Фаза 1: Детален опис
01.05.2021

ВОВЕД ВО БИОИНФОРМАТИКА

Детален опис на проблемот и поврзана литература

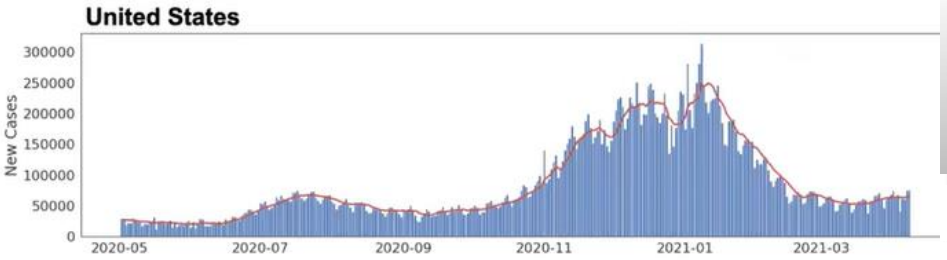
Предвидување на варијатни на глико-протеинот на SARS-Cov2 вирусот

Професор*:* д-р Невена Ацковска
Студент*:* Кирил Зеленковски, 161141

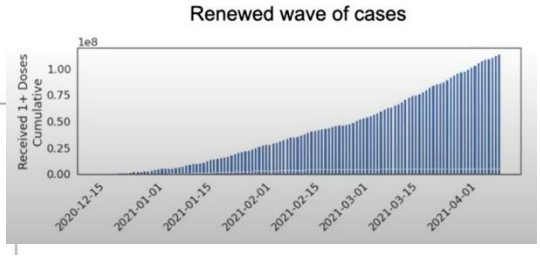
Генерален опис на проблемот

SARS-Cov2 вирусот нема да си оди во скора иднина. Во последната година излегоа голем број на предвивања [1] кои се обидуваат да го предвидат животниот век на овој вирус, кој само за 1 година нанесе масивна штета на нашата планета. Иако во своите истражувања истражувачи користат различни техники за предвидување, повеќето се согласуваат дека се далеку од модел кој би го опишал проблемот на епидемологија, т.е. ширењето на вирусот. Проблемот со ова не се лоши податоци, туку многу фактори што модел за една земја го прави да не одговара на друга земја. Моделите се само корисни во разбирање на однесувањето на вирусот, врз основа на нив треба да се аргументираат мерките што се превземаат (George Box: “All models are wrong, but some are useful.” [2]).

Следните податоци се земени од COVID CG иницијативата [3] каде на една кратка презентацијата ја покажуваат моментална состојба, Април 2021 (фигура 1) и сценарио каде се испитува бран од случаеви дури и кога 70 милиони се комплетно вакцинарни (со примени 2 дози) и 120 милиони со само примена 1 доза (фиугра 2).

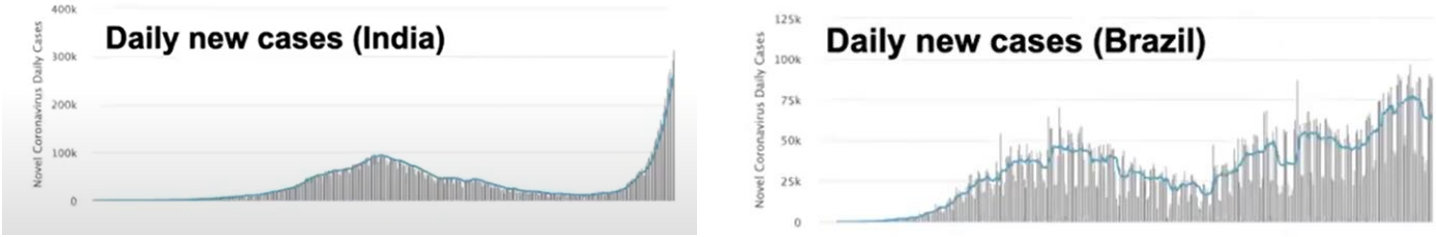


Фигура 1: Дистрибуција на нови случаеви за SARS-Cov2 до Април 2021



Фигура 2: Преглед на нов бран

Фигура 1 и 2 покажуваат дека и со целото вакцинирање не изгледа како да успорува вирусот, просекот на последните 2 недели е некаде 50 илјади случаеви само во Америка. Иако поголемиот дел од земјите во светот се надеваат на **имунитет на стадо** откако ќе започне масивно вакцинирање, тоа сценарио изгледа доста далечно - дали поради не-дејствување на вакцините (vaccine hesitancy), дали поради лоши епидемолошки модели и несоодветни чекори на управите на државите или сл.



