



**Vilniaus
universitetas**

Doktorantas:
Brendonas Stakauskas
2022-2026

Darbo vadovas:
Dr. Virginijus Marcinkevičius

2025/2026 metai
IV metai, II pusmetis



**Giliais neuroniniais tinklais grįstų
mašininio mokymosi metodų
taikymas viruso mutacijų trajektorijai
prognozuoti**

Atnaujintas objektas ir tikslas

Tyrimo objektas:

Bendros paskirties iš anksto apmokyti baltymų kalbos modeliai ir jų taikymas filogenetinei analizei.

Tyrimo tikslas:

Sukurti metodą filogenetinio medžio topologijai atkurti be iš anksto nustatyto evoliucinio modelio.

Atnaujintas problemos

- nežinoma, ar bendros paskirties baltymų kalbos modelių įterpiniai išsaugo reikalingą kiekį evoliucinės informacijos;
- nenustatyta, kaip turi būti apdorojama modelio išvestis siekiant išlaikyti evoliucinę informaciją;
- neaišku, kaip atstumai įterpinių erdvėje susiję su evoliuciniais atstumais.

Atnaujinti uždaviniai

- Įvertinti įterpinių gebėjimą atkurti kladų struktūrą baltymų sekose.
- Įvertinti, ar atstumai įterpinių erdvėje atitinka klasikinių filogenetinių modelių evoliucinius atstumus.
- Nustatyti, ar fiksuoto ilgio įterpinių pakanka tiksliai filogenetinei analizei.
- Sukurti metodą filogenetinio medžio topologijai atkurti iš baltymų kalbos modelių įterpinių ir palyginti jį su klasikiais ir mašininio mokymo metodais.



Studijų plano vykdymas

Studijų planas, vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose				Publikacijos				
			Nacionalinėse		Tarptautinėse		Su citavimo rodikliu		Be citavimo rodiklio		Buklė
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	
I (2022/2023)	2	2	1	0	0	1	0	0	1	1 (konf. darbų medžiagoje)	
II (2023/2024)	2	2	1	0	0	1	0	0	1	0	
III (2024/2025)	0	0	0	1	1	0	1	1	0	1 (konf. darbų medžiagoje)	Publikuota žurnaleJournal of Computational Biology.
IV (2025/2026)	0	0	0	0	1	0	1	1	0	0	Publikuota žurnale Scientific Reports
Iš viso:	4	4	2	1	2	2	2	0	2	2	

Publikacijos

Publikacijos 2025/2026 (2 pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
Journal of Computational Biology	B. Stakauskas, V. Marcinkevičius. Evaluating Protein Language Model Embeddings for Viral Clade Assignment	Įteikta : 2025-10-10 Publikuota: 2026-04-21 DOI:10.1177/15578666261443336	CA WoS su Impact Factor (1.6)
Scientific Reports	B. Stakauskas, P. Górecki. Phylogenetic Signal in Protein Language Model Embeddings: A Zero-Shot Analysis of Distance-Based Tree Inference	Įteikta: 2025-12-31 Publikuota: 2026-06-17 DOI: 10.1038/s41598-026-57699-5	CA WoS su Impact Factor (3.9)

Evaluating Protein Language Model Embeddings for Viral Clade Assignment

[Brendonas Stakauskas](#)   and [Virginijus Marcinkevičius](#) [View all authors and affiliations](#)

[Volume 33, Issue 7](#) | <https://doi.org/10.1177/15578666261443336> | [View article versions](#)

scientific reports

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [scientific reports](#) > [articles](#) > article

Article | [Open access](#) | Published: 17 June 2026

Analysis of phylogenetic signal in protein language model embeddings

[Brendonas Stakauskas](#)  & [Paweł Górecki](#)

[Scientific Reports](#) **16**, Article number: 22051 (2026) | [Cite this article](#)

Eil. Nr.	eLABa ID ↕	Autorius ↕	Antraštė ↕	Šaltinis ↕	Publikacijos rūšis ↕	Institucija ↕	Metai ↕	Kiti ID ir URL	Redagavimo data ↕	Būsena ↕	Veiksmai
1	297451884	Brendonas Stakauskas; Virginijus Marcinkevičius	Evaluating protein language model embeddings for viral clade assignment	Journal of computational biology, vol. 33, iss. 7	S1 - Straipsnis Web of Science DB	Vilniaus universitetas	2026	10.1177/15578666261443336	2026-09-18 08:46	Patvirtintas	 
2	317857501	Brendonas Stakauskas; Paweł Górecki	Analysis of phylogenetic signal in protein language model embeddings	Scientific reports, vol. 16, iss. 1	S1 - Straipsnis Web of Science DB	Vilniaus universitetas	2026	10.1038/s41598-026-57699-5 Pagrindinis failas	2026-08-31 14:44	Patvirtintas	 

Visų mokslinių tyrimų ir disertacijos rengimo etapai

1. Mokslinių tyrimų disertacijos tema apžvalga ir analizė
2. Mokslinio tyrimo vykdymas
 1. Tyrimo metodikos sudarymas
 2. Teorinis tyrimas
 3. Empirinis tyrimas
 4. Gautų duomenų analizė, apibendrinimas, išvadų parengimas
3. Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas
4. **Daktaro disertacijos parengimas ir svarstymas padalinyje (70-80 proc.)**
5. Daktaro disertacijos gynimas

Baltymų kalbos
modelių taikymas
evoliucinei analizei be
papildomo mokymo

Baltymų seka

- Aminorūgščių grandinė
- Sudarytas iš 20 skirtingų aminorūgščių
- Nulemta DNR: nukleotidų trejetas koduoja vieną aminorūgštį (yra sinonimų)
- Nulemia struktūrą ir funkciją

