An algorithm for inferring phylogenetic trees of languages using morphosyntactic feature data that are congruent to the gold standard reference phylogenies

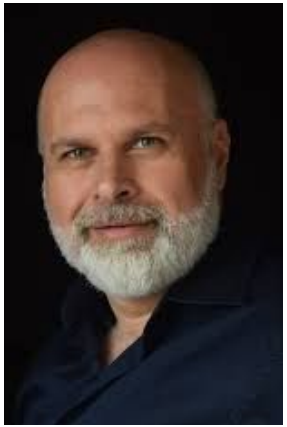
ICZEGAR, 15 October 2022 Lesvos

Pavlos Pavlidis

FORTH
INSTITUTE OF COMPUTER SCIENCE

The Team

Elena Anagnostopoulou



Manolis Ladoukakis

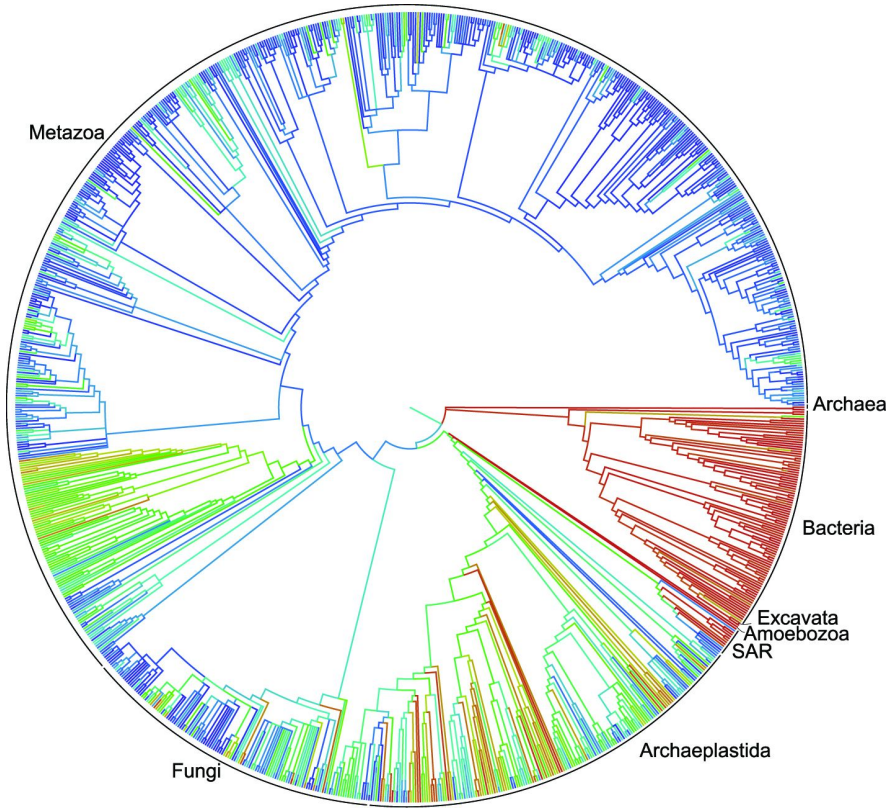
Margarita Makri

Dimitris Michelioudakis

Christos Zioutis

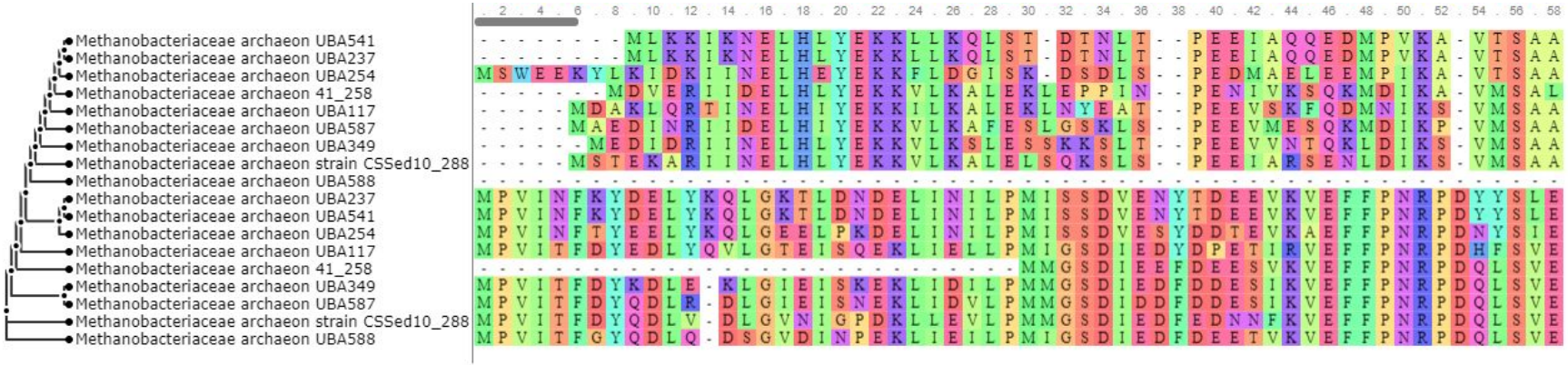
ΕΛΙΔΕΚ
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