Фигура 3: Дистрибуција на нови случаеви за SARS-Cov2 за Индија(лево), Бразил (десно)

Детален опис на проблемот

Но, SARS-Cov2 е **глобален** проблем! Вирус не препознава меѓудржавни граници и не разликува кого заразува. Иако земји како Италија, Америка ги поминаа своите поголеми бранови (Декември 2021), голем број на земји како Инидија [4], Бразил [5] моментално се во големи бранови од нови случаеви и допрва го доживуваат својот подем. (фигура 3).

Меѓутоа, зошто светот го интересира како и колку вирусот се шири во други земји?

Одговот на ова прашање е дека **вирусот мутира** (и тоа мутира премногу!). Бидејќи се шири меѓу луѓе со ралзични фенотипи на ДНА, тој колку повеќе луѓе заразува толку со поголема веројатност мутира. Овие мутации се од големо значење бидејќи доколку вирусот би го препознавале, би можеле да развиваме лекови и вакцини кои подобро ќе го таргетираат (со синтеза на антитела) и препознаваат вирусот, сличен пристап како тој во развивање на лекови против HIV [6].

Иако проблемот е брзината на мутирање, решението за предвидување на секоја следна варијанта или група под-варијатни на вирус е зависно од биолошки разлики во самиот геном. Нас овие биолошлки разлики во геномот ни се клучот, бидејќи иако се смета дека настануваат случајно, тие можат да го сменат однесувањето на вирусот (**silent** mutation, **missense** mutation, **nonsense** mutation).

Дефинирање на терминологијата:

- **Мутација:** е промена во самиот кодирачки регион која може да биде:
    - **point mutation** (*base substitutions*): промени од мал размер кои настануваат на РНА секвенцата за 1 до 2 нуклеотида еден ваков вид мутација е нуклеотидна замена
    - **frame shift** (*deletions* или *insertions*): се однесуваат на додавање (insertion) или бришење (deletion) на нуклеотиди во РНА секвенцата.

- **Варијанта (Variant/Strain):** се однесува на еволутивно слични геноми со заеднички мутации (збир од повеќе мутации)

Геномот на SARS-Cov2 вирусот е долг 30 kb (РНА вирус) и слично како и сите живи организми и вирусот е подлежен на природна селекција и тој се менува. Иако овој вирус има повеќе кодни региони [7] и секој си има своја клучна улога кога станува збора за репродукција на вирусот, јас се одлучив да ги разгледувам сите варијатни кои настануваат во генот S на вирусот, познат како глико-протеинот (**spike protein**).

Глико-протеинот сметам дека доколку се разбере може да му даде значење на ова истражување и информации зошто се достапни во следниот интерсен труд [8].

Податоци, опис на потенцијално податочно множество

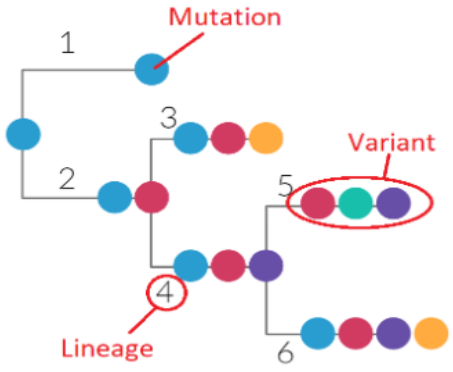
Предлог идеја премногу зависеше од множеството за учење и тренирање бидејќи не бев запознаен со форматот во кои доаѓаат податоците. По повеќе сурфње по интернет гледав од кој извор се податоците во повеќето dashboards и трудови што кружат околку интернетот, дали од престижни универзитети или фирми - повеќето влечат податоци од GISAID (не ги цитирав бидејќи навистина има премногу). GISAID [9] е овозможено од грантови од Германија, Америка и Сингапур.

Пред поднесување на оваа фаза аплицирав на страната за да можам да пристапам до самите варијанти, групи од мутации. По неколку дена добив пристап и страната е стварно фантастична има повеќе корисни алатки и скоро секој ден овие податоци се обновуваат (овој факт на нови updates, јас лично го гледам како мотивација за работа бидејќи навистина е морбидно колку брзо мутира, но for the sake of my 5 годишни студии ќе морам да почнам со фиксно множество за да не залутам).

Бидејќи ги разгледуваме тие варијанти, на самата страна има одел за **“emerging variants”**, каде множеството има: колона за име на варијанта, колона за колку геноми од истата се секвенционирани и колона за колку мутации има во таа варијанта (и кои се). Самата страна изведува како мала формула каде ги множи геномите * бр. на мутации за да добие како score за таа мутација, и тоа е опционална колона (може да биде корисна).