Filogenetinė analizė

1. Sequences

seq 1 M K T L L V A G S D Q A
seq 2 M K T L L V A G S E Q A
seq 3 M K S L L A G T D Q A
seq 4 M K S L I V A G T D H A
seq 5 M K S L I V S G T D H A

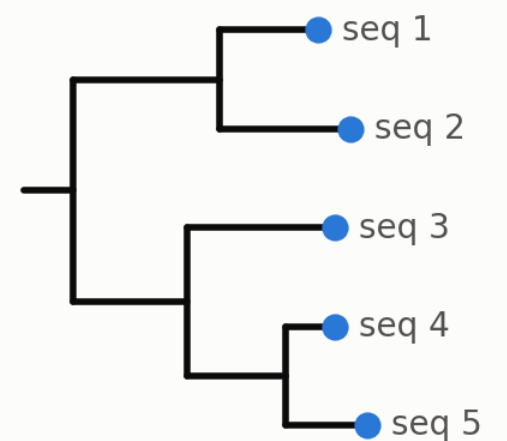
alignment
→

2. Multiple sequence alignment

seq 1 M K T L L V A G S D Q A
seq 2 M K T L L V A G S E Q A
seq 3 M K S L L - A G T D Q A
seq 4 M K S L I V A G T D H A
seq 5 M K S L I V S G T D H A

substitution model
+ tree search
→

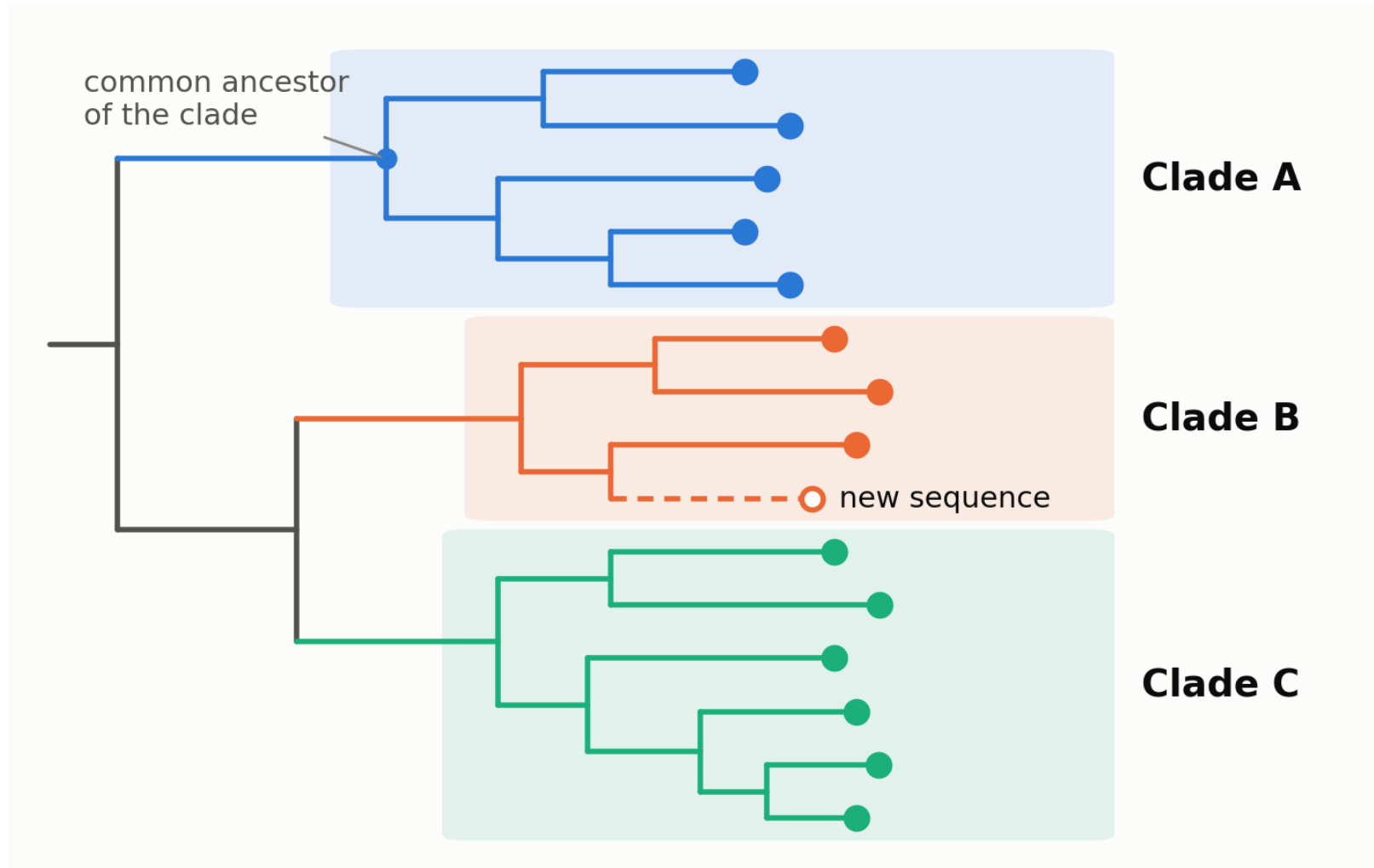
3. Phylogenetic tree



Filogenetinė analizė

- Atstumais grįsti metodai (NJ, UPGMA)
- Maximum Parsimony (MPBoot, TNT, PAUP*)
- Maximum Likelihood (RAxML, FastTree, IQ-TREE)
- Bajeso metodai (MrBayes, BEAST)

