FACULTY OF MEDICINE AND HEALTH SCIENCES

A DECOY-FREE APPROACH FOR IDENTIFYING PEPTIDES IN METAPROTEOMICS

Wout Cattrijsse
Student number: 01800516

Promotor: Prof. Dr. Lennart Martens
Co-promotor: Dr. Tim Van Den Bossche

Research Internship Biomedical Sciences

Academic year: 2022 – 2023

GHENT UNIVERSITY

ABSTRACT

Metaproteomics is a rapidly evolving field in which the functional activities of microbial communities in complex environments are studied. Mass spectrometry-based proteomics is the most used method of performing (meta)proteomics experiments. In such experiments, the number of false positives is estimated by calculating a false discovery rate using the target-decoy approach. However, in metaproteomics the large size of the databases causes a much decreased identification rate and very long computation time. Therefore, doubling the already large size of the target database seems contra-intuitive. In this thesis, a decoy-free machine learning model will be trained on publicly available data so the need for a decoy database can be removed. The goal of this model is therefore to be able to distinguish between correct and false peptide-to-spectrum matches, without the need for a decoy database. This will not only be faster but might also lead to a higher identification rate.

RATIONALE AND POSITIONING WITH REGARD TO THE STATE-OF-THE-ART

Metaproteomics is a rapidly evolving field that aims to understand the functional activities of microbial communities present in complex environments[1]. Such communities encompass, but are not limited to, the gut microbiome, the skin microbiome and environmental microbial communities such as those present in a soil. These functional activities can be identified by studying the full protein complement present in such a community, and described in, for example, metabolic pathways. Metaproteomics is therefore a powerful tool for understanding the complex interplay between microorganisms and their environment, and has important applications in fields such as environmental science, microbiology and human health[2].

Mass spectrometry-based metaproteomics has become an increasingly popular method in microbiome research. This approach involves extracting proteins from a sample, digesting them into smaller peptides and then analyzing these peptides using mass spectrometry. By comparing the resulting peptide spectra to theoretical spectra from *in silico* digested protein databases, researchers can identify specific peptides and infer the proteins that are present in the community[2]. However, the analysis of metaproteomic data is complicated by the high degree of complexity and variability of these communities, which can lead to challenges in accurate peptide and protein identification.

Mass spectrometry-based proteomics experiments often generate a large amount of data that need to be analyzed to identify peptides and infer proteins. When analyzing such data, a threshold needs to be stated for statistical significance (which means below this threshold, data points, or peptide-to-spectrum matches (PSMs), are considered significant or good hits; above they are considered to be bad hits). This threshold is defined as the False Discovery Rate (FDR) and is usually set at 1%, which is accepted as being the standard[3]. In proteomics, the FDR is estimated by controlling the proportion of false positives among all significant results, instead of controlling the probability of any single false positive.

Currently, the most popular way to determine the FDR is the Target-Decoy Approach (TDA). In this approach, a target protein sequence database (that contains all possible proteins in the sample plus contaminants) is concatenated with a decoy database (which contains reversed or shuffled versions of the target protein sequences). During peptide identification, the mass spectrometry data generated from the sample is searched against this concatenated database. Because the decoy database contains reversed or shuffled sequences, any hits to these sequences are considered

false positives. By analyzing the proportion of decoys over target sequences, researchers can estimate the false discovery rate (FDR) of the protein identification results[4].

The protein database selection has a big influence on the identification rate in proteomics. Ideally, a protein database contains all (and only) proteins that are expected to be present in the sample. In the case that the database does not accurately represent the contents of the sample, the spectrum might not be identified because the protein is not present in the database, or worse, a false ("forced") match might occur. This causes the overall identification rate to go down, and/or an increase in false positives. Especially in metaproteomics this problem becomes harder because we do not know the exact composition of the sample due to the inherent high complexity of metaproteomic samples. This causes indeed the low identification rates typically seen in metaproteomic experiments[5,6].

