From Metagenome to Metaproteome

Tom Naessens

Supervisor(s): prof. dr. Peter Dawyndt, ir. Bart Mesuere

Abstract— **Comparing and analysing genomes is usually done using BLAST [1]. We propose the Unipept Metagenomics Analysis Pipeline, a novel technique based on the conversion from metagenomics to metaproteomics and back, which makes use of the Unipept Metaproteomics Analysis Pipeline [5, 3]. To achieve this, a more specific version lowest common ancester (LCA) method is introduced for aggregating taxa. We also work out a proof of concept to modularise and extract the current visualisations in Unipept into a standalone Unipept visualisation framework, enabling users to inspect their data without uploading their data to the Unipept web application.**

Keywords—**metagenomics, metaproteomics, unipept, benchmarking**

I. Introduction

The Unipept platform is developed to map diversity in large and complex metaproteomic samples. The indexstructure of the underlying database has been finetuned to allow quick retrieval of all proteins containing a certain tryptic peptide. The taxon-specificity of a tryptic peptide can be determined based on the taxonomy information from UniProt[3].

Taxon-specificity of the tryptic peptide is successively derived from these occurrences using a novel lowest common ancestor approach that is robust against taxonomic misarrangements, misidentifications, and inaccuracies.

We introduce a new pipeline, the Unipept Metagenomic Analysis Pipeline or UMAP which allows the transformation of a metagenomics experiment into a metaproteomics experiment, using the functionality of Unipept described above. This pipeline is illustrated in Figure 1. For every DNA read in a metagenomics sample, genes are extracted with FragGeneScan[6] and are converted into proteins. These proteins are then split into tryptic peptides. Then, the Unipept Metaproteomics Analysis Pipeline is run to identify the consensus taxon for each tryptic peptide. The resulting taxa are afterwards aggregated using the LCA* algorithm.

II. LCA*

Unipept uses, as mentioned in the introduction, a robust implementation of the lowest common ancestor algorithm which is used for the taxonomic identification of one peptide. However, when the same algorithm is used to aggregate multiple peptides from the same organism, it does not always yield the most specific result.

Therefore, we introduce an adaptation of this algorithm which chooses more specific identifications over less specific ones when the input taxa lie on the same lineage. This new approach is illustrated in Figure 2 on the next page.

P. Dawyndt and B. Mesuere are associated with the Computational Biology reseach group from the Department of Applied Mathematics, Computer Science and Statistics, Ghent University (UGent), Ghent, Belgium. E-mail: Peter.Dawyndt@UGent.be, Bart.Mesuere@UGent.be.

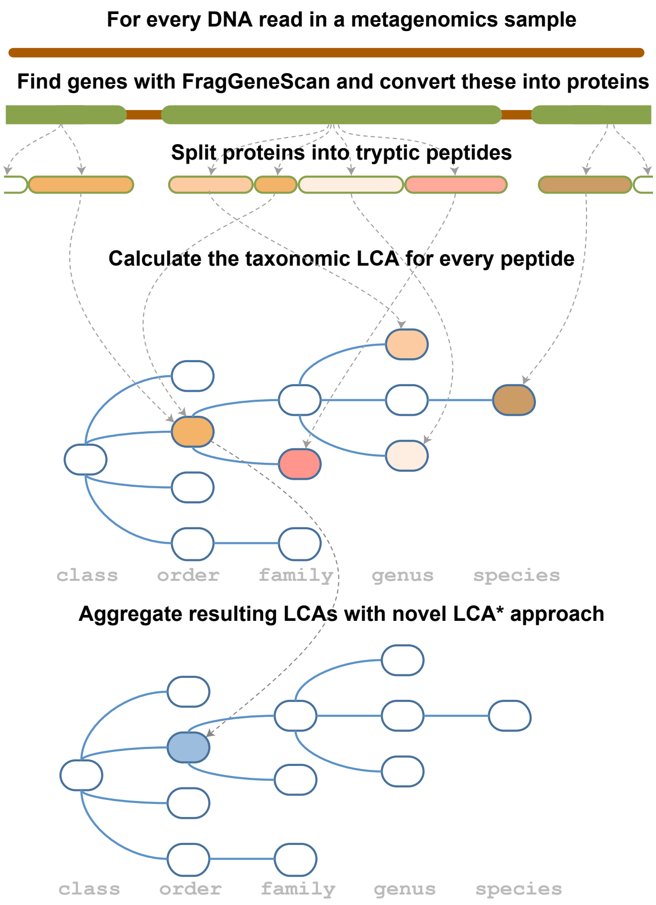
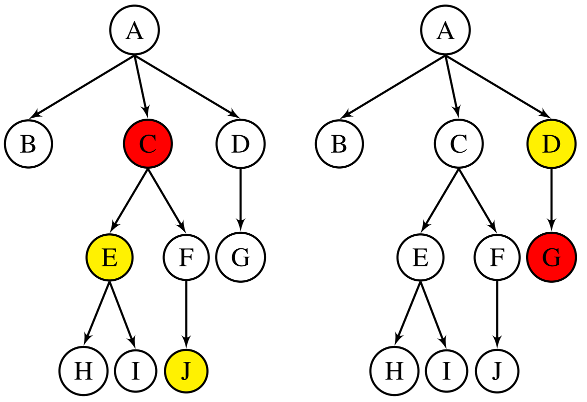