Filogenetinė analizė



Baltymų seka kaip kalba

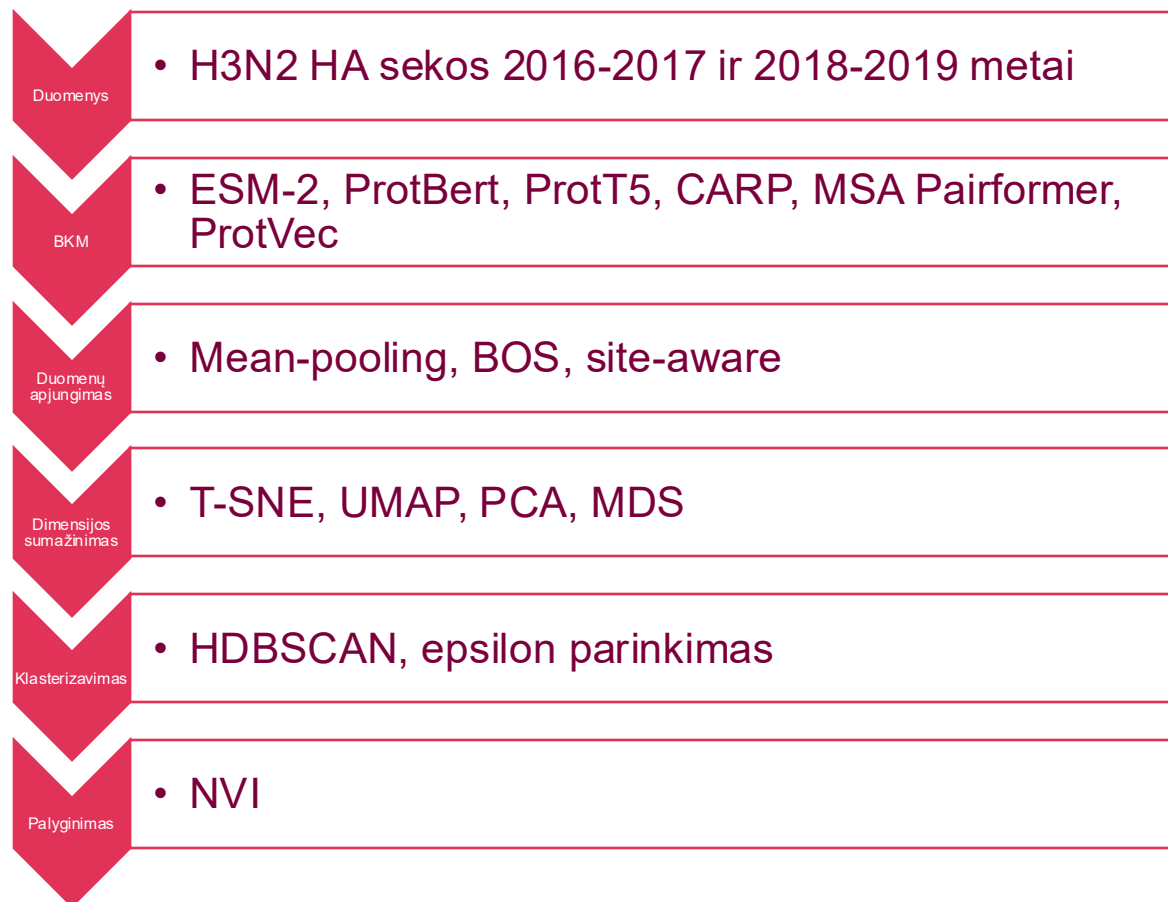
- Apibrėžta abėcelė iš 20 simbolių
- Žodžius ir frazes keičia motyvai ir domenai
- Turi atitikti taisykles, kad baltymas būtų funkcionalus
- Išlaiko reikšmę (nulemia struktūrą, funkciją)

Baltymų kalbos modeliai

- Kalbos modeliai pritaikyti baltymų sekoms
 - Enkoderiai ir dekoderiai
- Apmokyti be priežiūros
- Naudojami enkoderių sudaryti įterpiniai
- **Darbe naudoti:** ProtVec*, ESM-2, ESM-C, ProtBert, ProtT5, E1, CARP*, MSA Transformer, MSA Pairformer

