Vilniaus universitetas

Doktorantas:
Brendonas Stakauskas
2022-2026

Darbo vadovas:
Dr. Virginijus Marcinkevičius

2022/2023 metai
I metai, I pusmetis

Giliais neuroniniais tinklais grįstų mašininio mokymo metodų taikymas viruso mutacijų trajektorijai prognozuoti

Vilniaus universitetas

TURINYS

1. Studijų plano vykdymas
2. Problemos apibrėžimas, tyrimo objektas ir tikslai
3. Trumpas per pusmetį gautų mokslinių rezultatų pristatymas
4. Kito pusmečio darbo planas

Studijų plano vykdymas

1

Vilniaus universitetas

Studijų planas, vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose		Publikacijos					
					Konferencijos darbų medžiagoje			CA WoS su Impact Factor		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2022/2023)	**2**	**1**	**1**	**0**	**1**	**0**		**0**	**0**	
II (2023/2024)	2	0	1	0	0	0		0	0	
III (2024/2025)	0	0	1	0	0	0		1	0	
IV (2025/2026)	0	0	1	0	0	0		1	0	

Vilniaus universitetas

Ataskaitinio pusmečio darbo planas ir jo įvykdymas

Egzaminai 2022/2023 (I pusmetis)		
Planas	Įvykdyta	Būklė
Mašininis mokymasis	2023-03-02	Egzaminas išlaikytas.

Vilniaus universitetas

Visų mokslinių tyrimų ir disertacijos rengimo etapai

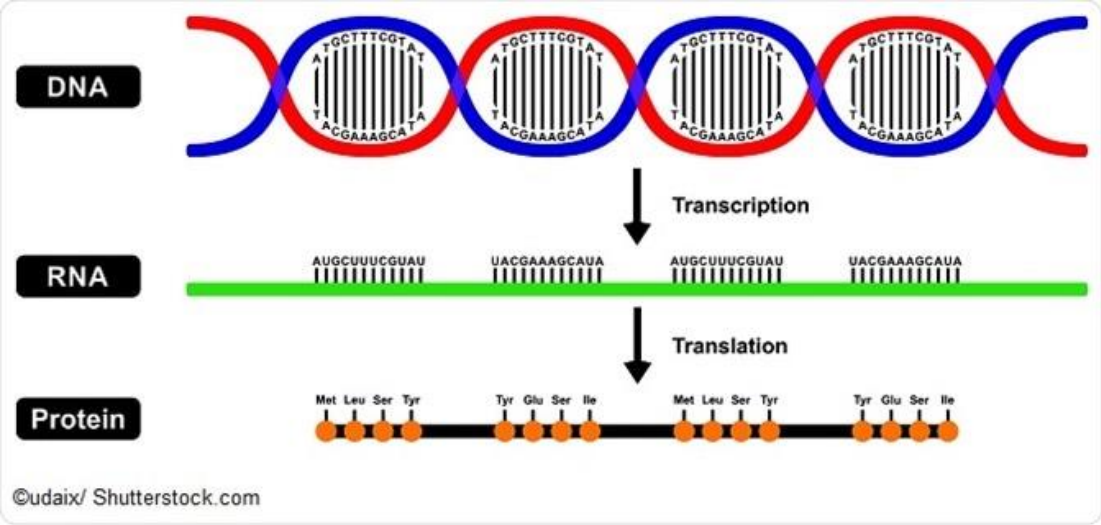
	Darbo pavadinimas	Atlikimo terminai	Pastabos
1.	**Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):** **1. Disertacijos tyrimo objekto detalizavimas.** **2. Giliais neuroniniais tinklais grįstų mašininio mokymo metodų (RNN, GRU, LSTM, Transformer ir pan.), naudojamų viruso genomo sekų mutacijoms prognozuoti, apžvalga ir analizė.**	**2022 m. spalio mėn. - 2023 m. vasario mėn.**	**Atlikimo etapai:** **1) Atliktas tyrimo objekto detalizavimas nustatant bendras disertacijos rašymo užduotis** **2) Atlikta algoritmų apžvalga ir susijusių darbų analizė**
	3. Metodų apžvalgos apibendrinimas ir pateikimas disertacijos analitinės dalies aprašyme. 4. Tyrimo tikslo formulavimas.	2023 m. vasario mėn. - 2023 m. rugsėjo mėn.	

Problemos apibrėžimas, tyrimo objektas ir tikslai

Vilniaus universitetas

Tyrimo objektas

Virusų baltymų sekos ir giliaisiais neuroniniais tinklais grįsti mašininio mokymo algoritmai skirti prognozuoti viruso mutacijas.



