MScDB: A mass spectrometry centric protein sequence database

Harald Marx[1], Thomas Rattei[2], Simone Lemeer[1], Bernhard Kuster[1]

1 Technische Universitaet Muenchen, Emil-Erlenmeyer-Forum 5, 85354 Freising
2 University of Vienna, Althanstraße 14, 1090 Vienna

The genomes of higher eukaryotes typically contain a number of indistinguishable gene products for shotgun proteomics. The loss of information between peptides and proteins through the proteolytic digest leads to ambiguities in the identification and quantification of proteins due to the set of shared peptides. This fact is also known as the protein inference problem. To alleviate the protein inference problem, previous approaches are based on the *a posteriori* information of the protein identification to generate a minimal list of proteins to describe the observed peptides.

This work describes a novel *a priori* approach resulting in MScDB a database solely constructed to address the requirements of mass spectrometry based proteomics experiments. In contrast to current protein sequence databases, MScDB features the differences of proteins on the peptide instead of the sequence level. For the construction of MScDB an *in silico* digest of one or multiple protein sequence databases is performed. The resulting peptide lists are grouped together using a clustering algorithm which calculates the distances in an all against all comparison. From each connected component cluster, one or more representative sequences are chosen to generate the *a priori* minimal list of proteins without losing any relevant information for the identification. The resulting database is suitable for search engines from different vendors that use mass spectrometry data.

MScDB is useful for a wide range of cases.

i) Reduce redundancies and increase consistency in protein sequence databases.
ii) The incorporation of all the versions of a source database for increased reproducibility in experimental results over time.
iii) Cross-species experiments, to constrain the search space to the relevant and indistinguishable sequences between species.

For the version 3.72 of IPI HUMAN with 86392 entries, the *in silico* digest of the MScDB pipeline generates a non redundant list of 83716 peptide lists in less than 20 s. The clustering algorithm runs in quadratic time, grouping the non redundant list in an all against all approach in approximately 32 min for 3.5 x 10^9 comparisons. The resulting output of MScDB consists of a MS searchable Fasta file and a XML file for proteins and peptides including all the information of the source database and also allowing the assignment of the identifications to their respective clusters.

The high-performance generation of MScDB requires few system resources and together with the availability of the pipeline source code efficiently allows users to construct databases adapted to their individual requirements.