For every DNA read in a metagenomics sample
Find genes with FragGeneScan and convert these into proteins
Split proteins into tryptic peptides
Calculate the taxonomic LCA for every peptide
class
order
family
genus
species
Aggregate resulting LCAs with novel LCA* approach
class
order
family
genus
species

Fig. 1: Illustration of the Unipept Metagenomics Analysis Pipeline.

III. Benchmarking the Unipept Metagenomics Analysis Pipeline

To test the accuracy of the UMAP, we run two toolchains on three sorts of genomes: *i)* fully sequenced genomes, to be able to measure the accuracy of the taxonomic identification, *ii)* simulated unknown genomes (using a kind of leave-one-out strategy), to measure the accuracy of the taxonomic identification on genomes that have not been sequenced yet, and *iii)* genomes simulated with wgsim[4] based on fully sequenced genomes of read length 250 with different read error percentages of 0, 1, 2 and 5, to measure the effect of read errors on the identification. The two toolchains that were developed can be seen in Figure 3 on the following page. The left toolchain is the UMAP itself, while the right toolchain introduces a filtering step to simulate the unknown genomes.

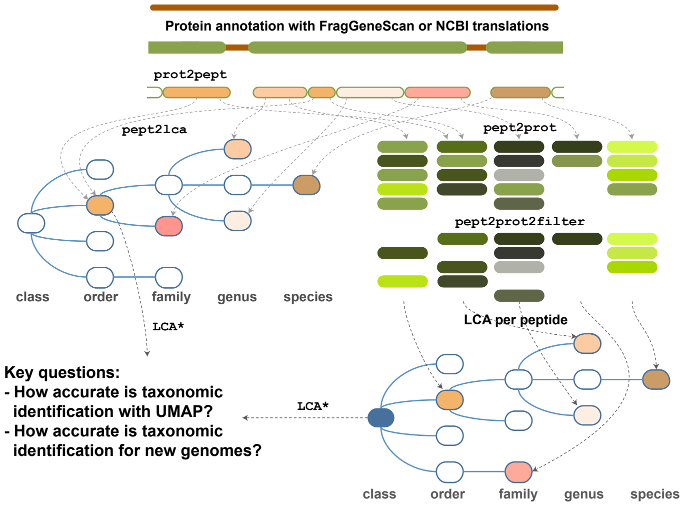
The UMAP has been found to identify fully sequenced genomes (*i*) very accurately with an average of 97% of the pro-



(a)

(b)

Fig. 2: Examples of the novel LCA* calculation. On the left, LCA*(E, J) is calculated, yielding node C as a result. On the right, the LCA of node D and G is calculated, yielding the more specific result G instead of the regular result D.



Key questions:
- How accurate is taxonomic identification with UMAP?
- How accurate is taxonomic identification for new genomes?

Fig. 3: The two toolchains used in the benchmarking process. The left toolchain is the UMAP itself, while the right toolchain introduces a filtering step to simulate the unknown genomes.

teins mapped to the species level. When the toolchain is being run on simulated unknown genomes (*ii*), a shift in specificity to the lesser accurate levels is observed. 38% of them could still be mapped to the species level. When reads were introduced into the genomes with wgsim (*iii*), we notice less specific results as the error percentage rises. The amount of proteins mapped to the species level drops just below 50% when the read error percentage crosses the level of 2%.