Studying evolution via phylogenies



A phylogeny describes evolution as divergence from common ancestors

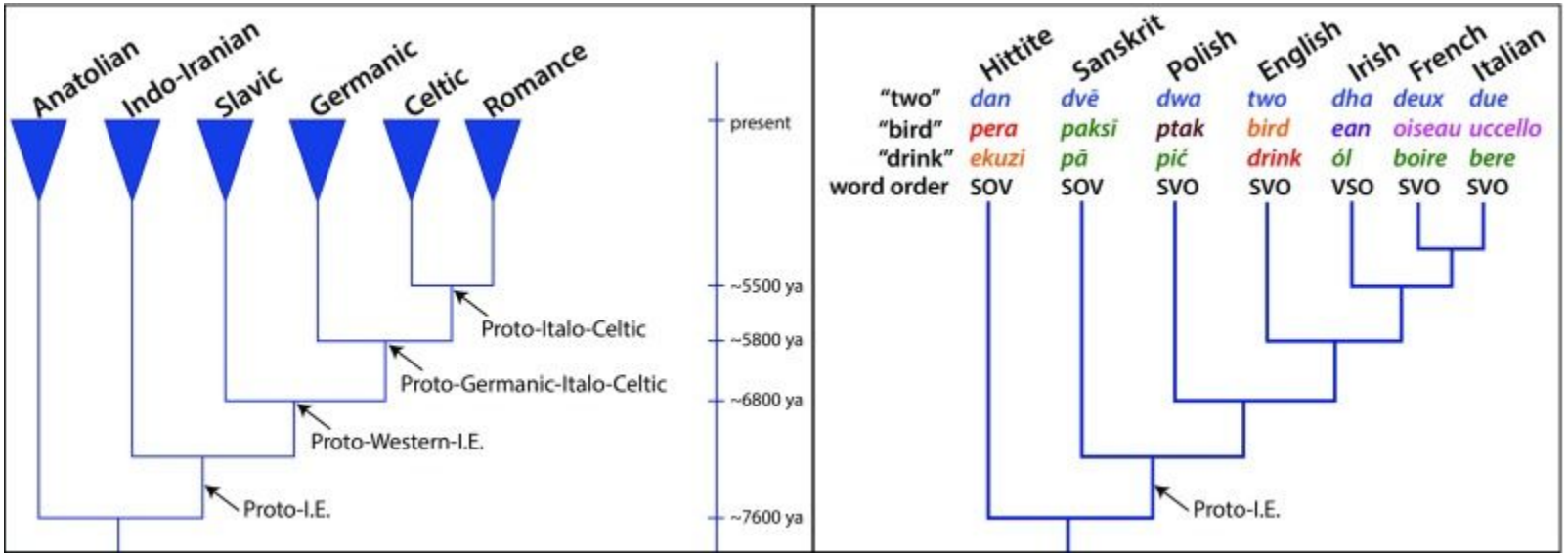
Constructing phylogenies



- Multiple Sequence Alignments
- Orthologue sites are aligned and a phylogenetic tree that fits the data is constructed

Evolution of languages

Languages evolve from common ancestors



Pagel 2017

Evolution of languages using cognates

Cognates are “orthologue” words

They have been inherited in direct descent from an etymological ancestor in a common parent language.

starve	father	night
sterben (German)	père (French)	νύχτα (Greek)
sterven (Dutch)	հայր (hayr; Armenian)	nishi (Bengali)
		natë (Albanian)
		nox/nocte (Latin)