*Ne transformeriai

Kladų priskyrimas



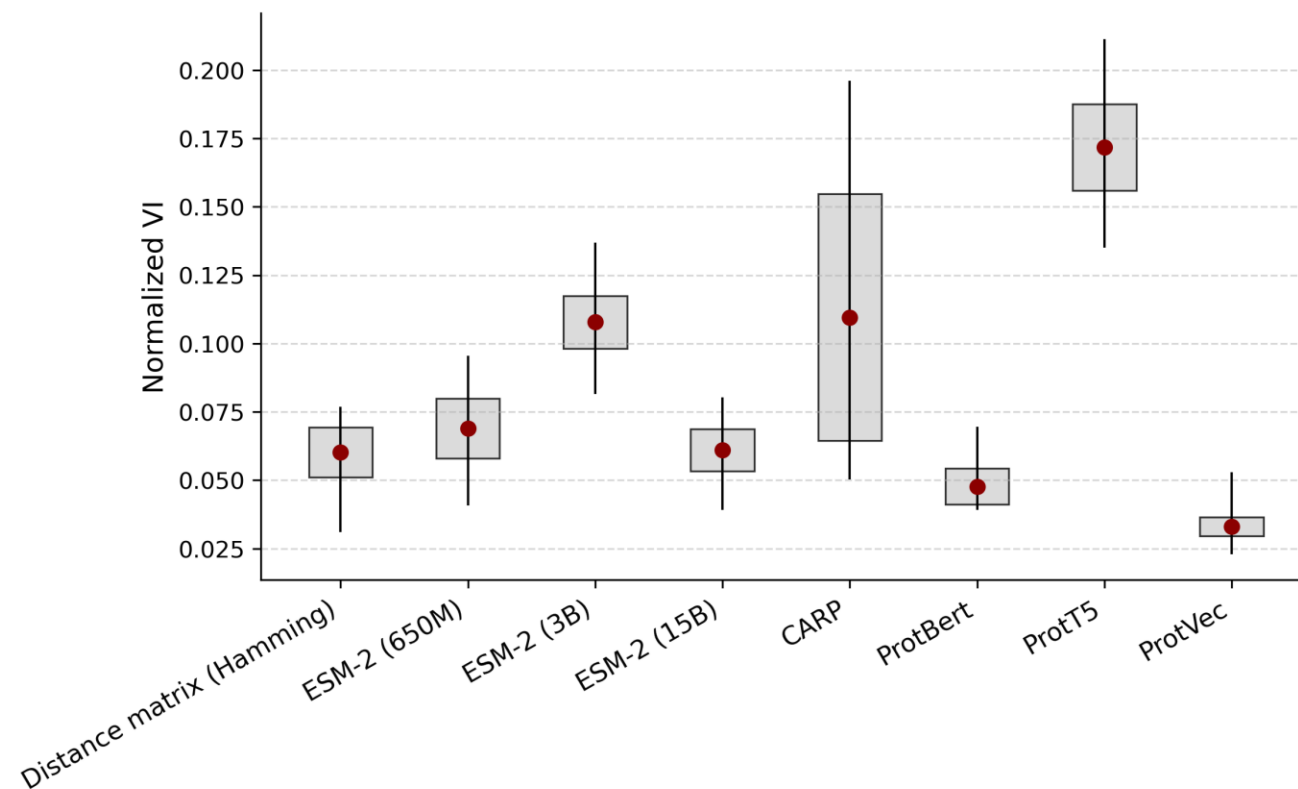
Duomenų rinkinys

Split	N	Clades	Within-clade Hamming	Between-clade Hamming
Train	1,523	11	0.0059	0.0158
Test	1,073	17	0.0054	0.0247

Atlikta 100 pakartojimų su atsitiktiniais duomenų poaibiais

Rezultatai

Model	NVI ↓
ProtVec	0.0405
ProtBert	0.0445
Baseline	0.0538
ESM-2 (650M)	0.0544
ESM-2 (15B)	0.0562
CARP	0.0728
ESM-2 (3B)	0.1058
ProtT5	0.1155

