Heidelberg University
Institute of Computer Science

Project proposals for 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

List of Abbreviations

AE	Auto Encoder
DNA	Deoxyribonucleic Acid
GAN	Generative Adversarial Networks
GCN	Graph Convolutional Network
GRU	Gated Recurrent Unit
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network

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 (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 (ML) 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 ML 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. Even before the SARS-CoV-2 pandemic, research was done in this area, e.g. from Salama et al. [9]. 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 ML model for predicting the probable next mutations. The ML model could be a standard neural network (as in [9]), 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 ML 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 genomes are available via GISAID [6]. Based on this data the Nextstrain[1] project provides a visualization how the SARS-CoV-2 genome evolves.

5. Computational resources

The needed computational resources and time depends on the concrete chosen ML approach. As described above due to a lack of time we haven't decided yet for one ML approach. For further research and the decision for one ML 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 [7] (represents pieces of DNA as vector).

[1] https://nextstrain.org/ncov/global?label=clade:19A

2 OpenVaccine - COVID-19 mRNA Vaccine Degradation Prediction

mRNA vaccines have the severe problem to degrade quickly as being unstable without intense refrigeration. This makes the preparation and shippment of such vaccines difficult and leads to high losses. Only little is known about which parts of the mRNA backbones are most likely to degrade. Thus we want to predict the degradation rate at each base of a possibly stable RNA molecule.

The Stanford University hosted a Kaggle challenge exactly for this problem [5] and provided parts of a high-quality EteRNA[2] data set consisting of over 3000 RNA molecules and their degradation rates at each base.

With over 1.600 teams having participated in this challenge many promising approaches were evaluated. For sequence modeling especially Recurrent Neural Network (RNN) approaches are most promising. [4] made use of a regularized LSTM model to predict the degradation rates of each base outperforming traditional tree-based algorithms. Besides Long Short Term Memory (LSTM)s [10] also made use of Gated Recurrent Unit (GRU)s and Graph Convolutional Network (GCN)s to solve the challenge, whereas the GRU approach performed best with an accuracy of 76%. Similar approaches are given in [8] and [11]. We want to reconstruct an own simplified solution for this challenge inspired by such RNN approaches and might add other components like an Auto Encoder (AE) for better denoised feature extraction. By dealing with such high-performance models and building an own simplified model one can better evaluate the pros and cons of the current best solutions for this problem.

The LSTM of [4] stated training durations of around 300 epochs which took about 84 minutes on an NVIDIA TESLA P100 GPU. Having a NVIDIA GTX 1080 available locally helps to achieve similar training times, also with the free NVIDIA K80 GPU on Google Colab one should not need to wait ages until training has finished.

The difficulty of this project is therefore to learn and deal with the mostly unknown RNN architectures and to achieve fair results given a simplified model as an outcome of this project. Also the comparison of different approaches and why one works best should not be neglected.

[2] crowdsourcing platform for RNA design: https://eternagame.org/

References

[1] GISAID (editor). *GISAID - Mission.* GISAID mission. 2021. 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". In: *arXiv* (2020). URL: https://arxiv.org/abs/2008.11790 (visited on 07/05/2021).

[3] Brian Hie et al. "Learning the language of viral evolution and escape". In: *Science* 371.6526 (2021). Publisher: American Association for the Advancement of Science, pp. 284–288. ISSN: 0036-8075. DOI: 10.1126/science.abd7331. URL: https://science.sciencemag.org/content/371/6526/284 (visited on 07/05/2021).

[4] Sheikh Asif Imran et al. "COVID-19 mRNA Vaccine Degradation Prediction using Regularized LSTM Model". In: *IEEE* (2021). DOI: 10.1109/WIECON-ECE52138.2020.9398044. URL: https://ieeexplore.ieee.org/document/9398044 (visited on 07/06/2021).

[5] Kaggle. *OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction.* OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction. 2020. URL: https://www.kaggle.com/c/stanford-covid-vaccine/discussion/182652 (visited on 07/06/2021).

[6] Amy Maxmen. "One million coronavirus sequences: popular genome site hits mega milestone". In: *Nature* 593.7857 (2021). Bandiera_abtest: a Cg_type: News Number: 7857 Publisher: Nature Publishing Group Subject_term: SARS-CoV-2, Databases, Epidemiology, 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).

[7] Patrick Ng. "dna2vec: Consistent vector representations of variable-length k-mers". In: *arXiv* (2017). URL: https://arxiv.org/abs/1701.06279 (visited on 07/05/2021).

[8] Talal S. Qaid et al. "Deep sequence modelling for predicting COVID-19 mRNA vaccine degradation". In: *PeerJ Computer Science* (2021). URL: https://doi.org/10.7717/peerj-cs.597 (visited on 07/06/2021).

[9] 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 (2016). DOI: 10.1186/s13637-016-0042-0. (Visited on 07/05/2021).

[10] Ankit Singhal. "Application and Comparison of Deep Learning Methods in the Prediction of RNA Sequence Degradation and Stability". In: *arXiv* (2021). URL: https://arxiv.org/abs/2011.05136 (visited on 07/06/2021).

[11] Gilles Vandewiele. *Predicting mRNA Degradation using GNNs and RNNs in the Search for a COVID-19 Vaccine.* 2020. URL: https://towardsdatascience.com/predicting-mrna-degradation-using-gnns-and-rnns-in-the-search-for-a-covid-19-vaccine-b3070d20b2e5 (visited on 07/06/2021).