The full results of the benchmarking process were published as a poster, shown on the next page, which was presented during the annual BIG N2N symposium, edition 2015[2].

IV. Unipept Visualisation Framework

We also set the first steps to create a modular and abstract visualisation framework based on the current visualisations in the Unipept web application. This visualisation framework allows users to inspect their results visually and interactively, without having to enter their data in Unipept. As a proof of concept, the treeview visualisation (as can be seen in Figure 4) was extracted from the Unipept code and put into a separate modular JavaScript plugin.

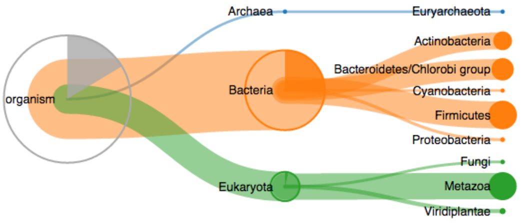


Fig. 4: Example of the treeview visualisation from Unipept

References

[1] Stephen F Altschul et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[2] *BIG N2N annual symposium | From Nucleotides to Networks*. http://www.bign2n.ugent.be/node/239. (Visited on 05/22/2015).

[3] UniProt Consortium et al. “UniProt: a hub for protein information”. In: *Nucleic Acids Research* (2014), gku989.

[4] H Li. *lh3/wgsim*. https://github.com/lh3/wgsim.

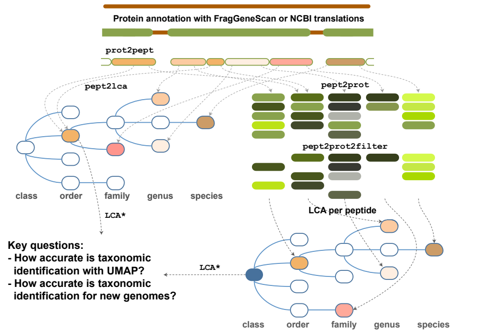
[5] Bart Mesuere et al. “Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples”. In: *Journal of proteome research* 11.12 (2012), pp. 5773–5780.

[6] Mina Rho, Haixu Tang, and Yuzhen Ye. “FragGeneScan: predicting genes in short and error-prone reads”. In: *Nucleic acids research* 38.20 (2010), e191–e191.

Benchmarking the Unipept Metagenomics Analysis Pipeline

T. Naessens, B.T. Habtemariam, R. Deklerck, S. Houbracken, M. Niklaus, I. Melckenbeeck
Promoter: Prof. Dr. Peter Dawyndt

Abstract



Key questions:
- How accurate is taxonomic identification with UMAP?
- How accurate is taxonomic identification for new genomes?

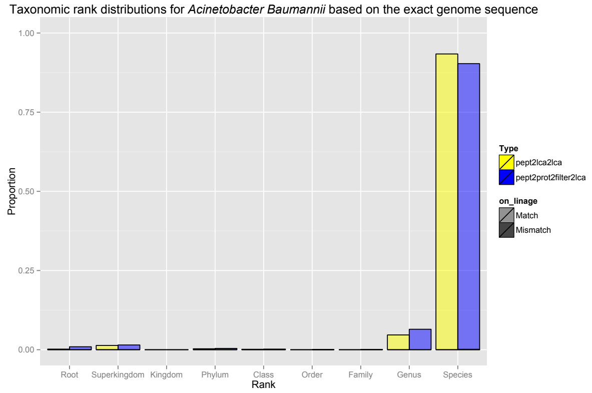
Context The Unipept Metagenomics Analysis Pipeline (UMAP) is an approach to solve the problem of taxonomic identification with metaproteomics. This is achieved by predicting all proteins on each DNA strand of a metagenomics sample, running the Unipept metaproteomics pipeline on these proteins, as indicated in the left hand side of the picture on the left, and aggregating them back to one resulting taxon. This last step is done using a novel LCA* algorithm, which exploits the fact that the proteins all originate from one DNA strand.

Approach To benchmark, we run the UMAP on both completely sequenced genomes and simulated reads from those genomes and compare the results with a separate analysis on those genomes, but where proteins that were found in Uniprot to be originating from that genome, are filtered out. This allows us to simulate what would happen if the UMAP is being run on unknown genomes while still producing comparable results.

