Vilniaus universitetas

Doktorantas:
Brendonas Stakauskas
2022-2026

Darbo vadovas:
Dr. Virginijus Marcinkevičius

2023/2024 metai
II metai, I pusmetis

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

Vilniaus universitetas

TURINYS

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

Problemos apibrėžimas, tyrimo objektas ir tikslai

1

Vilniaus
universitetas

Tyrimo objektas

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

Vilniaus universitetas

Tyrimo problemos

- Sąryšio tarp viruso proteinų sekų nustatymas;
- Duomenų aibės, atspindinčios istorines mutacijas, sudarymas;

Vilniaus
universitetas

Tyrimo tikslas

Sukurti giliaisiais neuroniniais tinklais ir natūralios kalbos apdorojimo algoritmais grįstą metodą leidžianti numatyti viruso mutacijos trajektoriją.

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.
- Sukurti duomenų aibę tyrimui.
- Pasiūlyti naują metodą viruso mutacijoms numatyti.
- Atlikti eksperimentinius tyrimus, palyginant pasiūlytą metodą su literatūroje aprašytais metodais.

Studijų plano vykdymas

2

Vilniaus universitetas

Studijų planas, vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose				Publikacijos				
			Lietuvoje		Užsienyje		Su citavimo rodikliu		Be citavimo rodiklio		Būklė
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	
I (2022/2023)	2	2	1	0	0	1	0	0	1	1 (konferencijos darbų medžiagoje)	
II (2023/2024)	**2**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	
III (2024/2025)	0	0	0	0	1	0	1	0	0	0	
IV (2025/2026)	0	0	0	0	1	0	1	0	0	0	

Vilniaus universitetas

Ataskaitinio pusmečio darbo planas ir jo įvykdymas

Egzaminai 2023/2024 (I pusmetis)		
Planas	Įvykdyta	Būklė
Fundamentalieji informatikos ir informatikos inžinerijos mokslų metodai	2024-01-24	Egzaminas išlaikytas.

Vilniaus universitetas

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
5. Daktaro disertacijos gynimas

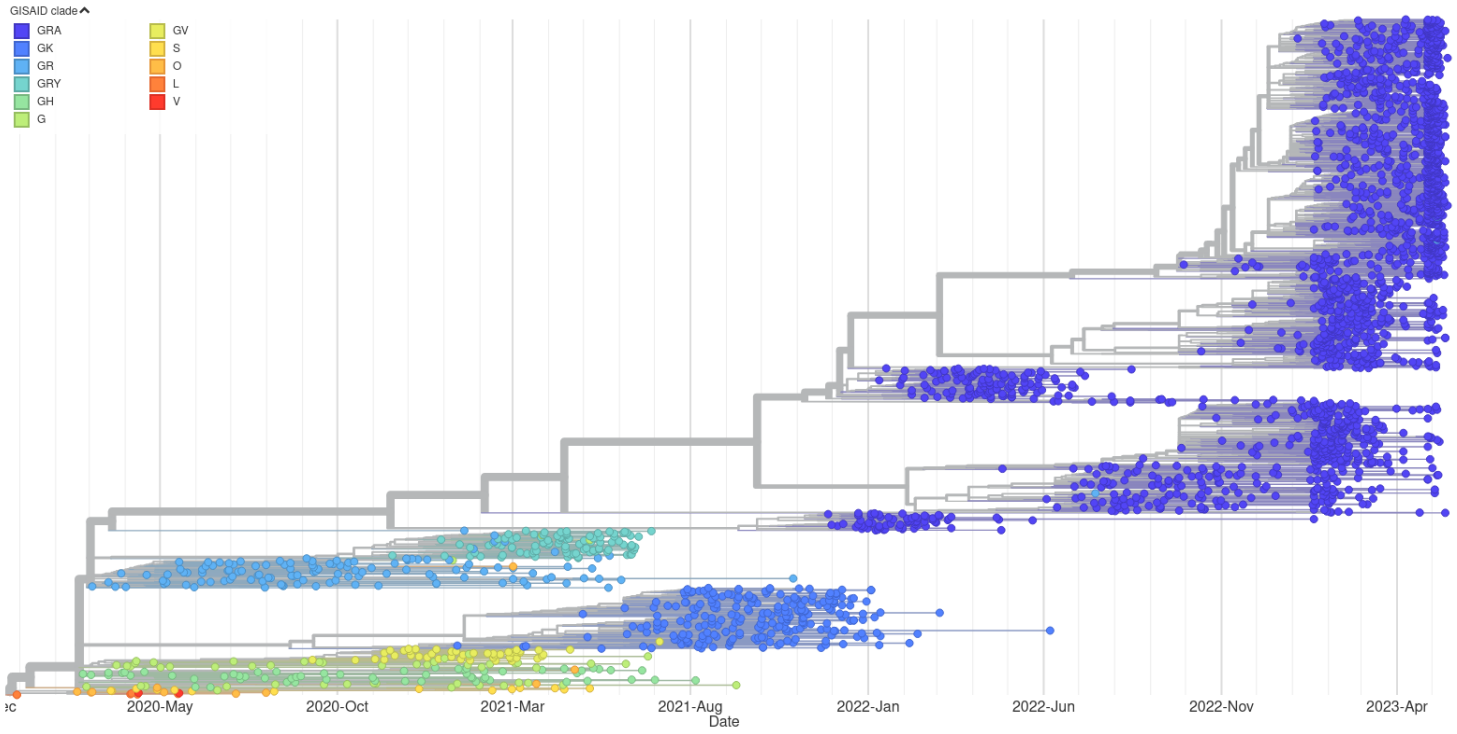
Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

