Large scale metaproteomic analysis – powered by a machine learning metric embedding approach

The study of the human microbiome - the entire community of all bacteria, viruses, and fungi colonizing for instance the human gut - has become a hot topic in life science, as the interplay of the microbial community has implications on human health. Many human diseases have been linked to the microbiome. Metaproteomics allows studying a complex protein mixture from several species and gives deep insights into such a microbiome. The standard way of analyzing it is by mass spectrometry: proteins are cut in small pieces and then weighted on a molecular scale (with suboptimal signal-to-noise performance), making the computational and statistical analyses of these sample a major challenges. The molecular scale result need to be robustly compared to millions of known protein subsequences to infer the content (and possible abnormalities) of a sample, Fig. 1.

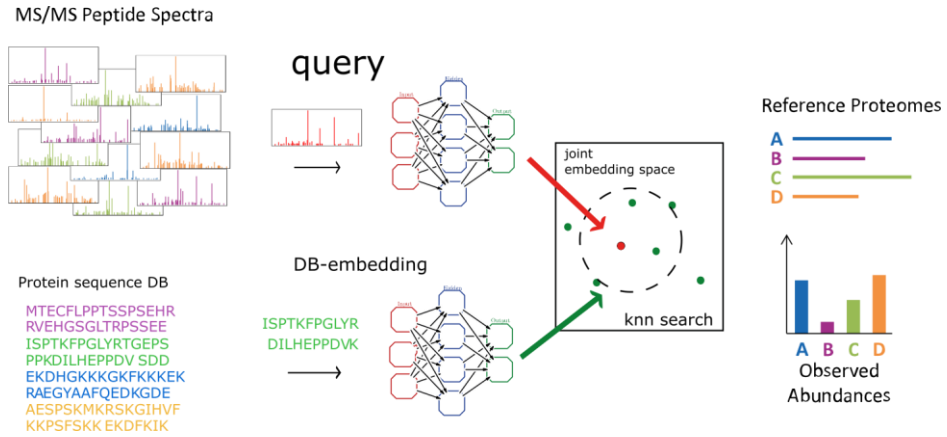

Figure 1: Sketch of the proposed procedure from mass spectra and sequences as input to the embedding.

In this master's project we will develop a new approach that is: directly searching a given spectrum against a very large protein sequence database (e.g. NCBIs non-redundant protein db) by machine learning a joint metric embedding, where the embedder is parametrized as a deep neural network.

The embedding is supposed to **jointly** embed **pairs of spectra-sequence**, such that spectra are directly comparable to sequences, by calculating their distance (e.g. manhattan distance) in embedding space, **see Fig 1**. In machine learning, such strategy is commonly used to embed pairs of e.g. image-text and hence match instances across two distinct domains (also called 'cross-modal retrieval').

low distance: between *embedding of a spectrum* **and a matching** *embedding of peptide sequence*

in contrast, a high distance for non-matching pairs:

high distance: between *embedding of a spectrum* **and any non-matching** *embedding of peptide sequence*

For training the joint-embedder we will use a public repository (http://www.ebi.ac.uk/pride/archive/projects/PXD010000) that contains a collection of 51 bacterial isolates consisting of approx. **1-million unique spectrum-peptide pairs.**

Once the embedder is trained we plan to perform a k-nearest neighbor search against embeddings from peptide sequences from the NCBI non-redundant protein sequence database. Ultimately, we

are interested to predict the taxonomy of a sample. To evaluate our approach we could estimate the taxonomy for a microbial mixture of a known composition, so-called mock community.

Additionally, we expect an extensive post-processing of the search results. For example, one would formulate constraints to exclude: **one-hit wonders** (e.g. proteins that have only one peptide assigned) or **redundant matches** (hits that are omni-present across distant species) to finally improve taxonomy estimation that stems from the distribution of resulting spectrum-sequence hits across the NCBI sequence database.

In this project, you will work with real-world data sets, for training and evaluation. Additionally, you will pioneer a multi-source data integration dealing with mass spectra and protein sequences. We offer existing code for pre-processing and training-pipeline (based on python, pyteomics and tensorflow).

We advocate including and realizing your own ideas. The project will be divided in submoduls. A first part will deal with data preparation and handling proteomic file formats. A large part will deal with stochastic gradient descent (SGD) keeping training instabilities, convergence issues under control. The core part will use deep learning for a joint embedding via metric learning and siamese-networks, including weight-tying/sharing. Hence, for training you will test several constrains such as large margin- or contrastive-losses. Finally, a last part will focus on an efficient k-nearest-neighbor search to query the resulting embedding space.

Biological expert-knowledge or experience with handling proteomic data is not required but we expect your interest in metaproteomic research as well as the ability to evaluate and visualize your results.

Contact

We offer this project for up to four students who will be supervised closely by Prof. Bernhard Renard and Tom Altenburg (tom.altenburg@hpi.de, F-E2.08).
If you have any questions, please do not hesitate to contact us.

References:

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804.

Steen, Hanno, and Matthias Mann. "The ABC's (and XYZ's) of peptide sequencing." *Nature reviews Molecular cell biology* 5.9 (2004): 699.

Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S., Renard, B. Y., Muth, T., & Martens, L. (2019). Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert review of proteomics*, *16*(5), 375-390.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic acids research* 35.suppl_1 (2006): D61-D65.

Matthew Monroe. DeNovo Peptide Identification Deep Learning Training Set, pride, V1; 2018. http://www.ebi.ac.uk/pride/archive/projects/PXD010000.

Tanca, A., Palomba, A., Pisanu, S., Deligios, M., Fraumene, C., Manghina, V., ... & Uzzau, S. (2014). A straightforward and efficient analytical pipeline for metaproteome characterization. *Microbiome*, *2*(1), 49.

Wang, Liwei, et al. "Learning two-branch neural networks for image-text matching tasks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018): 394-407.