

Linz, March 21, 2022

Opponent Review for the Bachelor Thesis:

**“Auto-Encoding Amino Acid Sequences with LSTM”
by Markus Promberger**

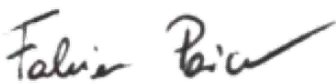
It was my pleasure to review the thesis which tackles the problem of protein sequence analysis with machine learning algorithms. Particularly, it deals with encoding of such sequences via recurrent neural networks and analysis of their latent space to find semantically similar sequences. First, the work looks into traditional Bioinformatics methods for clustering homologous amino acid sequences and highlights their drawbacks. Then, Machine Learning algorithms are introduced and the author elaborates on how they can be utilized to process protein sequences. Finally, it is shown that a recurrent autoencoder is capable of capturing semantics of analysed sequences, which allows classifying of unknown sequences to their most similar organisms. This provides a viable alternative to building phylogenetic trees.

In the introduction the approach is motivated and the data and used methods are explained. It should be more clearly motivated why machine learning techniques provide benefits in the realm of sequence analysis. For the sake of reproducibility, it should be mentioned whether the used dataset is publicly available or self-generated. The related work section is rather redundant, since it only lists a single other work that motivates the thesis.

The methods and results are explained in a clear manner. First, the traditional Bioinformatics pipeline for classifying protein sequences is explained and its drawbacks are highlighted. Then, the used Machine Learning algorithms are introduced and explained in a clear way. Experiments were conducted in a methodologically sound manner and the sensitivity of the algorithm to various hyperparameter choices are discussed. The shown results are promising considering the few amount of data used for training the algorithm. Anyways, it would have been interesting to investigate alternatives to the used one-hot encoding of amino acids.

Finally, the discussion comments on refinements regarding algorithmic implementation and highlights certain rather unusual results found by the hyperparameter search, which are worth a mention. Also, the future potential of the used techniques are highlighted, but could be elaborated on in more detail.

Overall I appreciate the scope of this work and I am pleased to recommend this Bachelor Thesis for acceptance at the University of South Bohemia.



Fabian Paischer, MSc