3

Vilniaus universitetas

Filogenetiniai medžiai

Filogenetinis medis, tai medis nurodantis evoliucinius ryšius tarp įvairių organizmų.



Vilniaus universitetas

Duomenų aibė

- Duomenų aibę sudaro 2429 Covid19 baltymų sekos.
- Sekų surinkimo laikotarpis: 2019-12-17 – 2023-05-15.
- Medžio sudarymo algoritmas IQ-TREE, pagal nextstrain sudarytus nustatymus.

Vilniaus universitetas

ProtVec

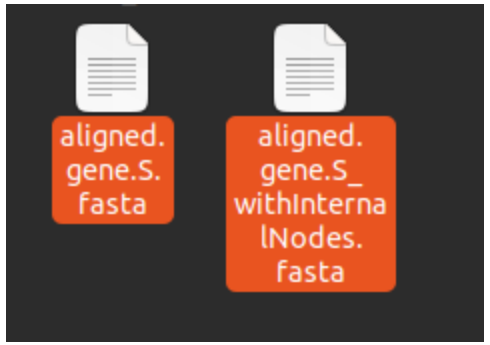
ProtVec, tai metodas leidžiantis atvaizduoti baltymų sekas kaip vektorius.

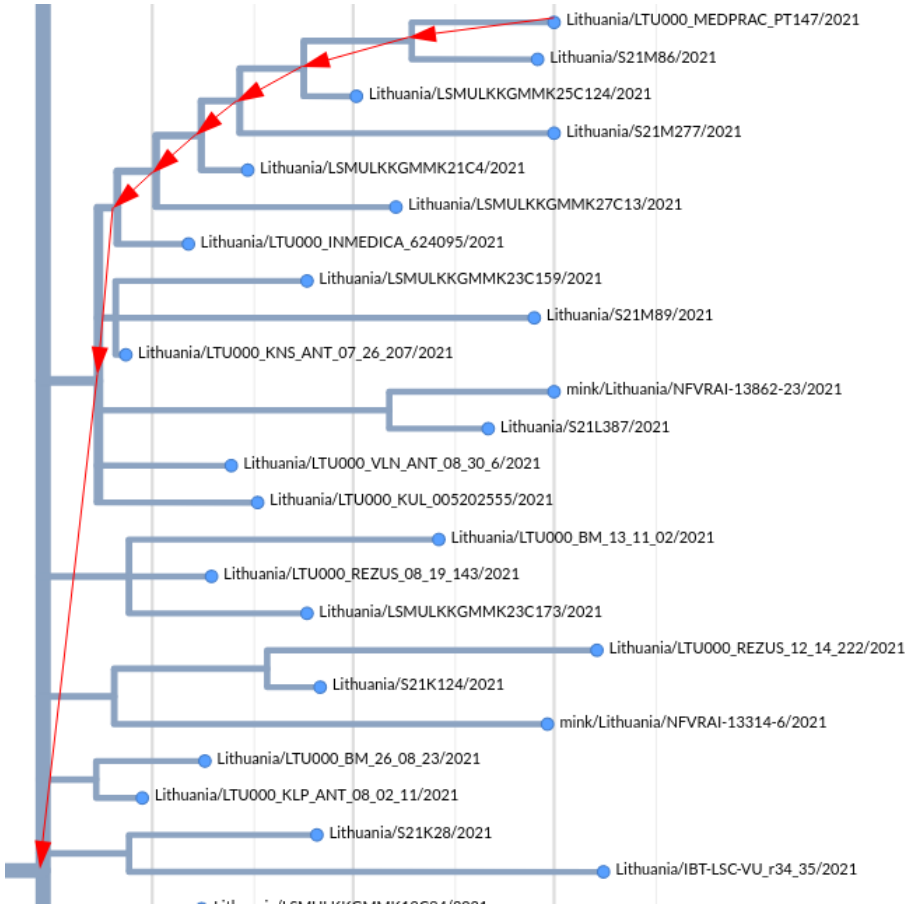
Tyrime naudojama iš anksto sudaryta amino rūgščių trigramų svorių aibė. Trigrama atvaizduojama kaip 100 reikšmių vektorius.

