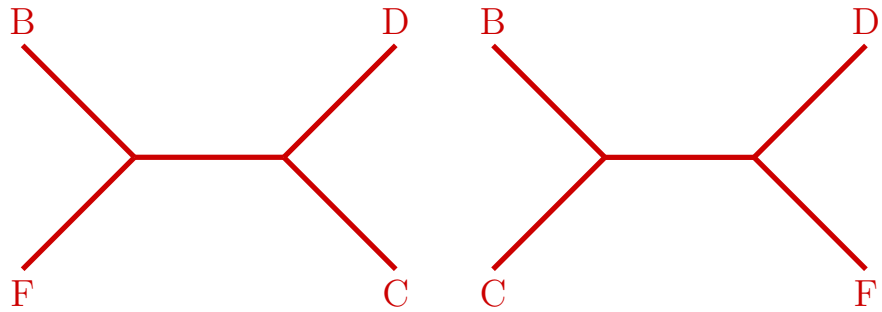
Expert Tree



Inferred Tree



Same configuration

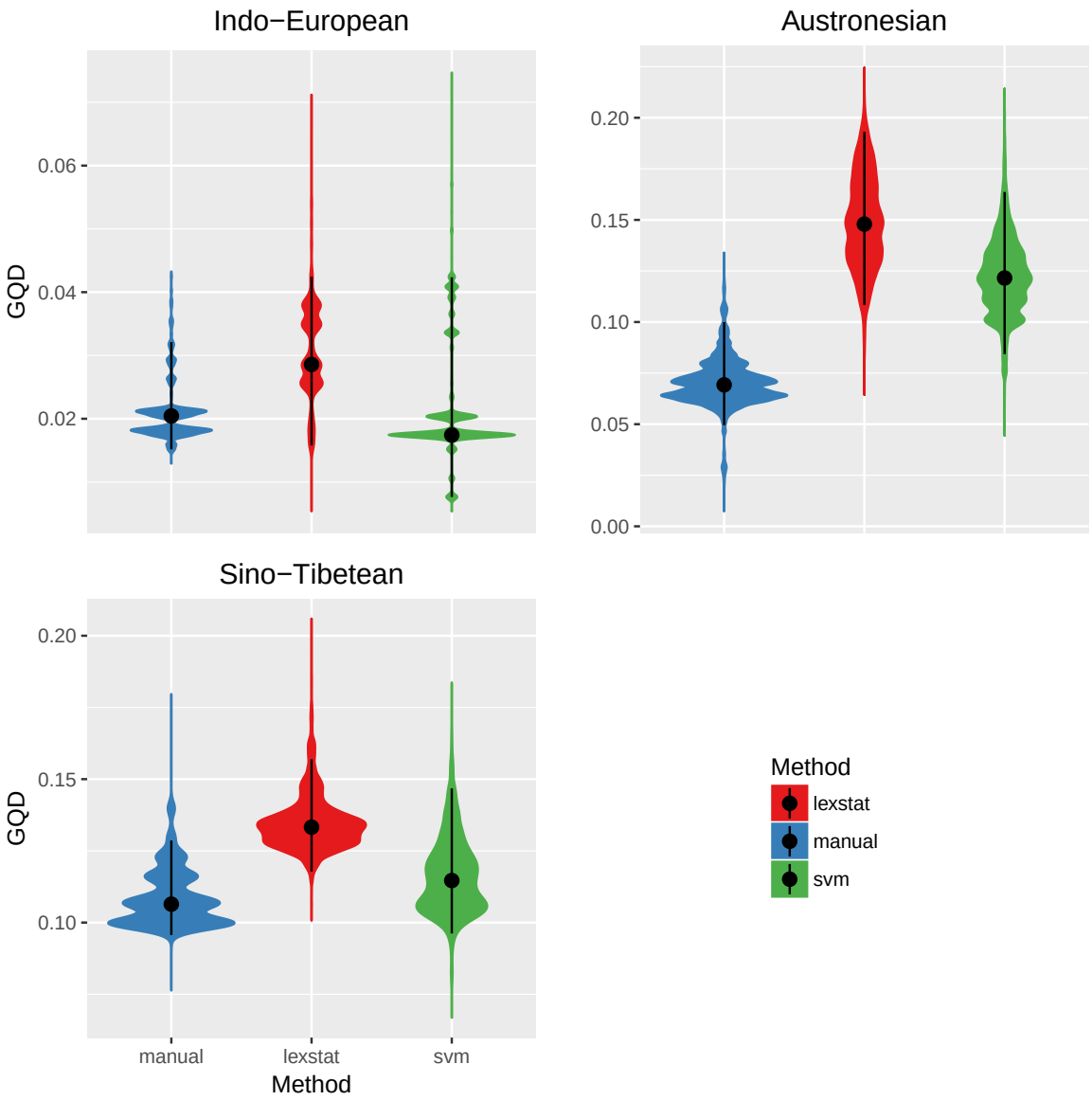


Different configuration

	manual	LexStat	SVM
Austronesian	**0.069 (0.050-0.100)**	0.133 (0.108-0.193)	0.121 (0.084-0.164)
Sino-Tibetan	**0.106 (0.096-0.129)**	0.133 (0.118-0.157)	0.115 (0.096-0.147)
Indo-European	**0.020 (0.015-0.032)**	0.029 (0.016-0.042)	0.017 (0.008-0.042)

Median and HPD Interval (95%)

Results


Indo−European
Austronesian
Sino−Tibetean
GQD
0.06
0.04
0.02
0.20
0.15
0.10
0.05
0.00
manual
lexstat
svm
Method
Method
lexstat
manual
svm

Conclusion

- The performance is close to manual judgments (especially for Indo-European)
- The supervised ACD method (SVM) performs better than the unsupervised method (LexStat)
