Algorithm Details

Implementation of metaproteomic analysis faces numerous challenges due to its objective. Identification of the unknown organism by proteomic methods is challenging since most of the relevant high throughput proteomic methods were developed for identification of proteins of a known organism with an already existing protein database. In metaproteomics prior to the standard proteomic analysis, the organism of origin must be identified. Here we present an algorithm implemented in a command line tool MetaDirectMS1 on Python that is capable of identifying the most presented organisms from a united multi-organism protein database such as the full uniprot database and quantifying its proteins in a proteomic sample based on ultrafast LC-MS analysis.

To choose candidate organisms for presence in the sample from the initial database the process we called “blind search” was implemented. During blind search the algorithm matches peptide MS1 profiles (so called features) from the experimental sample on theoretical peptides from the reduced initial protein database by its m/z values. Accordingly, to find any organism as a candidate by such procedure, that organism should be presented in the initial protein database. Biosaur2 (https://doi.org/10.1002/rcm.9045) software is used to extract features of the peptides presented in the sample.

That brings to the necessity in the huge uniprot-like sized initial databases, which is not effective from the perspective of the time consumption and matching accuracy. To overcome that issue MetaDirectMS1 is constructing a reduced protein database on the stage of database processing. The algorithm groups all the organisms from the initial protein database into groups consisting of all subspecies of one species and in each group two leaders based on the greatest number of proteins according to the swissprot database and to the uniprot database are chosen. Next, all well-studied swissprot proteins and 2000 randomly chosen TrEMBL proteins for each group leader from the initial database are included in the reduced database. This operation allows to reduce database size for blind search stage while preserving the characterizing set of proteins for each organism. Peptide retention time values, which are often used for matching features for already identified peptide-spectrum-matches, are not used here because of low accuracy gain due to wide search space. Next, for each protein in the leading species database the probability of a random protein having that number of matched peptides is computed using the survival function for a binomial discrete random variable based on the number of matched peptides.

The number of proteins with probability of random match under a certain automatically optimized threshold are considered as a score for the organism. The purpose of the probability threshold here is not to find reliably identified proteins on certain criteria like 1% false-discovery-rate but to rank organisms based on overall quantity of its protein matches. Thus, based on such scoring a fixed number of top scored organisms are considered candidates for presence in the sample.

Then, a protein database of all organism candidates is constructed for each file and standard search (so-called preliminary) using LC-MS1 search engine ms1searchpy (https://doi.org/10.1021/acs.jproteome.0c00863) against the file-specific protein database is performed. Briefly speaking, peptide features are matched to theoretical peptides from the database using high precision retention time predictions and calibrated m/z accuracy coupled with protein identification with FDR control based on binomial distribution probability counting of the number of identified peptides related to protein. Afterwards the probability scores are then recalculated for each taxonomic group selected at a given taxonomic level

found among the candidate organisms. That allows to overcome drawbacks of the protein database combined from proteins of multiple organisms and to estimate the number of reliably identified proteins inside each taxonomy group.

To perform the final search (so-called precise), a combined protein database is assembled based on the number of proteins identified for each organism in the file-specific search. The purpose of additional filtering is to include in the combined protein database organisms that are consistently identified in individual replicates and to exclude underrepresented organisms in order to minimize the size of the final database. To do so, the algorithm by default includes in the combined database all organisms with the number of identified proteins higher than 2% of all identified proteins at least in one file. Then, the final search, called precise, is performed on each file against the combined protein database using the same search engine. The quantitative analysis for protein differential expression supporting multiple comparison groups is performed using directMS1quantmulti, based on the protein identification results of the precise search.