words	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11
AAA	-0.17406	-0.095756	0.059515	0.039673	-0.375934	-0.115415	0.090725	0.173422	0.29252	0.190375	0.094091
ALA	-0.114085	-0.093288	0.1558	-0.037351	-0.121446	0.084037	0.023819	0.093442	0.143256	0.044627	-0.105535
LLL	-0.075594	-0.100834	-0.046616	-0.20898	-0.008596	-0.038612	-0.04936	0.06072	-0.062662	-0.155879	-0.095449
LAA	-0.137546	-0.135425	0.121566	-0.038295	-0.212129	0.040009	0.078545	0.029837	0.138343	0.049377	0.025048
AAL	-0.156112	-0.133524	0.114426	-0.020264	-0.058513	0.057005	0.076881	0.054781	0.129436	0.019448	0.043217
ALL	-0.056191	-0.144594	0.043214	-0.146754	-0.058094	0.024076	0.074966	-0.028923	0.056939	0.104797	-0.111064
LLA	-0.17789	-0.001898	0.032638	-0.053407	-0.036736	-0.021239	-0.013052	-0.026865	0.029256	0.061432	-0.122784
LAL	-0.188611	-0.002185	0.108836	-0.126098	-0.001931	-0.017215	-0.056647	0.043682	0.039895	-0.084752	-0.080592
SSS	0.012405	-0.368833	-0.368951	-0.212781	-0.227907	-0.243726	0.056344	-0.137908	0.083584	-0.105236	-0.143811
EAL	-0.143734	-0.07476	-0.033182	0.061196	-0.107497	0.11289	0.055947	0.131137	0.055672	0.135703	0.010345
AAG	-0.122047	-0.131051	0.082267	0.007056	-0.133866	-0.047755	-0.057543	0.106205	0.158156	0.066523	0.112016
LGL	-0.040655	-0.029622	0.017907	-0.132046	-0.045142	-0.052407	0.039318	0.144508	-0.027922	-0.110411	-0.071683
LLS	-0.010711	-0.156172	-0.041087	-0.072868	0.022658	0.042833	0.019424	-0.093487	-0.003759	0.092095	0.074979
ELL	-0.093557	0.014243	-0.014263	-0.075895	0.02139	0.066242	0.030806	0.030263	0.026134	0.059581	-0.079635
LLE	-0.162262	-0.111966	0.040836	0.094973	-0.067124	0.059648	0.099326	-0.007085	-0.113252	0.077397	-0.076037
LLG	-0.12907	0.113013	0.114456	-0.17476	0.002945	0.02806	-0.108037	0.042189	-0.135708	-0.045802	-0.132305
SLL	0.005315	0.016392	-0.119716	-0.285225	-0.036248	0.122715	0.173636	-0.0147	-0.06203	-0.043801	-0.045284
VAA	-0.027963	-0.149745	0.025364	0.019901	-0.215134	-0.071495	0.016492	-0.085392	0.219921	-0.004116	0.01965
LAE	-0.083166	-0.067017	0.069444	0.002721	-0.124963	0.123592	0.035736	0.067787	0.066908	0.123455	-0.035818