Results To summarise obtained results, we have found that the UMAP performs very well for known genomes, where on average 97% is mapped on the species level. For simulated reads where no error was introduced, this number is reduced to 74% and lowers the more error is introduced. On the simulated unknown genomes, 38% is mapped to the species level.

Benchmarking results for *Acinetobacter baumannii*

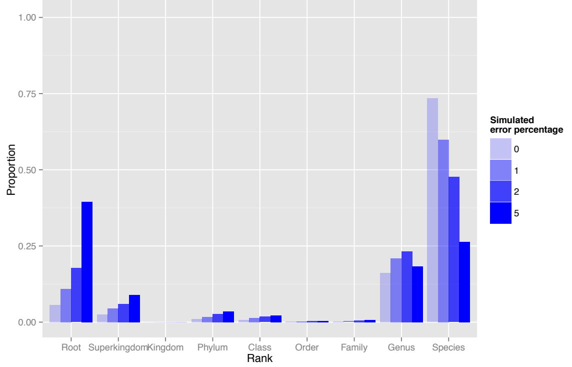
Taxonomic rank distributions for *Acinetobacter Baumannii* based on the exact genome sequence



The barplot above shows the taxonomic rank distribution of the results of both toolchains for the taxonomic identification of the peptides from the *Acinetobacter Baumannii* organism. The yellow bars show the result with the default UMAP toolchain, where the blue bars correspond with the different approach where the proteins from the original sequence have been filtered out in the identification process.

A specific level of identification is obtained for both toolchains. We also see a shift to the less specific ranks when we filter out the proteins already occurring in the originating genome. This shift is expected as this filter step causes a loss of specific information.

Taxonomic rank distributions for *Acinetobacter Baumannii* based on the simulated reads

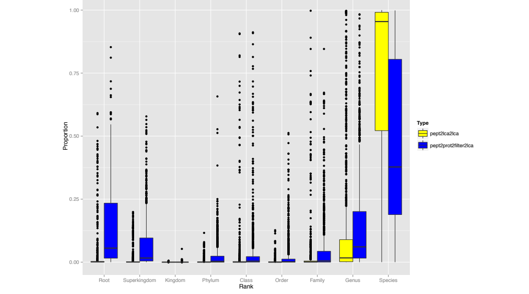


The above barplot shows the distribution of identifications found at the different levels in simulated reads on one genome with read lengths of 500 with 0%, 1%, 2% and 5% error rates.

As can be seen in the plot, the identification of the proteins is about 20% less accurate when using reads with 0% than when using the exact genome sequence. When using reads without errors, almost 75% of the proteins are mapped to the species rank and 20% to genus. Introducing errors in the reads predictably hampers the identification of the peptides, resulting in a worse specific identification when the error rate increases.

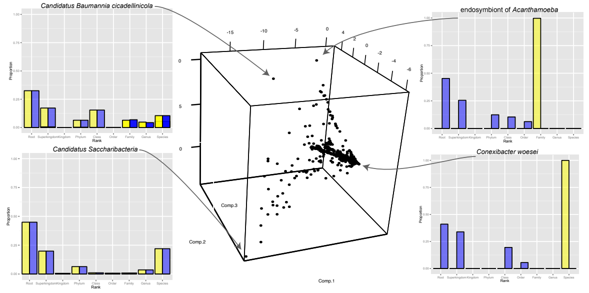
Total aggregated results for 1145 genomes

Per rank boxplot distribution plot of the percentage of taxonomic identifications for the aggregated results



Identification of proteins from 1145 genomes with (blue) and without (yellow) filtering out proteins that were found to be originating from the initial genomes. Without filtering, the identifications are on average for 97% mapped to species rank, with the remaining being mapped at genus rank. Enabling the filter step results in a general shift to the less specific ranks, which is expected as specific information is being filtered out.

3D PCA classification plot of the diversification of taxonomic idenftification in the aggregated results



Above is the PCA plot of the identification results for 1145 completely sequenced genomes. This plot clusters the genomes by the way their proteins are classified. Most are clustered along one line, but some outliers can be seen. The most extreme outliers have been accompanied by their correspondig bar plots. The reason for these outliers can vary from wrongly identified or classified genomes in the source database, genomes with very less or very much specific peptides or proteins, etc.