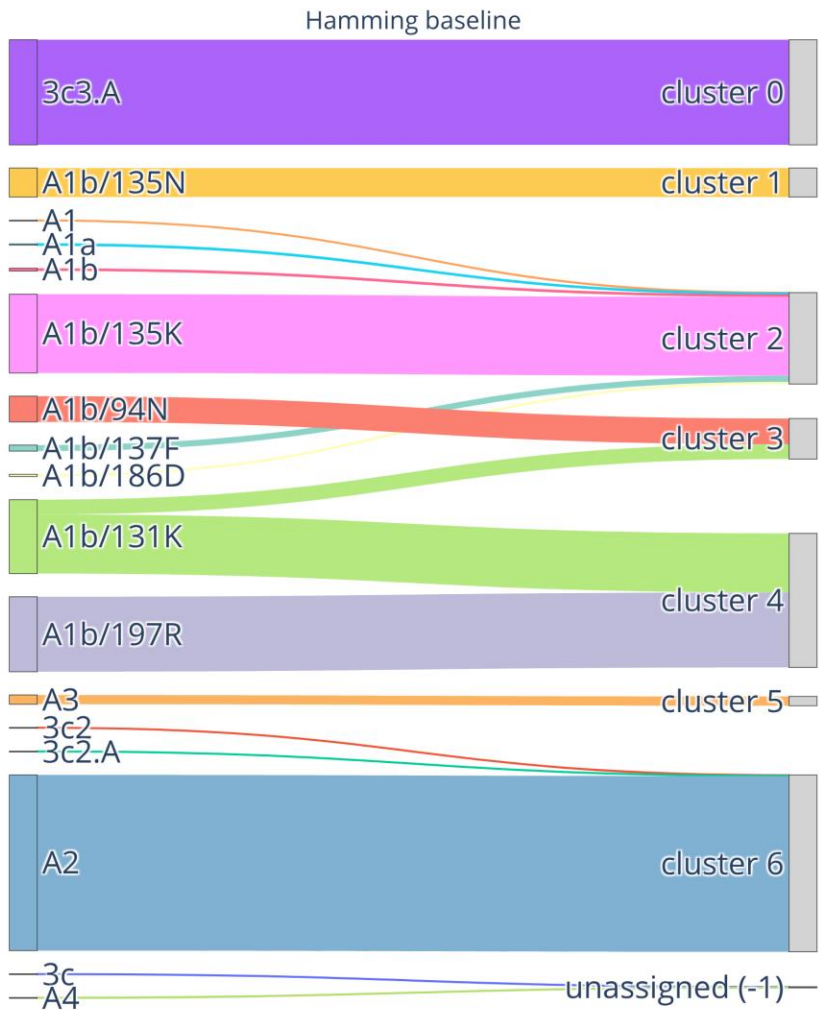
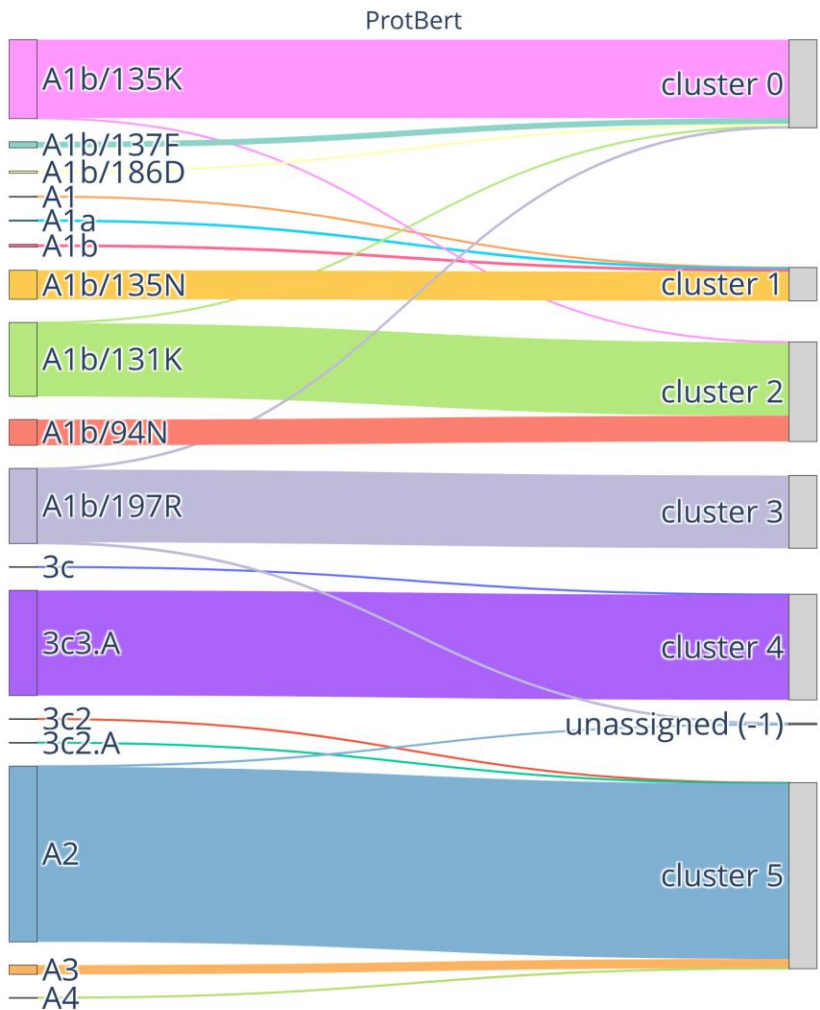
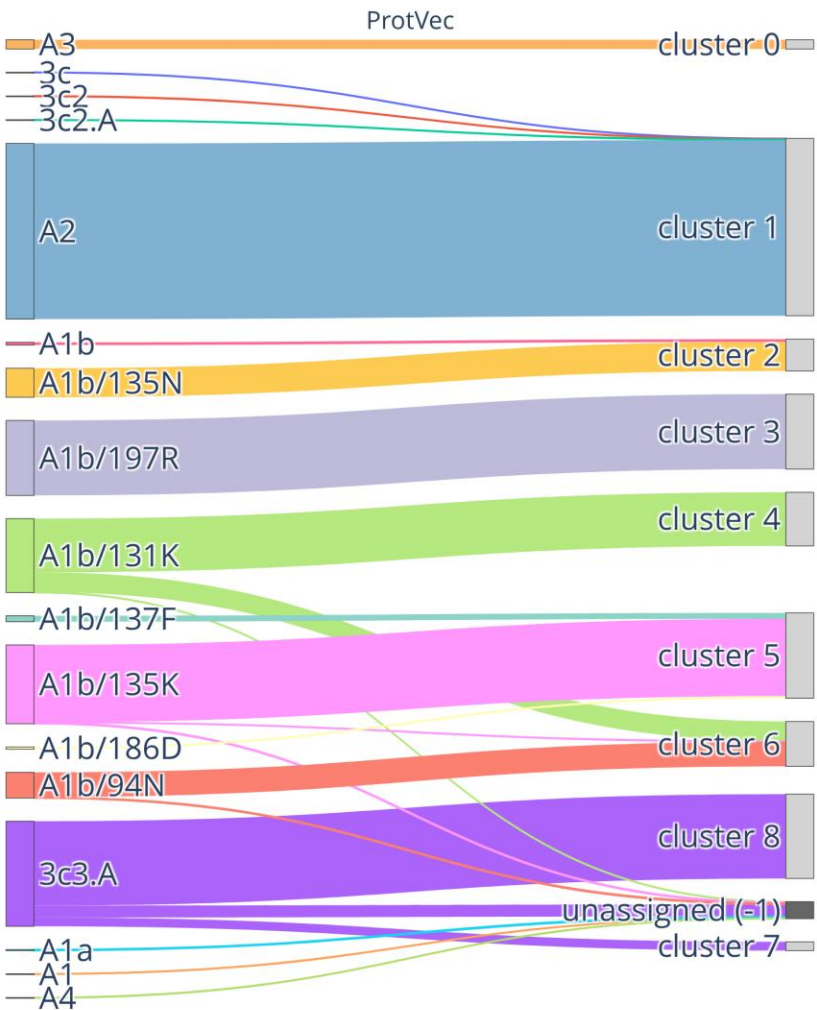


Atmesti duomenys

2016-2017

Clade	<i>N</i>	ProtVec	ProtBert
A1	359	54	3
A2	273	19	3
A1a	181	25	–
A1b	156	29	–
A3	138	29	–
3c2.A	137	44	–
3c3.A	117	–	–
A1b/135K	64	16	–
A1b/135N	48	14	–
A4	41	–	–
3c3	9	1	–
Total	1523	231	6

Clade	<i>N</i>	ProtVec	ProtBert
A2	321	–	1
3c3.A	192	22	–
A1b/135K	144	2	–
A1b/197R	137	–	2
A1b/131K	135	1	–
A1b/135N	53	–	–
A1b/94N	47	2	–
A3	17	–	–
A1b/137F	11	–	–
A1b	5	–	–
A1b/186D	4	–	–
A1a	2	2	–
A4	1	1	–
A1	1	1	–
3c	1	–	–
3c2.A	1	–	–
3c2	1	–	–
Total	1073	31	3



Tarpinės išvados

- Įterpiniai atkuria kladų struktūrą panašiai kaip Hamingo atstumo metodas
- Didesnis modelis nereiškia geresnio rezultato
- Geriausi rezultatai turi išlygų: ProtVec atmeta dalį sekų, ProtBert sulieja mažus kladus

Filogenetinio medžio rekonstravimas

Duomenų rinkiniai:

- 500 skirtingų baltymų šeimų iš PANTHER DB
- Generuotos sekos naudojant SimPhy ir AliSim.

