Heidelberg University
Institute of Computer Science

Project proposal for the lecture Advanced Machine Learning

Project ideas in the area of Covid-19 research

`https://github.com/nilskre/AML-covid-project`

Team Member: Felix Hausberger, 3661293,
Applied Computer Science
eb260@stud.uni-heidelberg.de

Team Member: Nils Krehl, 3664130,
Applied Computer Science
pu268@stud.uni-heidelberg.de

Contents

1 Introduction 2

2 Proposal 1: Machine Learning based prediction of the next SARS-CoV-2 variants and simulation of vaccine effectiveness 3

3 Conclusion 5

List of Abbreviations

DNA	Deoxyribonucleic Acid
GAN	Generative Adversarial Networks
NLP	Natural Language Processing

1 Introduction

tbd

2 Proposal 1: Machine Learning based prediction of the next SARS-CoV-2 variants and simulation of vaccine effectiveness

1. Idea and research question

There is hope for an end of the pandemic due to the vaccines. All vaccines are developed against the wild type of the virus. Through the mutation processes the danger, that a variant occurs, against which the vaccines are no longer effective, arises. This process is known as viral escape. This proposal aims to detect this development faster to enable a faster and better response to new variants.

So there are two scientific questions (which can also be separated):

1. Can the next possible SARS-CoV-2 variants be predicted by a Machine Learning model?
2. Based on the predicted possible next variants can the effectiveness of the vaccines be simulated?

2. Related work

Already some works exist in the area of predicting virus mutations using Machine Learning techniques. Hie et al. [3] applied methods developed for Natural Language Processing (NLP). According to their work escape mutations look different to the immune system, but have the same viral infectivity. The analogy from the NLP area are word changes, which change the meaning of a sentence, but the grammaticality remains. Through their work, they have managed to generate a connection between natural language and viral evolution.

Even before the SARS-CoV-2 pandemic, research was done in this area, e.g. from Salama et al. [6]. They used neural networks for predicting new mutations and rough set techniques for detecting patterns in mutations. Furthermore they validated their approach for the Newcastle virus and achieved an accuracy of 75%.

Generative Adversarial Networks (GAN) achieve great results in image generation. Berman et al. applied a GAN in [2] to generate Deoxyribonucleic Acid (DNA) sequences.

3. Approach

After preprocessing the raw SARS-CoV-2 genomes we would like to train a Machine Learning model for predicting the probable next mutations. The Machine Learning model could be a standard neural network (as in [6]), a

GAN (as in [2]) or a neural network language model (as in [3]). Due to a lack of time we haven't decided yet for one Machine Learning model.

4. Data sources

For predicting the next possible SARS-CoV-2 variants the SARS-CoV-2 genome development from the past can be used as data source (genomic time series data). Data sources are available through the GISAID initiative [1]. In April 2021 over one million SARS-CoV-2 genome are available via GISAID [4]. Based on this data the Nextstrain project provides a visualization how the SARS-CoV-2 genome evolves: Nextstrain

5. Computational resources

The needed computational resources and time depends on the concrete chosen Machine Learning approach. As described above due to a lack of time we haven't decided yet for one Machine Learning approach. For further research and the decision for one Machine Learning approach we keep this in mind.

6. Probable difficulties

Challenges could arise due to the amount of data (one SARS-CoV-2 genome consists of about 30.000 bases => TS size = data instances * 30.000). This can be encountered with less data instances or a compression of SARS-CoV-2 genome e.g. through DNA2vec [5] (represents pieces of DNA as vector).

3 Conclusion

tbd

References

[1] GISAID (editor). *GISAID - Mission.* GISAID mission. URL: https://www.gisaid.org/about-us/mission/ (visited on 07/05/2021).

[2] Daniel S. Berman et al. *MutaGAN: A Seq2seq GAN Framework to Predict Mutations of Evolving Protein Populations.* 2020. arXiv: 2008.11790 [q-bio.QM].

[3] Brian Hie et al. "Learning the Language of Viral Evolution and Escape". In: *Science* 371.6526 (2021), pp. 284–288. ISSN: 0036-8075. DOI: 10.1126/science.abd7331. eprint: https://science.sciencemag.org/content/371/6526/284.full.pdf. URL: https://science.sciencemag.org/content/371/6526/284.

[4] Amy Maxmen. "One Million Coronavirus Sequences: Popular Genome Site Hits Mega Milestone". In: *Nature* 593.7857 (7857 Apr. 23, 2021), pp. 21–21. DOI: 10.1038/d41586-021-01069-w. URL: https://www.nature.com/articles/d41586-021-01069-w (visited on 07/05/2021).

[5] Patrick Ng. *Dna2vec: Consistent Vector Representations of Variable-Length k-Mers.* 2017. arXiv: 1701.06279 [q-bio.QM].

[6] Mostafa Salama, Aboul Ella Hassanien, and Ahmad Mostafa. "The Prediction of Virus Mutation Using Neural Networks and Rough Set Techniques". In: *EURASIP Journal on Bioinformatics and Systems Biology* 2016 (Dec. 2016). DOI: 10.1186/s13637-016-0042-0.