- Trees from manually annotated cognates are not perfect either
  1. Rethink tree inference algorithms for linguistics?
  2. Rethink cognate annotation?

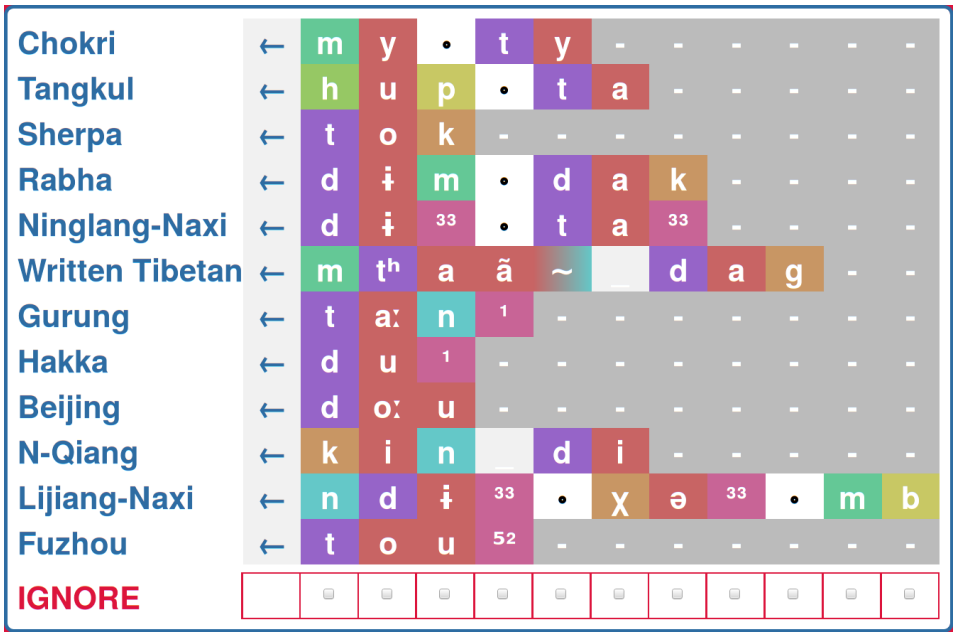
Outlook

How can we improve?

- Enhance the quality of our datasets (Sino-Tibetan is highly problematic, IELex will be superseded by COBL, ABVD can be improved by looking into sub-branches).
- Test expert cognate trees vs. autocogs on better data (COBL?).
- Enhance annotation of lexical differences (partial cognates are needed, List 2016, morphological differences in data should be annotated).
- Enhance phonological annotation (standardized cross-linguistic transcription systems).
- Enhance lexical coverage in the data (increase size of concept lists, increase internal coverage per language).