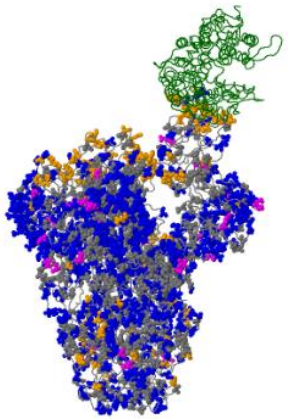
Ова е доста raw множество и почнав да го средувам (малку одзема време бидејќи има премногу записи).

Затоа иницијално ќе почнам да работам само на еден **lineage**, така ќе можам да видам и кое решение е подобар пристап. **Lineage** доаѓа како лоза (потекло) на секоја нова варијанта. Фигура 4 е графички приказ за разликата помеѓу сите овие термини. Јас ќе предвидувам група од мутации (варијанта) и ќе ја потоа споредувам со веќе постоечките варијанти кои се водат како **variants of concern** [10]**.**



Фигура 4. Мутација, варијанта, лоза [11]

Ова е доста интересно бидејќи откако ќе се подели множеството на тренирачко и тестирачко, со метрики ќе гледам колку/дали се совпаѓаат предвидените варијатни со веќе постоечките. Подоцна откако ќе научи добро идејата е да предвидува непостоечки.



Фигура 5: 3D модел од B.1.1.7

Пример за лоза која е доста важна е **Британската варијанта**, **B.1.1.7.** Се смета за доста битна (има преголем score) бидејќи се има проширено во преку од 50 земји, константно мутира и според следниот труд [12] сметат дека има 30% поголема стапка на смрттност (но имаше и други статии што го побиваат ова [13]).

(Предлог) Решение со користење на Трансформери

На влез се читаат РНА секвенци (од тој регион: 21563-25384) , сите се варијанти со соодветни мутации и на излез се добива:

- Една идна мутација (прв чекор)
- Повеќе идни мутации (втор чекор)

Иако проблемот може да се реши со класификација, сепак прво сакам да се обидам со техника од длабоко учење познати како **трансформери** [14]. Оваа sequence-to-sequence архитектура (Seq2Seq) е предложена во светот на *обработка на природни јазици* (NLP) и во позадина е невронска мрежа што трансформира дадена секвенца од елементи, секвенца од зборови во реченица, и ја трансформира во нова секвенца – реченица.

Сега, иако овие модели претежно ги користат за превод од еден јазик во друг самата структура на моделот ме доведе до следната идеја.

Seq2Seq модели се составени од Енкодер и Декодер. Енкодерот чита влезна секвенца и ја мапира во повеќедимензионален вектор. Овој апстрактен вектор потоа го храни Декодерот што го трансформира во излезна секвенца. Овој излез може да биде во друг јазик, симбол, копија на самиот влез и сл.

Јас кога го замислувам моделот ме потсетува на следното сценарио: Енкодер и Декодер зборуваат два јазика, едниот е нивниот мајчин различен за двајцата (пример Германски и Македонски) додека нивниот втор е јазик имагинарен и заеднички за двајцата. За да се

преведе нештото од Германски во Македонски, Енкодерот мора да конвертира од Германска реченица во имагинарен и потоа Декодерот го чита тој имагинарен и го преведува во Македонски.

Е сега зошто јазици и целиот овој свет е корисен за предвидување на варијанта? Мојата замисла за одговорот е повторно прашање: **Како знаеме кое делче од секвенцата е подлежно да се смени во следната варијанта?**

И тука ова се решава со самиот начин како архитектурата учи. Доколку замислиме сценарио каде ни Енкодерот ни Декодерот немаат заеднички имагинарен јазик, за да го научат они мора да тренираат (моделот се тренира) со многу примери и благодарение на GISAID, нив ги имаме на одлив. Тука само е во прашање претпроцесирањето што треба да се добие множество кое би имало влезна и излезна секвенца, кои би се ределе врз основа на time-stamp (влезна пред излезна) и би биле за почеток САМО од една лоза. Одам со една лоза бидејќи полесно е да scale-аш ако е помало, и не сум сигурен колку е оптимално за моделот добро да се тренира со премногу примероци, поради тоа овој број може да варира.

Едноставен избор за Енкдодер и Декодер од ваква архитектура се единечни LSTM за секој од двата. За полесно да ги разбираме трансформерите има еден технички детал што е доста важен за да ја компелемнтира идејата а тоа е деталот за **внимание**.

Механизмот за **внимание** на самиот модел гледа на влезна секвенца и одлучува во секој чекор од учењето кои други делови од секвенцата се важни. Ова малку звучи апстрактно, но прост пример помага:

Кога ја читате реченицава, секогаш се фокусирате на зборот што го читате но во исто време вашиот ум ја држи целата содржина од најбитните клучни зборчиња во меморија за да може да даде значење и да не бидат случајни.