Modeliai:

ProtVec, ProtBert, ProtT5, ESM-C, E1

MSA Transformer, MSA Pairformer, E1*

Medžio rekonstravimas:

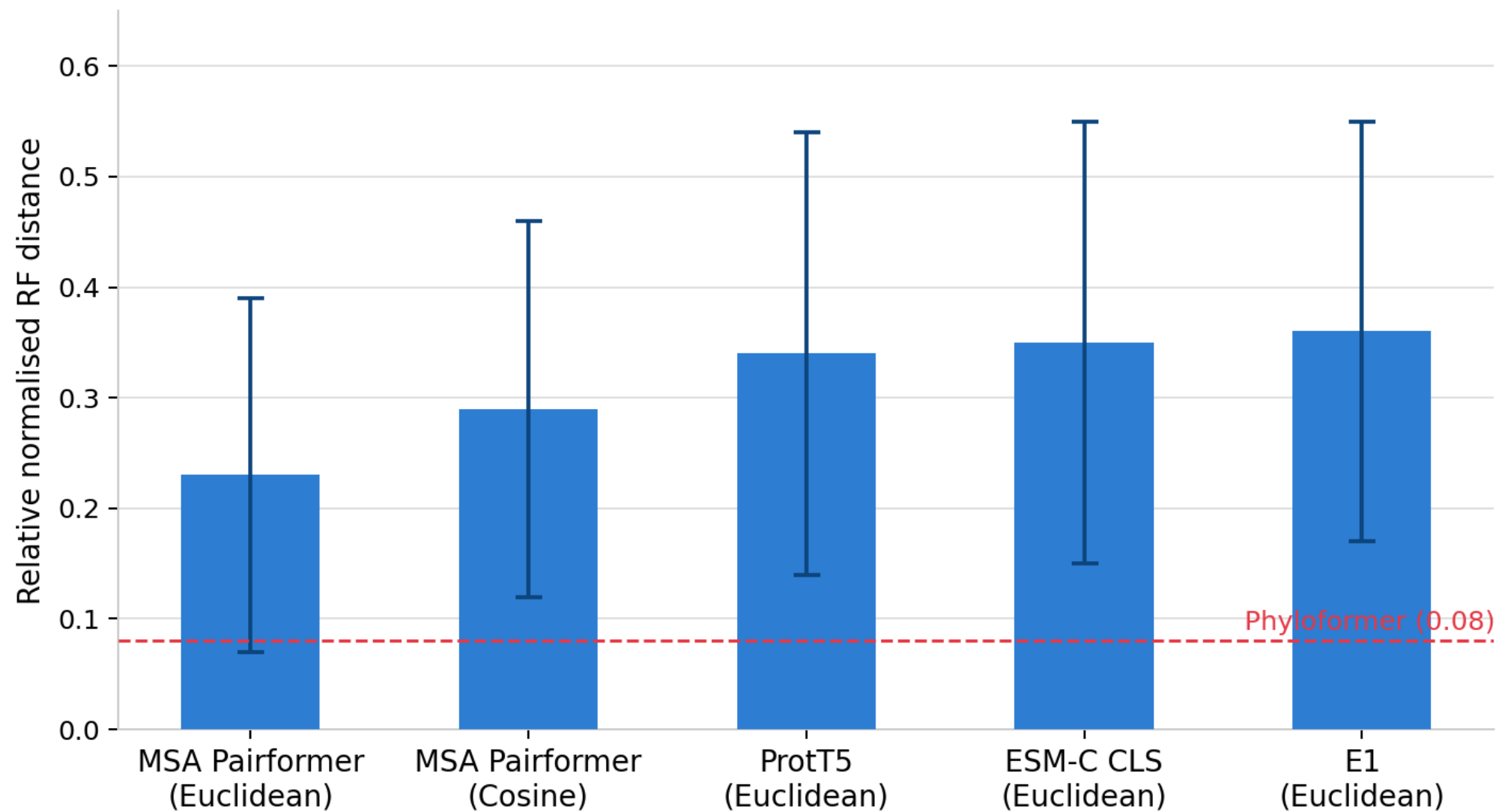
Įterpiniai -> Atstumai tarp sekų -> NJ medis

Duomenų vektoriaus sudarymas:

- Vienas vektorius sekai (vidurkis, CLS)
- Atstumai suagreguoti per pozicijas
 - Slenkantis langas
 - Padengiamumas
 - Atsitiktiniai poaibiai

Vertinimas:

- Topologiškai: RF, Nye, SPI
- Šakų ilgiai: KF
- Lyginama su: NJ, IQ-TREE, FastTree, Phyloformer, Phyla, NeuralNJ