Evolution of languages using cognates

There are thousands of cognates that can be used to build phylogenetic trees

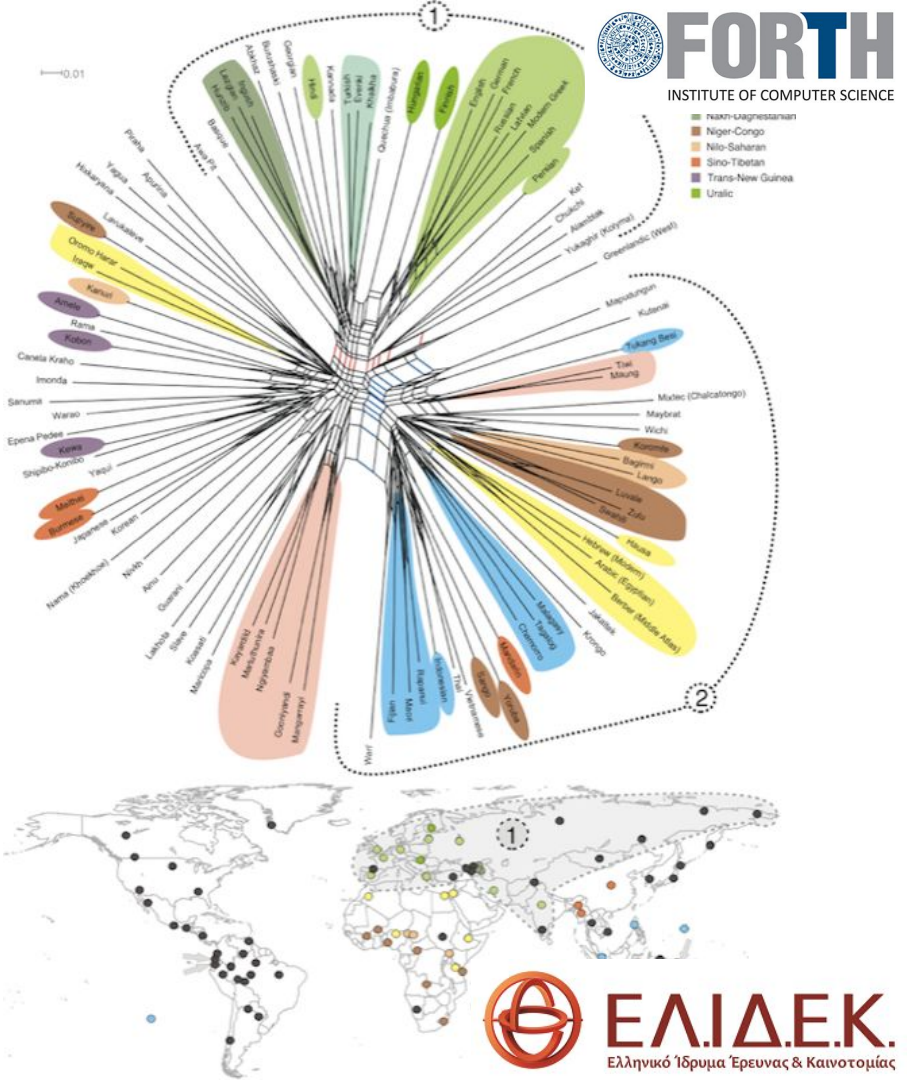
	Meaning 1		Meaning 2	
	lexeme	class	lexeme	class
Language A	mhɨm	a	cɨŋ	x
Language B	mhɨm	a	kɨt	y
Language C	lɔ:t	b	kət, lpəc	y, z
Language D	?	?	lpət	z

	Cognate set				
	1a	1b	2x	2y	2z
Language A	1	0	1	0	0
Language B	1	0	0	1	0
Language C	0	1	0	1	1
Language D	?	?	0	0	1

