

Screening for taxon-specific peptides using dynamic index structures

Bart Mesuere¹, Griet Debyser², Peter Vandamme², Bart Devreese², Peter Dawyndt¹

1 Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281, B-9000 Ghent, Belgium

2 Ghent University, Department of Biochemistry and microbiology, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium

Intense antibiotic usage and disrupted digestion of cystic fibrosis patients probably has a dramatic effect on the composition and function of their gut microbiome. It is therefore appealing to determine the exact composition of this complex microbial community. In this research we explore a novel computational proteomics approach to explore the composition of microbial communities by using dynamic index structures for identifying taxon-specific peptides.

All complete bacterial genomes sequences available from the NCBI Reference Sequence (RefSeq) collection are processed in order to construct a unique peptide database. We extract all annotated protein sequences from each genome and perform an *in silico* trypsin digest to collect a list of unique peptides. The latter step is computationally intensive but can be speeded up considerably by using a prefix tree as a supporting data structure. As a side effect one can easily traverse the tree to find peptides that are strain-, species- or genus-specific. Given the dynamic nature of prefix trees, the data structure can easily be updated whenever new complete genomes become available or existing annotations are altered, without the need to rebuild the complete data structure.