GGG	-0.123306	-0.093683	-0.108203	-0.231365	-0.523704	-0.403804	0.195372	0.061233	-0.113784	-0.081564	0.088243
AGL	-0.207143	0.034644	0.121295	-0.163401	0.087782	-0.036605	-0.01754	0.088209	0.131121	0.061844	0.080362
EEL	-0.051608	-0.087294	0.036072	0.000638	-0.006934	-0.015743	0.032682	0.084566	0.002506	0.107209	0.002186
AVA	-0.076361	-0.062268	0.144362	0.14062	-0.120104	0.057678	0.04253	-0.070264	0.161969	0.080845	0.037832
EEE	-0.099348	-0.035743	0.048511	-0.011389	-0.166738	-0.203926	0.298709	0.013724	-0.076794	0.02071	-0.06938
LSL	-0.072752	-0.069518	-0.036016	-0.106899	-0.145442	-0.053543	0.055813	0.017788	-0.041808	-0.067715	-0.145315
VLA	-0.159771	0.186847	0.13649	-0.009145	-0.118425	0.008066	0.092235	-0.080168	0.011427	0.034528	-0.102418
ALG	-0.110935	-0.020543	0.044947	-0.013864	0.030869	0.02336	-0.035395	0.055567	0.088909	0.064803	0.01046
AAV	-0.132446	-0.147385	0.087288	0.032358	-0.09355	0.010073	0.015097	0.044157	0.209611	0.082991	0.03442
LAG	-0.171906	0.005839	0.09819	-0.043466	0.027992	0.048256	0.113637	0.059491	-0.015636	0.04267	-0.01567
VLL	-0.162151	0.005962	-0.072247	-0.07607	-0.015651	0.0627	0.099056	-0.043566	0.059558	-0.0715	-0.074845
GLL	-0.173185	0.012069	0.043597	-0.18772	-0.005353	-0.051068	-0.022771	-0.042247	-0.018772	0.018083	-0.112315
EAA	-0.078988	-0.184197	0.127382	0.031843	-0.239929	0.023499	0.129351	0.047405	0.153995	0.173007	0.002077
AGA	-0.18963	-0.203938	0.161371	0.005892	-0.202696	-0.147961	0.020732	-0.103878	0.093181	0.049163	0.005256
LLD	-0.076499	-0.017439	0.070412	-0.087081	-0.033162	-0.040791	0.193116	0.011308	-0.02188	0.192121	-0.114374
ELA	-0.111599	-0.066261	0.045726	-0.048047	-0.025858	-0.00024	-0.00333	0.052254	0.081465	0.122557	-0.04747
LEE	-0.060691	-0.081675	-0.015018	0.029915	-0.078151	-0.007045	0.1492	0.091358	-0.05752	0.046129	0.030288
ALE	-0.134804	-0.078882	-0.079002	0.125755	-0.100296	0.073504	0.098199	0.014848	0.014328	0.150866	-0.07598
AVL	-0.124822	-0.091698	0.014471	-0.079913	-0.100369	0.003308	0.121022	0.028039	0.058126	0.107839	-0.092217