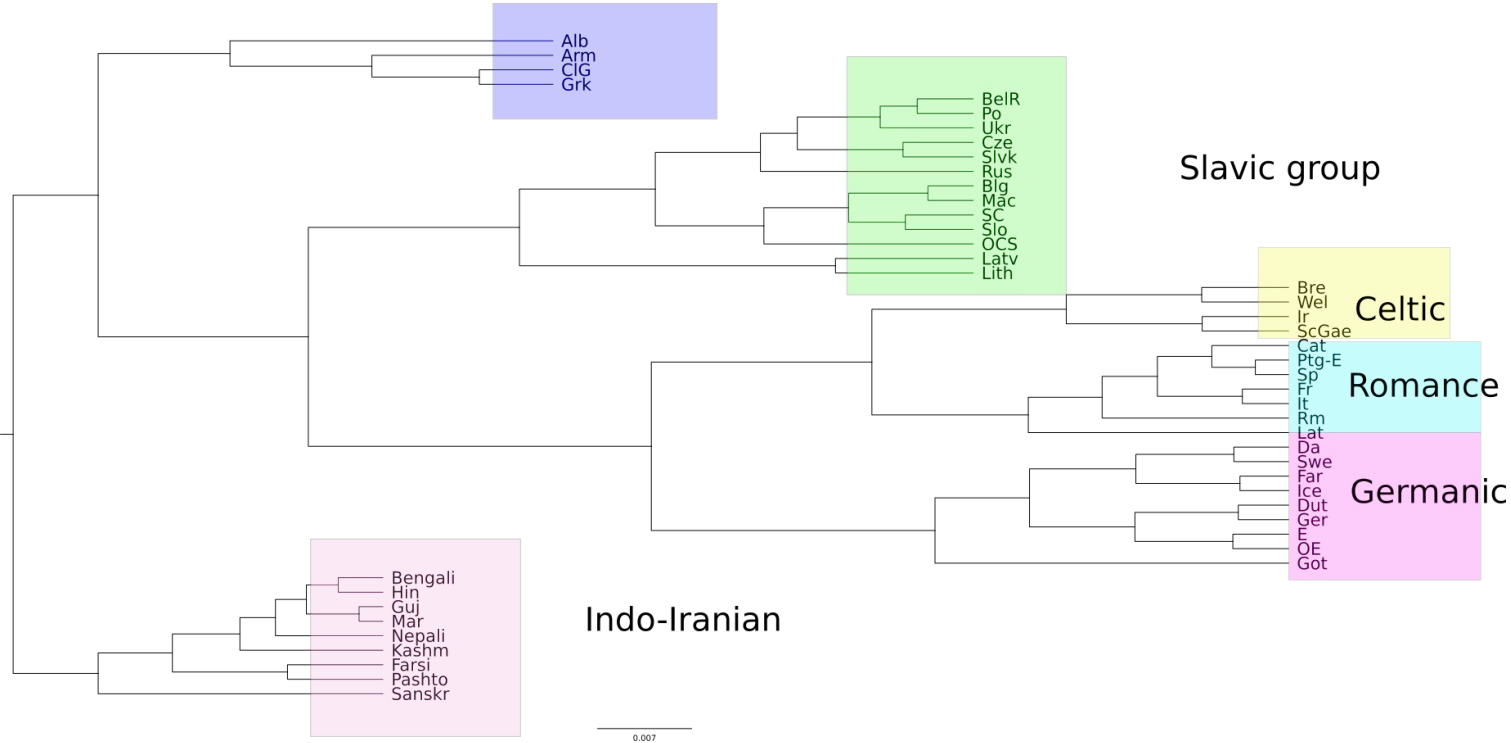
Distance based methods: NJ method; UPGMA method
Model based: The MK evolutionary model (Maximum Likelihood; Bayesian)