A first approach to increase the identification rate in metaproteomics is to perform a two-step search. Here, as the name suggests, the searching is done in two steps. In the first step, the initial database is searched without any FDR limitation and all the proteins that are identified in this step are brought together into a new database. Then a second search is performed with a stringent FDR threshold against this new database. But this two-step approach holds some pitfalls, as shown in the study performed by Muth *et al.* (2015). This method effectively decreases the database size and results in an increased identification rate, but comes with the cost of an increased number of false positives[6]. Currently, new multi-step search methods do exist such as the search-all-asses-subset method from Sticker *et al.* (2017)[7]. In this method first all experimental MS2 spectra are searched against all peptides that are potentially present in the sample. Then all PSMs that emanate from proteins you are not interested in are removed and only then in a third step the FDR is calculated[7].

A second approach to increase the identification rate in metaproteomics is to perform post-processing steps with machine learning systems such as Percolator or MS^{2}Rescore[8,9]. Percolator is a semi-supervised machine learning algorithm that uses the PSMs derived from the decoy database as negative examples and the high-scoring PSMs derived from the target database as positive examples. On the basis of this distinction between negative and positive examples, a support vector machine is trained to be able to discriminate between positive and negative PSMs. MS^{2}Rescore uses Percolator to perform its post-processing, but it will also add features from MS^{2}PIP and DeepLC to distinguish even better between target and decoy hits[8,9].

In metaproteomics, where the protein sequences come from multiple organisms, the TDA approach may not be suitable. The TDA approach starts from the Equal Chance Assumption which states that target mismatches and decoy matches are equally likely to happen[10]. However, due to the large databases used in metaproteomics the target and decoy sequences will become too similar causing their score distributions to overlap too much (as can be seen in Figure 1) with a low identification rate as a result[6]. In this thesis, we will tackle this by applying a decoy-free approach based on the binary classifier Nokoi. Nokoi is a semi-supervised machine learning system that uses the low-ranked PSMs as substitutes for decoy PSMs. From these low-ranked PSMs several features were extracted and these were used to compute a feature vector that serves as the negative example and which describes the PSMs at the spectrum and at the peptide level. For the positive example a feature vector was calculated in the same manner but using the highly confident rank 1 PSMs[11]. It was already demonstrated that Nokoi shows good performance and broad applicability, and could therefore possibly also be applied to metaproteomics[11].

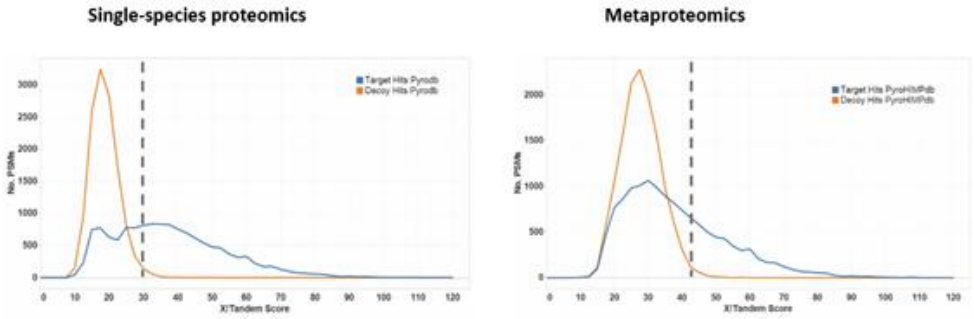

Figure 1 adapted from Muth et al. (2015), Comparison of score distributions of target and decoy hits in single-species and metaproteomics

SCIENTIFIC RESEARCH OBJECTIVES

The overall objective of this project is to train and validate a machine learning model that can serve as an alternative to the target-decoy approach in mass spectrometry-based (meta)proteomics.
To achieve this overall objective, I first will (re-)investigate the classical target-decoy approach and the multi-step approach on benchmark datasets. If necessary, I will have to construct these datasets myself. These results can later serve as comparison against the decoy-free method. Secondly, I will develop the decoy-free model, based on the Nokoi algorithm. The goal of this part is to train a semi-supervised machine learning model and tune the hyperparameters so that the model achieves an optimal performance that can be nicely extrapolated to new, validation data.