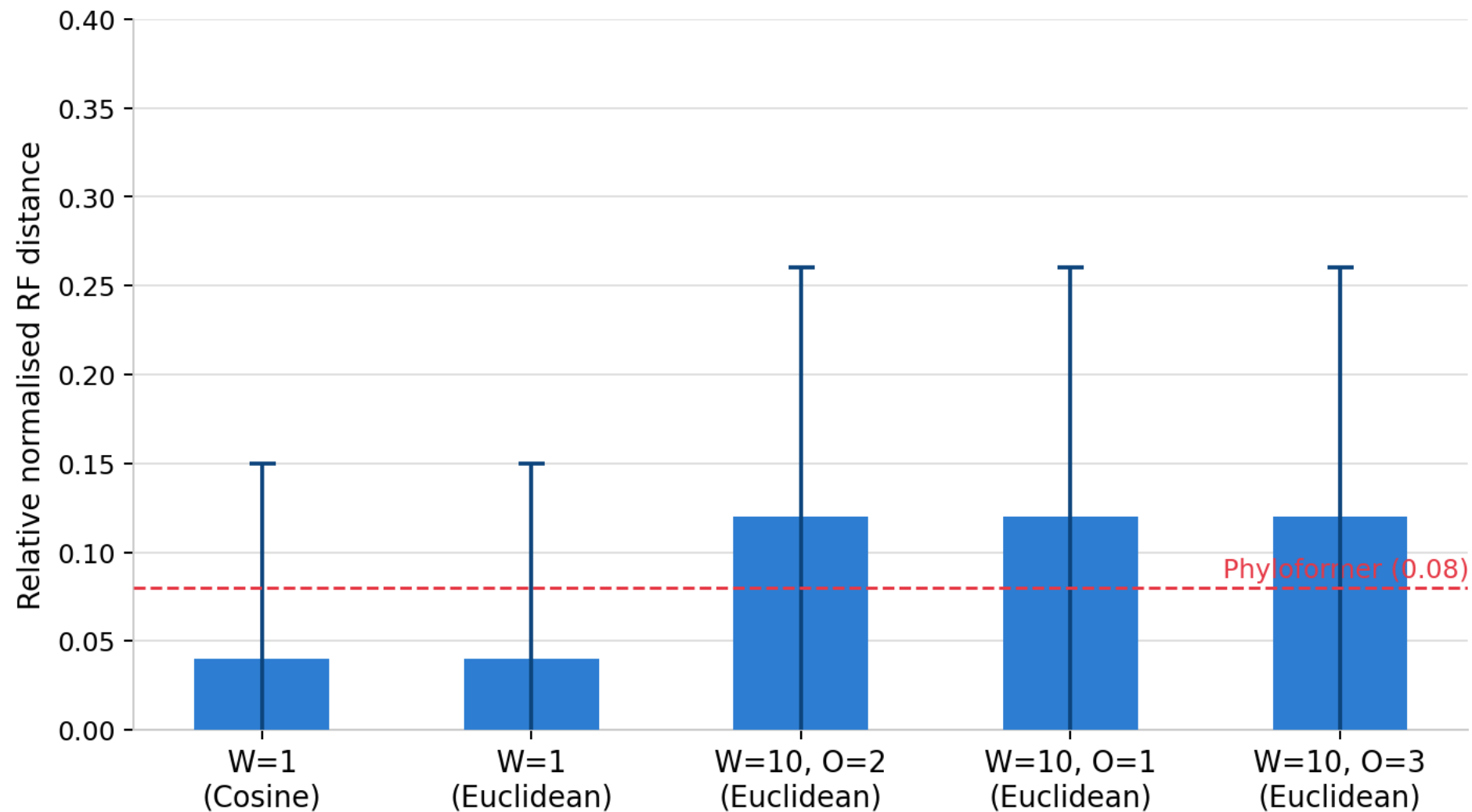
Slenkantis langas

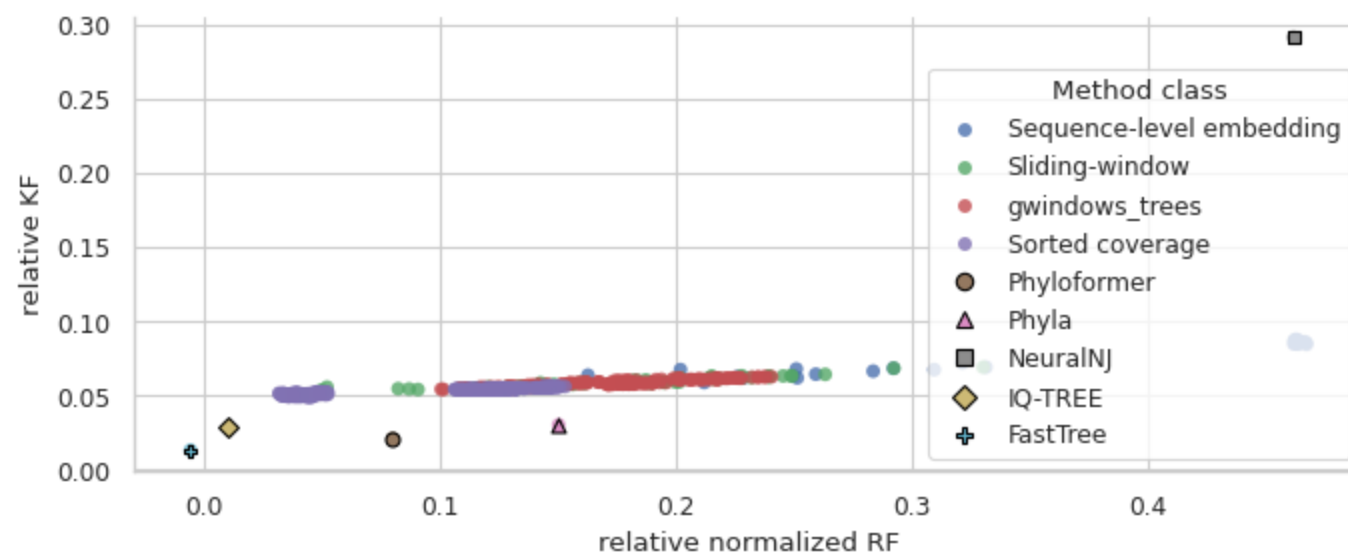
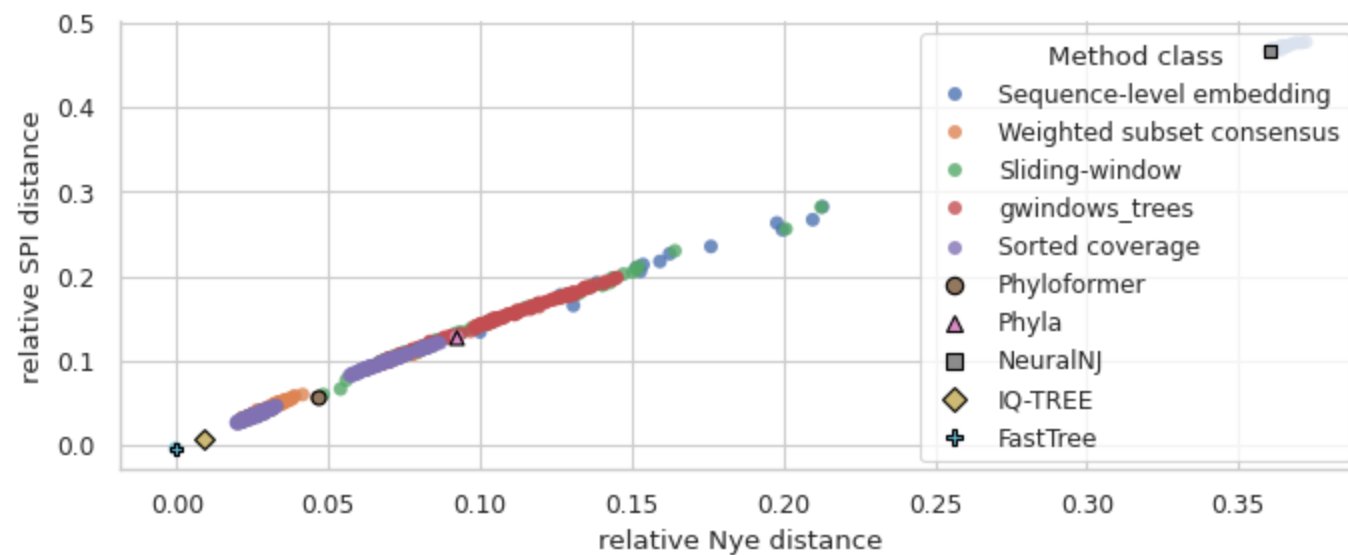
- Sudaromi įterpiniai kiekvienai pozicijai
- Sudaromi persidengiantys langai
- Apskaičiuojamas atstumas kiekvienam langui
- Susumuojami visi gauti atstumai
- Išvedamas vidurkis