Outlook

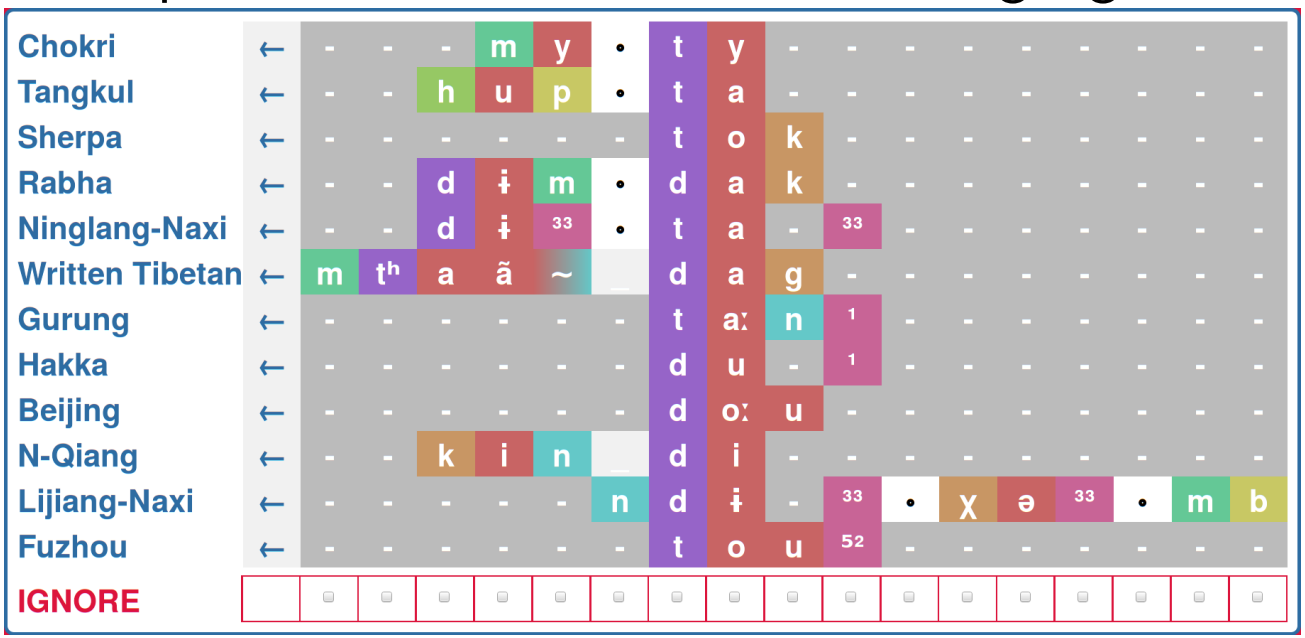
Concept “all” in some Sino-Tibetan languages:



Annotating alignments with the EDICTOR (List, 2017).