Baltymai, tai iš 20 skirtingų amino rūgščių sudarytos sekos. Paprasčiausia baltymo struktūra išreiškiama simbolių eilute.

Tyrimo metu nagrinėjama pirminė baltymo struktūra – simbolių seka (kalbiniame kontekste analogiška atskiroms raidėms).

Replikuojantis virusams atsiranda klaidų. Ilgainiui dėl šių klaidų sutrinka imuninis atsakas (virusas nėra identifikuojamas) ar nebeveikia vaistai (veiklioji medžiaga negeba prisijungti prie baltymo).

Vilniaus universitetas

Tyrimo problemos

- **Sąryšių nustatymas tarp viruso proteinų mutacijų sekų;**
- Sąryšių aibės, atspindinčios istorines mutacijas, sudarymas;
- Efektyvaus mašininiu mokymusi grįsto metodo sudarymas mutacijų prognozavimui;

Mašininio mokymo modelio apmokymui reikalingos duomenų aibės sudarymas nėra trivialus uždavinys. Aibė turi būti sudaryta taip, kad atspindėtų ryšius tarp viruso variantų. Literatūroje pateikiami keli metodai aibės sudarymui:

1. Grįstas klasteriais (Yin et al. 2020)
2. Grįstas filogenetiniu medžiu* (Zhou et al. 2023)

*Kai kurioms duomenų aibėms šie duomenys gali būti nepriskirti

Vilniaus universitetas

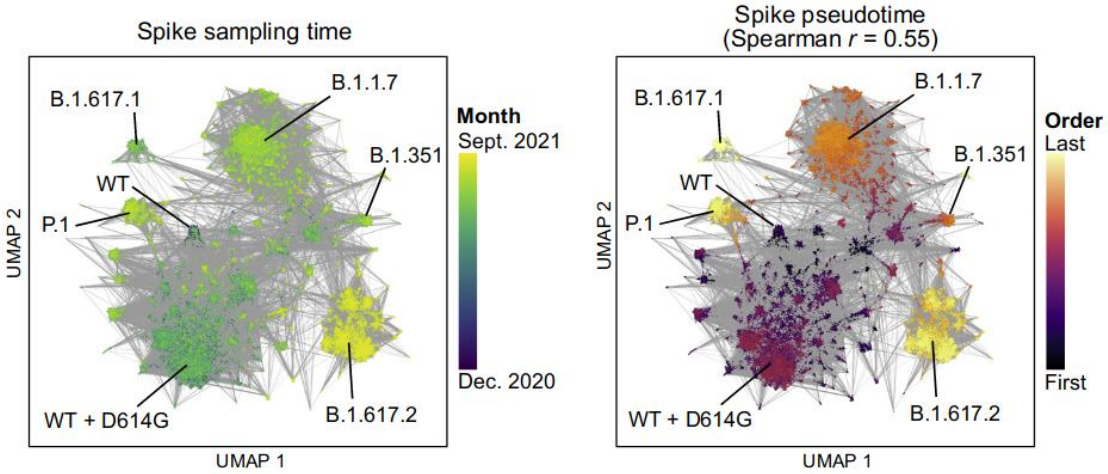
Preliminarūs tyrimo uždaviniai

- Atlikti literatūros analizę, išanalizuoti *state-of-the-art* algoritmus viruso baltymo mutacijų prognozavimui.
- Atkartoti literatūroje pateikiamų metodų rezultatus.
- Sukurti metodą duomenų aibės, atspindinčios istorines mutacijas, sudarymui.
- Surinkti duomenų aibę tyrimui.
- Pasiūlyti naują mašininiu mokymusi grįstą metodą viruso mutacijoms numatyti.
- Atlikti eksperimentinius tyrimus, palyginant pasiūlytą metodą su literatūroje aprašytais metodais.

Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

3

Apžvelgti dabar literatūroje taikomi metodai



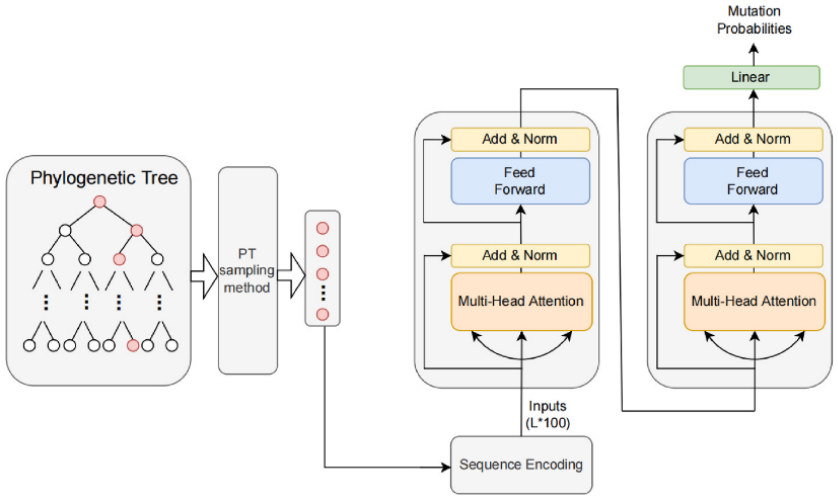
Hie et al. 2022 m. straipsnyje pasinaudojęs proteino kalbos modelio priskirtomis reikšmėmis išreiškia baltymo mutacijos „pseudolaiką“, kuris koreliuoja su tikruoju mutacijos atsiradimo laiku.

Brian L. Hie, Kevin K. Yang, and Peter S. Kim.

Evolutionary velocity with protein
language models predicts evolutionary dynamics of diverse proteins.

Cell Systems,
13(4):274–285.e6, April 2022.

Apžvelgti dabar literatūroje taikomi metodai



Zhou et al. 2023 m. straipsnyje siūlo pasinaudoti filogenetine informacija sudarant modelio įvesties aibę. Modelio rezultatas – binarinė reikšmė nusakanti ar parinktoje pozicijoje įvyks mutacija ar ne.

Binbin Zhou, Hang Zhou, Xue Zhang, Xiaobin Xu, Yi Chai, Zengwei Zheng,
Alex Chichung Kot, and Zhan Zhou.

TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution.

Computers in Biology and Medicine,
152:106264, January 2023.

Identifikuoti duomenų rinkiniai

Vilniaus universitetas

Article name	Virus	Source	Dataset size
Long Short-Term Memory Neural Networks for RNA Viruses Mutations Prediction [7]	Newcastle disease Influenza	Liu et al. [6] Bao et al. [2] (NCBI IVD)	83 Newcastle 4609 H1N1
Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks [15]	Influenza (H1N1, H3N2, H5N1)	Bao et al. [2] (NCBI IVD)	8470-H1N1 7703-H3N2 2213-H5N1
The prediction of virus mutation using neural networks and rough set techniques [11]	Newcastle disease	Sayers et al. [12] (NCBI GenBank)	22 Korea 45 China
TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution [15]	SARS-CoV-2 Influenza (H1N1, H3N2, H5N1)	Shu et al. [13] (GISAID) Song et al. [14] (RCoV19) Bao et al. [2] (NCBI IVD)	7 million SARS-COV-2 8470-H1N1 7703-H3N2 2213-H5N1

Taikomų metodų rezultatų palyginimas

Vilniaus universitetas

Method	Accuracy	Precision	Recall	F-score	MCC
SVM	0.530	0.519	0.588	0.551	0.063
LR	0.542	0.530	0.575	0.552	0.085
RF	0.544	0.534	0.561	0.547	0.089
RNN	0.609	0.581	0.720	0.643	0.226
LSTM	0.648	0.619	0.731	0.670	0.302
Tempel	0.648	0.618	**0.743**	**0.675**	0.305
TEMPO	**0.655**	**0.658**	0.614	0.636	**0.309**

Mutacijų nustatymo uždavinys nėra plačiai nagrinėjamas, nėra sudarytų etaloninių duomenų aibių.
Zhou et al. lygino savo rezultatus su panašiu Yin et al. siūlomu metodu, gauti labai panašūs rezultatai.

Binbin Zhou, Hang Zhou, Xue Zhang, Xiaobin Xu, Yi Chai, Zengwei Zheng,
Alex Chichung Kot, and Zhan Zhou.

TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution.

Computers in Biology and Medicine,
152:106264, January 2023.

Kito pusmečio darbo planas

4

Vilniaus universitetas

II pusmečio darbų planas:

Planuojama veikla	Atlikimo terminai
• Metodų apžvalgos apibendrinimas ir pateikimas disertacijos analitinės dalies aprašyme. • Tyrimo tikslo formulavimas.	2023 m. vasario mėn. - 2023 m. rugsėjo mėn.
• Dalyvavimas konferencijoje Lietuvoje.	2023 m. rugsėjo mėn.
• Mokslinių tyrimų disertacijos tema apžvalga (konferencijos darbų medžiagoje).	2023 m. rugsėjo mėn.