Master's Thesis - Project plan

Disease prediction from plasma proteomics
using Natural Language Processing

Enric Cristòbal Cóppulo, s202562

February 23, 2022

1 Description

One of the current approaches for analyzing the presence of some specific disease is by doing a protein identification from protein plasma mass spectrometry data. Currently, the mainstream bottom-up proteomics peptide identification is based on matching experimentally generated MS2-spectra to in-silico predicted MS2-spectra. The peptides are assembled back to proteins which are then aggregated to protein groups. And protein quantification is based on peptide levels. However, during these steps a lot of information is lost and around 50% of MS2-spectra cannot be assigned to a specific peptide. Therefore, the aim of this project is to investigate whether the use of Natural Language Processing models on mass spectrometry data allows the creation of meaningful embeddings allowing to jump the previously mentioned steps for the current approach, and instead go straight to the disease prediction from the embedding. This will not only allow to keep all the available information on the analysis, but also accelerate the prediction.

Details:

- ECTS Points: 30
- Start date: 24/01/2022
- Hand-in date: 24/06/2022
- Project developed in: Center for Protein Research (KU)
- Collaborative institutes: Department of Health Technology (DTU)
- Supervisors: Henrik Nielsen (henni@dtu.dk)
- External Supervisors: Simon Rasmussen (simon.rasmussen@cpr.ku.dk), Lili Niu (lili.niu@cpr.ku.dk)

2 Project plan

Date	Risk	Task
		January
24/01 - 30/01	1	Literature analysis for Transformers and NLP
		February
31/01 - 06/02	1	In-depth Transformer and BERT model architecture understanding
07/02 - 13/02	1	Proteomics mass spectrometry field and NLP coding
14/02 - 20/02	2	Analysis of data and data preparation
21/02 - 27/02	3	BERT vanilla model creation
		March
28/02 - 06/03	3	Embedding and model training
07/03 - 13/03	3	Model testing and first results analysis
14/03 - 20/03	2	Try new data structures for training and model fine-tuning
21/03 - 27/03	4	Decision making based on results
		April
28/03 - 03/04	3	Consider new architectures to use as baseline
04/04 - 10/04	1	Writing code
11/04 - 17/04	3	Train model on bigger database
18/04 - 24/04	3	Analyze and assess new classifier performance
25/04 - 01/05	2	Benchmarking against other procedures
		May
02/05 - 08/05	2	Optimize model
09/05 - 15/05	1	Clean code
16/05 - 22/05	3	Discuss model performance and potential final approach
23/05 - 29/05	2	Obtain first final results for the report
		June
30/05 - 05/06	2	Final code preparation for repository creation
06/06 - 12/06	1	Report writing
13/06 - 19/06	1	Report writing
20/06 - 24/06		Hand-in

2.1 Description

This is a quite open project and as such, and as specified on the project plan, there will be some important decisions to make along the way regarding model performance that might influence the tasks' time-line depending on the findings. However, the main frame could be divided in 4 sections:

2.1.1 January & February - NLP and mass spectrometry proteomics introduction

In this first part of the project I will gather as much information as possible through literature research as well as from my supervisors' resources. The aim here is to deeply understand the world surrounding NLP (specially Transformers and BERT model) for the model construction as well as the mass spectrometry to become familiar with my project dataset.

2.1.2 March - BERT model assessment

During the month of March I will be mainly focused on the development and assessment of BERT model, our first and main choice to analyze its behaviour in front of this new type of data. Here I will develop some fine-tuning to improve our results as well as try different ways to organize the data for the model input. As this is a Language model applied to "numerical" data, which we believe is very new in the field, no correct approach has been defined, and thus there is a lot of room for improvement by doing trial and error. Lastly, after this month and after been fully focused on BERT model, its performance will be assessed and a key decision for this project will be taken so we rather continue improving its performance or we consider to go deeper on new architectures.

2.1.3 April - Model optimization and benchmarking

Regardless of the assessment output, in April I will first consider new potential architectures that could be trained on our dataset to lately have some kind of benchmarking to compare our model performance. Besides, the model performance will be potentially improved by training on bigger databases to allow our model to generalize and learn a bigger "mass spectrometry language". This will definitely allow us to assess whether any of the models used is really learning some useful embedding.

2.1.4 May - Final steps and code preparation

In this part, a final decision on the main model to focus on should have been made and the main action would be to finalize the model optimization for the obtainment of the final results. Some further discussions could occur to consider applying new things based on the knowledge obtained up to this point. Furthermore, I will start cleaning the code and creating a good documentation of it to enable potential future users to easily understand it.

2.1.5 June - Report writing and hand-in

This last section will be basically wrapping-up all the work done, create the repository for my code as well as obtain the final results to include to the final report. Most of the time will be spent on report preparation before the hand-in.

2.2 Risk analysis

In the project plan, the risk of each task is ranked on a scale from 1 (little risk) to 4 (high risk) as a measure of the likeliness of the task taking longer than the designated time or the presence of critical decisions for deciding the project direction.

The main risks in this project is the novelty of this approach as although there is one study that succeeded on using this type of data for developing some "mass spectrometry language" understanding, no successful results are ensured.