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

List of Abbreviations

tbd tbd

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. [2] applied methods developed for NLP (Natural Language Processing). 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. [4]. TODO: approach

GANs (Generative Adversarial Networks) achieve great results in image generation. Berman et al. applied a GAN in [1] to generate DNA sequences.

DNA2vec [3]

3. Approach

DNA2vec und dann in NN (GAN)? vaccines effectivness unsupervised wie in paper learning the language?

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). There are numerous

The Nextstrain project has visualized TODO

TODO: data sources GSAID?

In terms of data quality, when working with genomic data one should always have in mind, that genotyping errors are possible.

5. Computational resources

6. Probable difficulties

TODO: ca. 30000 Basen besitzt das coronavirus -> sehr groß

3 Conclusion

tbd

References

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

[2] 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.

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

[4] 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.