Овој механизам работи слично на секвенцата. Пример доколку зборуваме за преводот од погоре, тогаш овој механизам би бил Енкодерот кога преведува запишува клучни зборчиња поврзани со преводот што се важни за семантиката, и ги пренесува на Декодерот како плус пред да почне да преведува. Тие нови зборчиња му ја олеснуваат работата на Декодерот бидејќи знае точно кои делчиња од реченицата се битни за контекстот да не се изгуби.

Со тоа во процесот на учење, моделите стават **маски** (ознаки за контекстот) кои се стават на самата влезна секвенца во првата како фаза (multi-head) и ова всушност се прави за се да избегне потенцијални “идни” грешки со елементите од секвенцата (повеќе за ова во другите фази).

Овие маски би биле самите мутации за влезната варијанта, што ни се колона (feature) во множеството од влезни секвенци и ги поставуваме со цел декодерот да учи нагласено на тие места за идната секвенца.

Имам познавање минимално за длабоко учење, но како идеја личи доста револуционерно бидејќи труд кој користи трансформери за предвидување на идни варијанти на SARS-Cov2 сеуште нема. Можеби ќе ми треба повеќе време да процесирам дали идејата е добра, има примери каде се користи во генетика, но повеќе е кон транскрпција [15].

Продолжување на проблематика, алтернативни модели

1. Класификација на мутации

SARS-Cov2 е **РНА** базиран вирус, што се многу различни од **ДНА** вируси во контекст на повисоки рати на мутирање со што имаат повисока стапка на адаптивност

Откако ќе се предвидат мутациите може да се направи и класификација врз база на каква мутација настанала. Има неколку видови на мутации кои настануваат на **РНА** секвенцата погоре што ги објаснива и ова може да се комбинира со проблемот за што се случува откако ќе се добијат, т.е. какво е нивното влијание врз основа на нивното значење:

- **silent mutation:** промена на кодон со што резултатната амино секвенца останува непроменета
- **missense mutation:** промена која влијае на резултатна амино секвенца
- **nonsense mutation:** промена на кодон која продуцира стоп кодон кој потоа во генската транслација предизвикува нефункционални читања

2. Rough set theory за предвидување на РНА мутации

Техника што предвидува потенцијални замени што може да настануваат во примарната **РНА** секвенца [16]. Во секоја итерација на учење:

- влез: РНА секвенца од генерација (мутација)
- излез: РНА секвенца од следна генерација (мутација)

Features на влезот се сите нуклеотиди во **РНА** секвенцата и истите кореспондираат со features на излезот. Моделот го храниме со пораменти **РНА** секвенци од последователни генерации (**GISAID**, outbreak.info) од истиот вирус (истата РНА) што се секвенционирани во слична околина (нас ни треба човек)
Крајната цел е да се предвиди следната РНА секвенца зимајќи ги во предвид сите претходни РНА секвенци.

Референци

[1] Analysis and prediction of COVID-19 trajectory: A machine learning approach, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7744840/

[2] George E. Box, https://www.lacan.upc.edu/admoreWeb/2018/05/all-models-are-wrong-but-some-are-useful-george-e-p-box/

[3] COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest, https://elifesciences.org/articles/63409

[4] India tops 200,000 dead amid coronavirus surge, https://www.dw.com/en/india-tops-200000-dead-amid-coronavirus-surge/a-57356631#:~:text=Coronavirus%20cases%20are%20surging%20in,3%2C000%20fatalities%20in%20one%20day.

[5] Covid: Brazil has more than 4,000 deaths in 24 hours for first time, https://www.bbc.com/news/world-latin-america-56657818

[6] Fifteen to Twenty Percent of HIV Substitution Mutations Are Associated with Recombination, https://jvi.asm.org/content/88/7/3837

[7] NCBI, SARS-CoV-2, complete genome, https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/

[8] The sugar code: Why glycans are so important, https://pubmed.ncbi.nlm.nih.gov/28709806/

[9] GISAID, https://www.gisaid.org/

[10] Outbreak.info, https://outbreak.info/situation-reports/caveats

[11] Phylogeny of SARS-CoV-2: https://www.nature.com/articles/s41598-021-82938-2

[12] Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study, https://www.bmj.com/content/372/bmj.n579

[13] "Nervtag", Peter Horby, Catherine Huntley, Nick Davies, John Edmunds, Neil Ferguson, Graham Medley, Calum Semple, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/961037/NERVTAG_note_on_B.1.1.7_severity_for_SAGE_77__1_.pdf

[14] Attention Is All You Need, https://arxiv.org/abs/1706.03762

[15] An Attention-Based Model for Transcription Factor Binding Site Prediction, https://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-83.pdf

[16] The prediction of virus mutation using neural networks and rough set techniques, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4867776/