Outlook

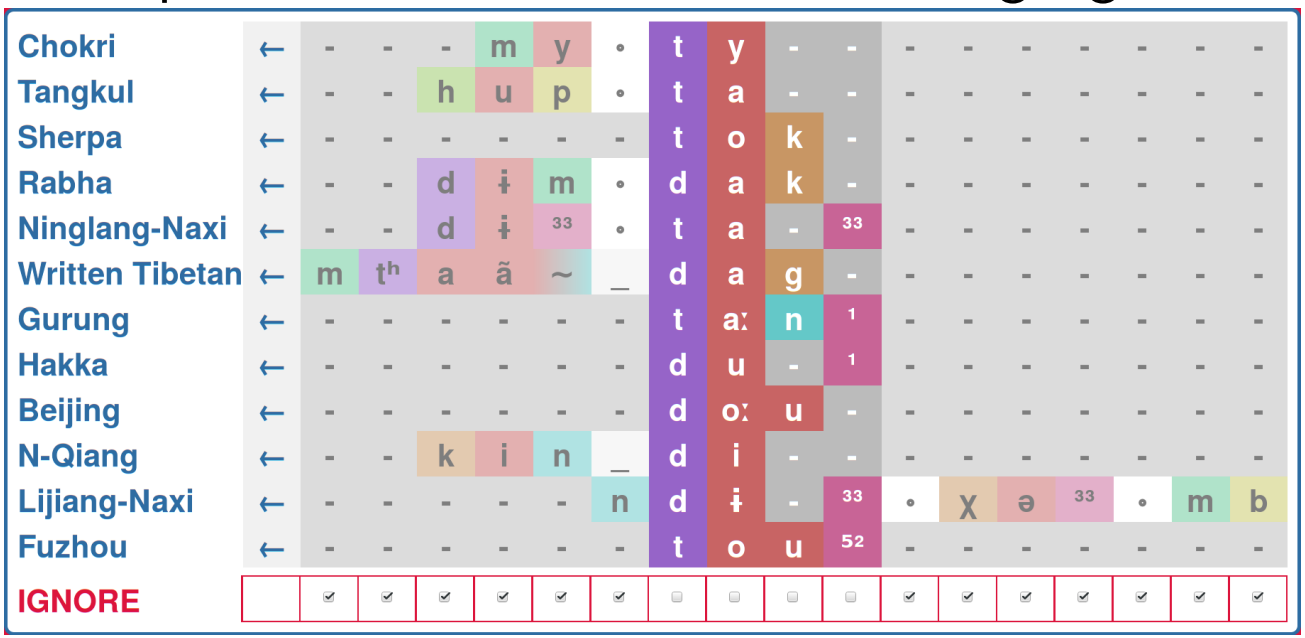
Concept “all” in some Sino-Tibetan languages:



Annotating alignments with the EDICTOR (List, 2017).

Outlook

Concept “all” in some Sino-Tibetan languages:



Annotating alignments with the EDICTOR (List, 2017).

Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the ACL*, pages 79–85, 1998.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960, Aug 2012.

Michael Dunn. Indo-European lexical cognacy database (IELex). URL: http://ielex.mpi.nl/, 2012.

Simon J. Greenhill, Robert Blust, and Russell D. Gray. The Austronesian Basic Vocabulary Database. *Evolutionary Bioinformatics*, 4:271–283, 2008.

Rebecca Grollemund, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43): 13296–13301, 2015.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2015. URL: http://glottolog.org.

Gerhard Jäger and Pavel Sofroniev. Automatic cognate classification with a Support Vector Machine. In Stefanie Dipper, Friedrich Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13th Conference on Natural Language Processing*, volume 16 of *Bochumer Linguistische Arbeitsberichte*, pages 128–134. Ruhr Universität Bochum, 2016.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E17-1113.

Johann-Mattis List. Lexstat. automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 117–125, Stroudsburg, 2012.

Johann-Mattis List. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf, 2014.

Johann-Mattis List. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136, 2016. doi: http://dx.doi.org/10.1093/jole/lzw006. URL http://jole.oxfordjournals.org/content/1/2/119.

Johann-Mattis List. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia, 2017. Association for Computational Linguistics. URL http://edictor.digling.org.

Johann-Mattis List and Robert Forkel. *LingPy 2.5*. Max Planck Institute for the Science of Human History, Jena, 2016. doi: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. URL http://lingpy.org. URL: http://lingpy.org.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the ACL 2016 Short Papers*, 2016.

Johann-Mattis List, Simon Greenhill, and Russell Gray. The potential of automatic cognate detection for historical linguistics. *PLOS ONE*, 2017.

Ilia Peiros. *[Dataset on Sino-Tibetan languages encoded in STARLING in the file] sintib.exe*. Russian State University for the Humanities, Moscow, 2004. URL http://starling.rinet.ru/download/SINTIB.exe.

Simone Pompei, Vittorio Loreto, and Francesca Tria. On the accuracy of language trees. *PLOS ONE*, 6(6):1–11, 06 2011. doi: 10.1371/journal.pone.0020109. URL https://doi.org/10.1371/journal.pone.0020109.

F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, 61(3):539–42, 2012. ISSN 1063-5157 (Print). doi: 10.1093/sysbio/sys029.

Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4): 1118–1123, 2008.