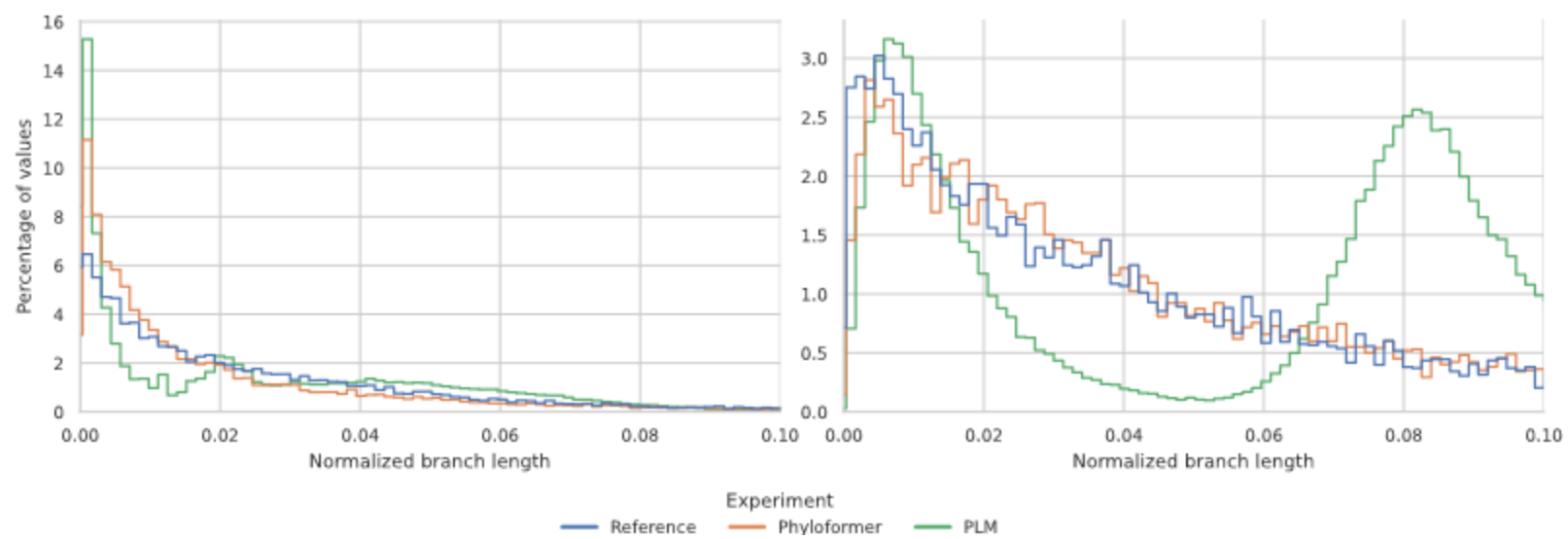
Ekstremalus atvejis: $N=1$

Padengiamumas

- Kiekvienai pozicijai (stulpeliui) priskiriamas svoris
 - Vienetinis
 - Skirtingų aminorūgščių skaičius
 - Šenono entropija
- Imami svarbiausi stulpeliai iki nustatyto svorio







Išvados

- Įterpiniuose egzistuoja filogenetinis signalas
- Įterpinių vidurkinimas šį signalą susilpnina
- Agreguojant atstumus pagal pozicijas galima atkurti medžio topologiją tiksliau, nei Phyloformer, Phyla ar NeuralNJ metodais
- Atstumai įterpinių erdvėje neatitinka evoliucinių atstumų