Vilniaus universitetas

Istorinių duomenų sudarymas





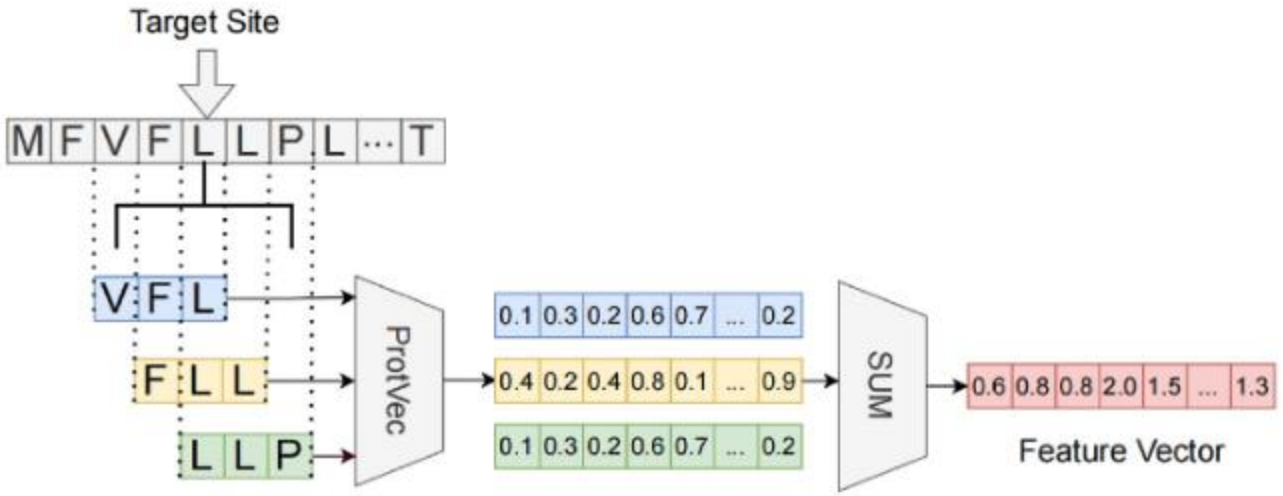
Vilniaus universitetas

Duomenų įvestis modeliui

- Duomenų aibė sudaroma taip, kaip straipsniuose.
- Tiriamas spyglio genas.
- Duomenų aibės apimtis: 5 (skaičius tiriamų įstorinių sekų) x 1269 (spyglio geno ilgis - kraštinės reikšmės) x 2426 (sudarytų sekų skaičius) x 100 (ProtVec įterpinio dydis). ~12gb
- Vieną duomenų įvestį sudaro 3 persidengiančių trigramų suma. Laikome, kad centrinis elementas (kuris yra visose trijose trigramose) yra stebimoje pozicijoje (žr. Iliustraciją kitoje skaidrėje).
- Duomenų išvestis: 2 dimensijos, taip arba ne (ar stebimoje pozicijoje atsiranda mutacija)

Vilniaus universitetas

Duomenų įvestis modeliui



Zhou, B., Zhou, H., Zhang, X., Xu, X., Chai, Y., Zheng, Z., Kot, A. C., & Zhou, Z. (2023). TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution. In Computers in Biology and Medicine (Vol. 152, p. 106264). Elsevier BV. https://doi.org/10.1016/j.compbiomed.2022.106264

Vilniaus universitetas

Duomenų aibės problema

Mūsų

```
_, counts = np.unique(Y_train, return_counts=True)
train_counts = max(counts)
train_imbalance = max(counts) / Y_train.shape[0]
_, counts = np.unique(Y_test, return_counts=True)
test_counts = max(counts)
test_imbalance = max(counts) / Y_test.shape[0]

print('Class imbalances:')
print(' Training %.3f' % train_imbalance)
print(' Testing  %.3f' % test_imbalance)
```

```
Class imbalances:
 Training 0.999
 Testing  1.000
```

Autorių

Covid19

```
Class imbalances:
 Training 0.502
 Testing  0.510
```

Gripas

```
Class imbalances:
 Training 0.853
 Testing  0.856
```

Sprendimo būdai

	Oversample	Undersample	Weighted
Iteracijos nr.	437	340	30
Accuracy	0.711	0.682	0.621
Precision	0.676	0.753	0.929
Recall	0.810	0.540	0.262
F_1 score	0.737	0.629	0.408
MCC	0.431	0.380	0.348

Kito pusmečio darbo planas

4

Vilniaus universitetas

Mokslinė veikla:

- Viruso mutacijų analizėje ir kompiuterinėje filogenetikoje naudojamų metodų teorinis tyrimas.
- Giliais neuroniniais tinklais grįstų mašininio mokymo metodų taikymo, viruso sekų mutacijoms prognozuoti, teorinis tyrimas.

Vilniaus
universitetas

Publikacijų rengimas

Mokslinių tyrimų disertacijos tema apžvalga (konferencijos darbų medžiagoje)

Vilniaus
universitetas

Dalyvavimas konferencijose

Dalyvavimas tarptautinėje konferencijoje Lietuvoje. ECML

Vilniaus
universitetas

Dalyvavimas vasaros mokykloje

Dalyvavimas KTU 8-oje tarptautinėje doktorantų vasaros mokykloje.