RESEARCH METHODOLOGY AND WORK PLAN

1. Data collection

In this thesis, mass spectrometry-based (meta)proteomics data will be collected from multiple sources. The training and test data for the model will be retrieved from an in-house reprocessing effort of human, mouse and *E. coli* data from the PRIDE database (https://www.ebi.ac.uk/pride/) with ionbot[12].
To validate the results, I will test the model on datasets from the CAMPI benchmark study and an inflammatory bowel disease (IBD) dataset. First, I will use the datasets from the CAMPI study, which is a community-driven, multi-laboratory paper that provides publicly available benchmarking datasets. The CAMPI study provides two different samples, including a simplified laboratory-assembled human intestinal gut sample (a mock community known as SIHUMIx) and a human fecal sample[13]. This dataset is available via the PRIDE (https://www.ebi.ac.uk/pride/) partner repository with the data set identifier PXD023217, and contains 21 mass spectrometry files for the SIHUMIx sample and 21 for the human fecal sample. Second, the model will be tested on the IBD dataset from Lehman *et al.* (2019)[14]. This dataset, which is also publicly available on PRIDE (PXD010371), contains 77 raw files with in total 3.3 million fragment ion spectra.

2. Data processing

The raw mass spectrometry files downloaded from PRIDE (https://www.ebi.ac.uk/pride/) will be converted to mascot generic format (MGF) with the ThermoRawFileParser[15]. The obtained datasets will be randomly split into a training and validation set. Some biologically relevant datasets will be withheld during the training stage and will only be used later as a test set. During the spectral

matching, the databases that make up the search space will be the same as the databases used in the paper from which the data was acquired but without the addition of the decoy databases, only the target databases will be used.

The search parameters used in the paper by Van Den Bossche *et al.* (2021) are as follows: up to two missed cleavages are allowed, 10 ppm precursor mass tolerance, and 0.02 Da fragment ion tolerance. Only precursors with charge +2 or +3 are analyzed[13]. In the paper by Lehman *et al.* (2019) the following parameters for the protein database search were applied: the enzyme used for protein digestion is trypsin with one missed cleavage allowed, only the monoisotopic mass is taken into consideration, carbamidomethylation is set as a fixed modification and methionine oxidation is set as a variable modification, the peptide tolerance is ±10 ppm and the MS/MS fragment tolerance is set to ± 0.5 Da, 113C, peptide ions with a charge of +2 or +3[14].

3. Model training

The code for this machine learning model will be written in the Jupyter Notebook with Python as the main programming language. I will train the model on mass spectrometry data so that it can accurately distinguish between good and bad PSMs without the use of a decoy database.

4. Model validation

The model will be tested against the results of the benchmark datasets, after which I can validate that the identification rate of the decoy-free model is on par with, or better than, the classical target-decoy or multi-step approach without decreasing sensitivity or specificity.

5. Workflow management and code availability

Since I will need to efficiently handle a vast amount of data during this thesis, I will incorporate the use of Nextflow as my workflow framework and use this to build and execute the data analysis pipelines. All the code will be written in the Jupyter Notebook and will be made available on GitHub. GitHub will mostly be used as a digital lab notebook and as a repository for all the code to make sure it is easily retrievable.