Cognates result in accurate phylogenies

Phylogenies using cognates capture historical relations between languages

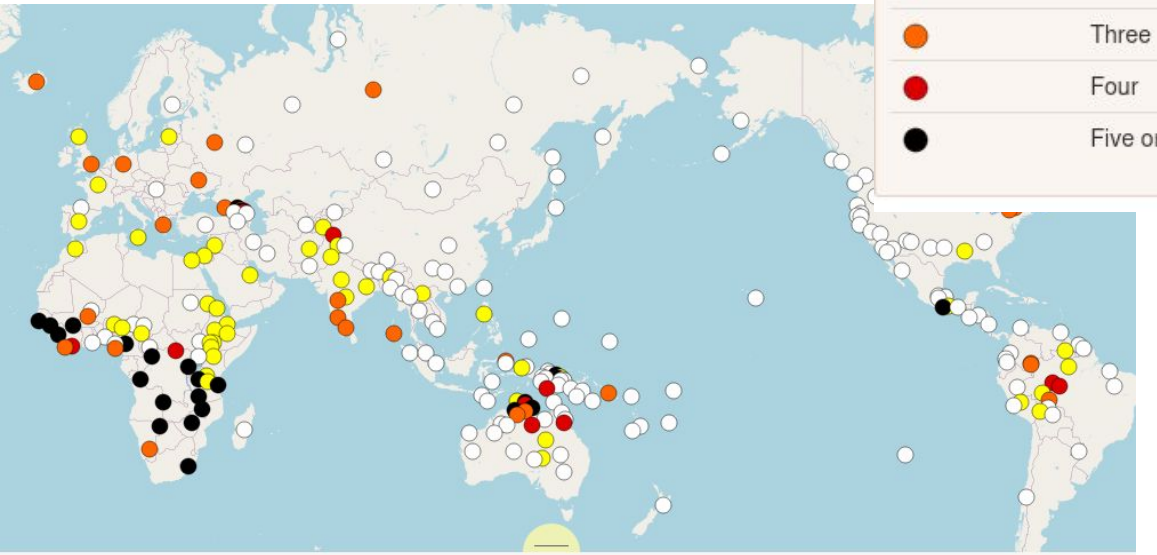


Accurate phylogeny of the IndoEuropean Family of languages



Morphosyntactic features

Feature 30A: Number of Genders



Values

○	None	145
●	Two	50
●	Three	26
●	Four	12
●	Five or more	24

Morphosyntactic features of languages

Language Name	20A	21A	21B	22A	23A	24A	25A	25B	26A	27A	28A	29A	30A	31A
Italian	1	5	2		4	2	5	2	2	3	3	2	2	2
Spanish	1	1	2		3	2	5	2	2	3	3	2	2	2
Catalan	1	1	2		3	2	5	2	2	3	3	2	2	2
French	1	5			4	2	5		2	3	3		2	2
Portuguese (Euro	1	5			4	2	5		2	3	3		2	2
Portuguese (Brazi	1	5			4	2	5		2	3	3		2	2
Romanian	1	1			3	2	5		2	3	3		3	2
Latin		2			2	2	2		2	1	3		3	2
Classical Greek, K		2			2	3	5		2	1	3		3	2
New Testament G		2			2	3	5		2	1			3	2
Salento Greek	1				2	3	5		2	3	3		3	2
Calabria Greek A	1				2	3	5		2	3			3	2

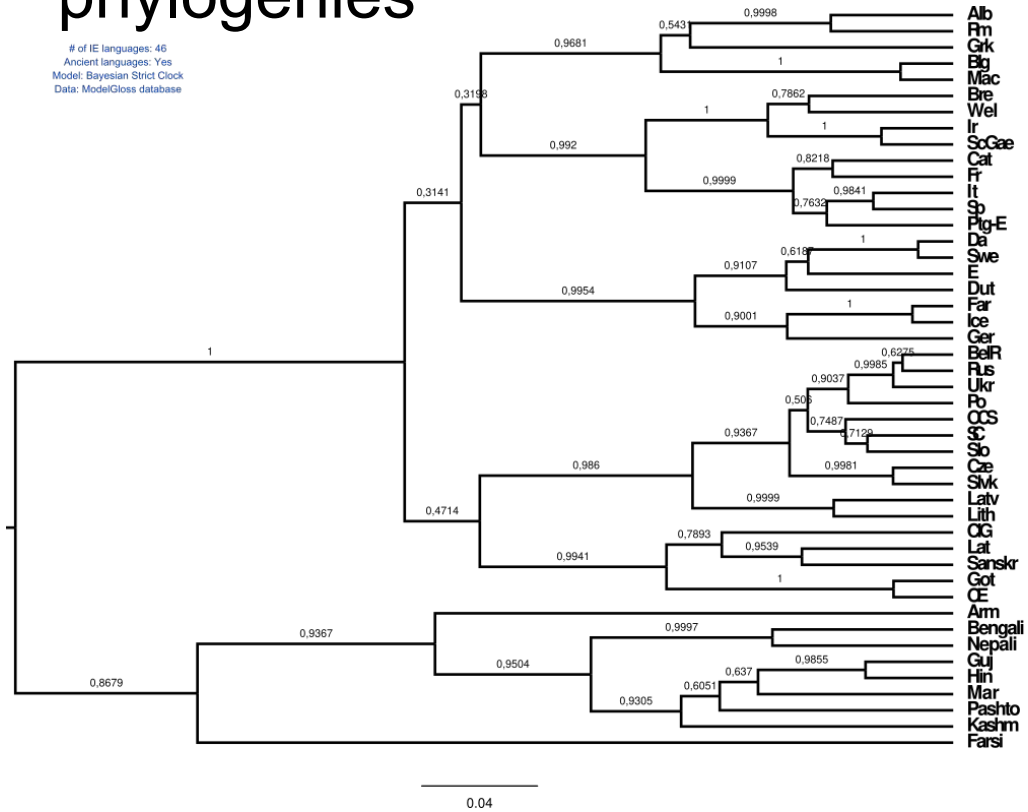
The morphosyntactic benefits and drawbacks

Morphosyntactic features can reconstruct **deep phylogenetic relations**

Morphosyntactic features drawbacks:

1. **horizontal transfer (between languages that are not related historically)**
2. **convergent evolution (independent evolution)**

Morphosyntactic features fail to reconstruct accurate phylogenies



Problems:

The effect of Geography

Alb, Rm, Grk, Blg, Mac are together (Sprachbund). They are not related historically

English should be attached to German/Dutch, not Danish/Swedish (Vikings + homoplasies)

Old languages

(Sanskr/Got/OE/Lat/CIG) should not be all together

Pashto should be together with Farsi

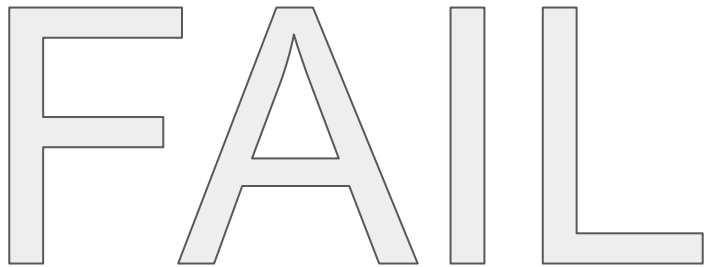
Can we identify features (sites) that produce a phylogeny as close as possible to the “cognate phylogeny”?

Idea 1:

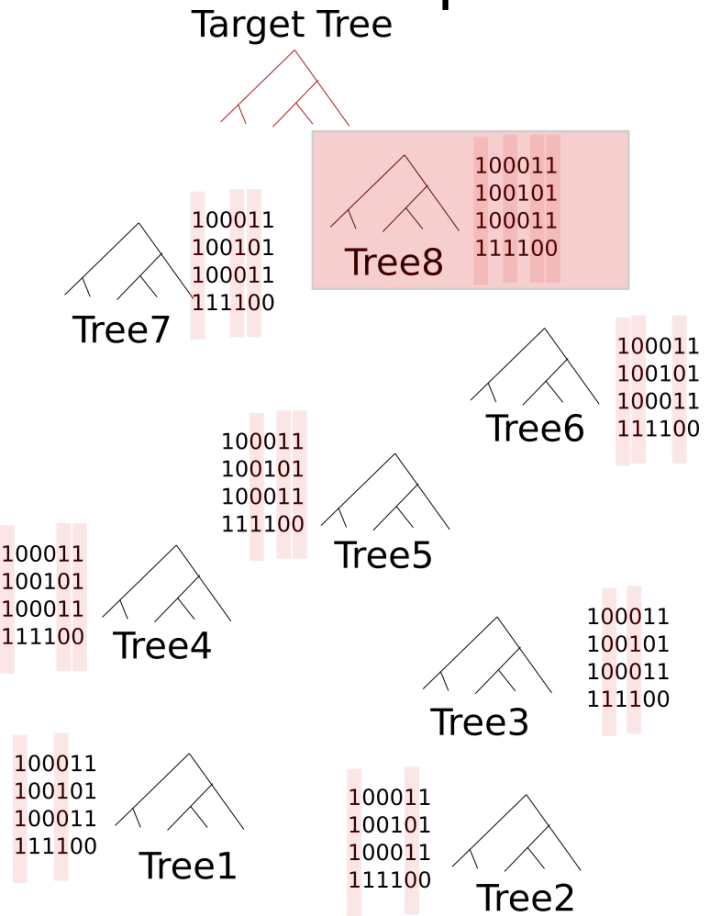
Use the cognate tree as a reference

Evaluate the likelihood of the cognate tree for each morphosyntactic site

Keep the sites that result in high likelihood



Can we identify features (sites) that produce a phylogeny as close as possible to the “cognate phylogeny”?



Is it possible to find the subset of sites that produce a phylogeny as close as possible to the target phylogeny?

ΕΛΙΔΕΚ
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

Can we identify features (sites) that produce a phylogeny as close as possible to the “cognate phylogeny”?