6. Gantt chart

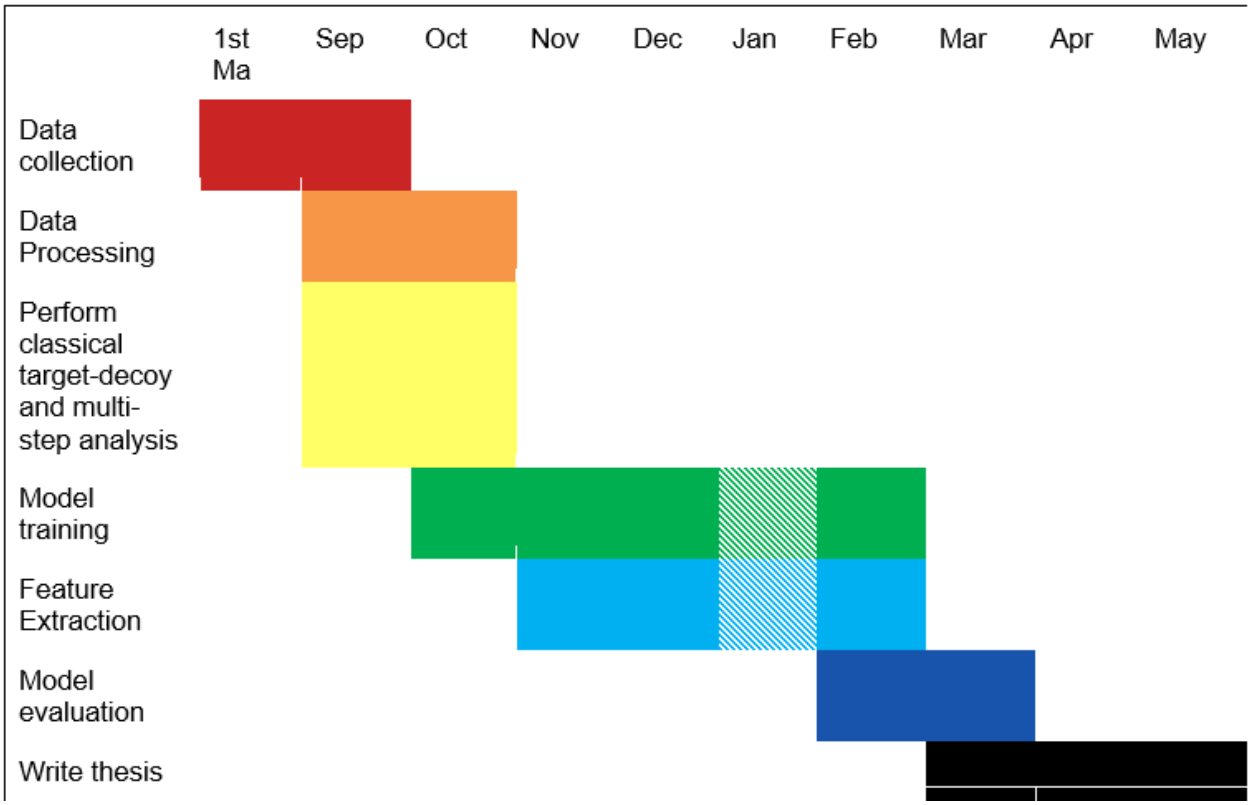

Figure 2 Gantt chart, January is shaded because of the exam period

REFERENCES

1 Wilmes, P. & Bond, P. L. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology* **6**, 911-920, doi:https://doi.org/10.1111/j.1462-2920.2004.00687.x (2004).

2 Hettich, R. L., Sharma, R., Chourey, K. & Giannone, R. J. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Current Opinion in Microbiology* **15**, 373-380, doi:https://doi.org/10.1016/j.mib.2012.04.008 (2012).

3 Barnouin, K. Guidelines for experimental design and data analysis of proteomic mass spectrometry-based experiments. *Amino Acids* **40**, 259-260, doi:10.1007/s00726-010-0750-9 (2011).

4 Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of proteome research* **7**, 40-44, doi:10.1021/pr700739d (2008).

5 Muth, T., Benndorf, D., Reichl, U., Rapp, E. & Martens, L. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems* **9**, 578-585, doi:10.1039/C2MB25415H (2013).

6 Muth, T. *et al.* Navigating through metaproteomics data: A logbook of database searching. *PROTEOMICS* **15**, 3439-3453, doi:https://doi.org/10.1002/pmic.201400560 (2015).

7 Sticker, A., Martens, L. & Clement, L. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nature Methods* **14**, 643-644, doi:10.1038/nmeth.4338 (2017).

8 Declercq, A. *et al.* MS²Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *bioRxiv*, 2021.2011.2002.466886, doi:10.1101/2021.11.02.466886 (2022).

9 Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923-925, doi:10.1038/nmeth1113 (2007).

10 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).

11 Gonnelli, G. *et al.* A Decoy-Free Approach to the Identification of Peptides. *Journal of Proteome Research* **14**, 1792-1798, doi:10.1021/pr501164r (2015).

12 Degroeve, S. *et al.* ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. *bioRxiv*, 2021.2007.2002.450686, doi:10.1101/2021.07.02.450686 (2022).

13 Van Den Bossche, T. *et al.* Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nature Communications* **12**, 7305, doi:10.1038/s41467-021-27542-8 (2021).

14 Lehmann, T. *et al.* Metaproteomics of fecal samples of Crohn's disease and Ulcerative Colitis. *Journal of Proteomics* **201**, 93-103, doi:https://doi.org/10.1016/j.jprot.2019.04.009 (2019).

15 Hulstaert, N. *et al.* ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *Journal of Proteome Research* **19**, 537-542, doi:10.1021/acs.jproteome.9b00328 (2020).