Start with a random subset of sites and build a phylogenetic tree

100011
100101
100011
111100

Tree1

Tree 1 is far from the target

Step 1: Propose a small change on the subset of sites

100011
100101
100011
111100

Tree2

Tree 2 is a bit closer to the target → KEEP the subset

Step 2: Propose a small change on the subset of sites

100011
100101
100011
111100

Tree3

Tree 3 is further than Tree 2→DISCARD the subset (keep subset2)

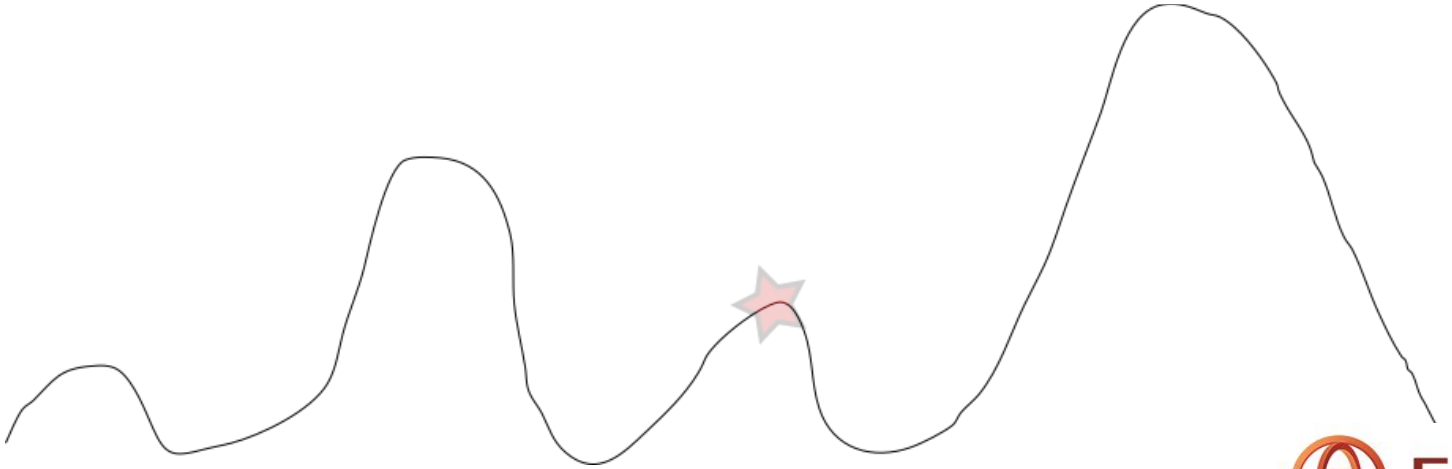
ΕΛΙΔΕΚ
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας

Hill Climbing Algorithm -- Tips and Tricks

This is a classical hill-climbing algorithm for optimization

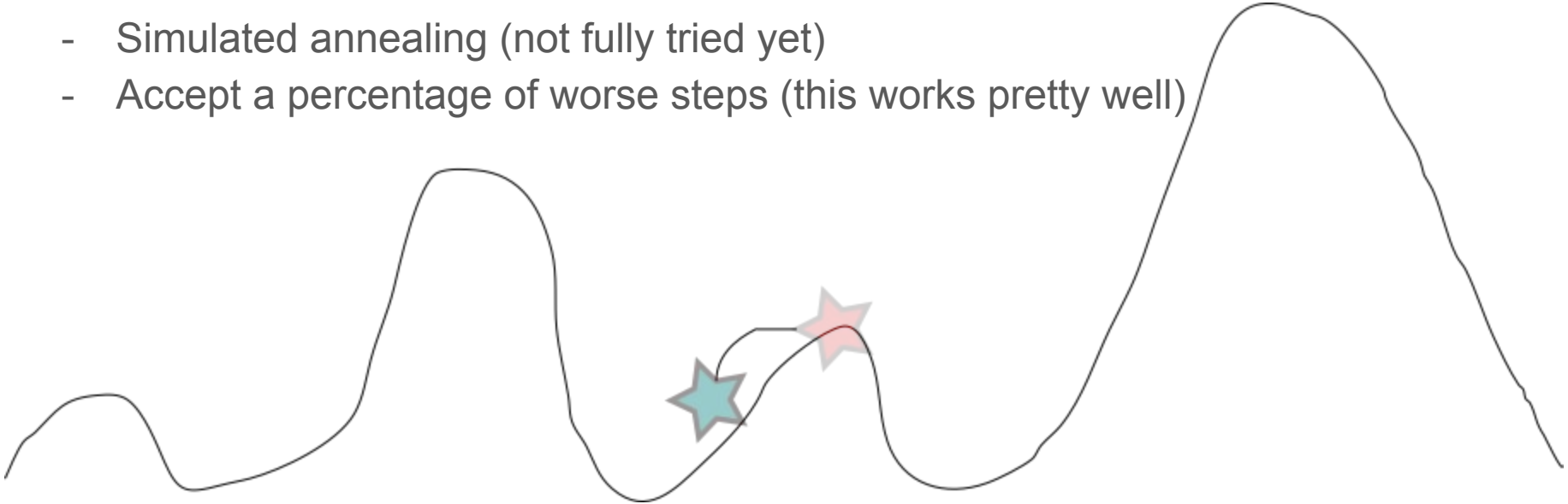
DRAWBACK Local Optimum

Whatever proposal no better tree is produced → the algorithm is stuck

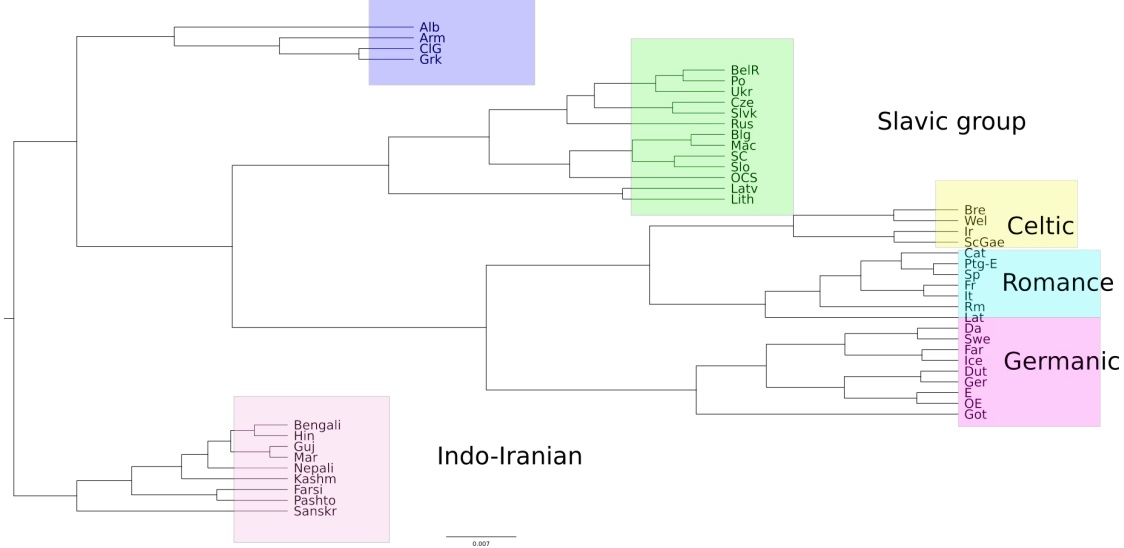


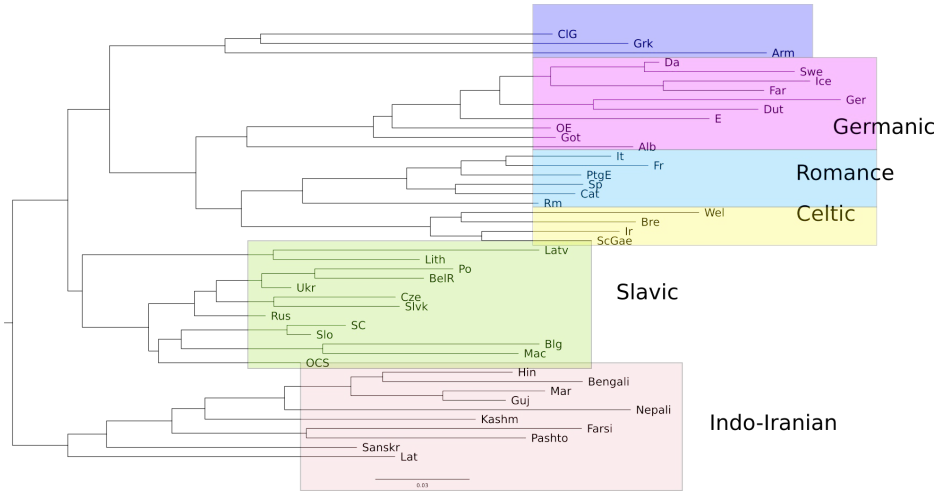
Hill Climbing Algorithm -- Tips and Tricks

- Simulated annealing (not fully tried yet)
- Accept a percentage of worse steps (this works pretty well)



Results





This tree is constructed with just 38 features out of 425

FORTH
INSTITUTE OF COMPUTER SCIENCE

Thank you!

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant” (Project Number: HFRI-FM17-44).

ΕΛ.Ι.Δ.Ε.Κ.
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας