A new tool for protein identification by PMF

Tiengo A[(1)], Barbarini N[(1)], Troiani S[(2)], Rusconi L[(2)], Magni P[(1)]

[(1)] Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Pavia
[(2)] Biotechnology dep., Nerviano Medical Sciences, Nerviano

Motivation

Protein identification can be performed through different approaches. One of them is the Peptide Mass Fingerprinting (PMF), which combines the mass spectrometry (MS) data with the search in a suitable protein database. MS is able to measure with high precision the mass-charge ratio (m/z) of charged molecules. There are several algorithms and software tools developed for protein identification, but due to the amount of data generated by MS, this analysis remains a hard task from a computational point of view. This work presents a Perl procedure, called MsPI (Mass spectrometry Protein Identification), for protein identification by PMF approach.

Methods

PMF allows to generate a list of candidate proteins for a biological sample by comparing the acquired m/z with the ones stored in a database of proteins digested in silico, provided that the amino-acid sequence of the protein in the sample is already in the database. The PMF workflow can be subdivided in three steps: 1. the sample preparation and the spectrum acquisition; 2. the generation of the protein database; 3. the matching of the acquired spectrum against the generated database. 1. The protein components are separated from other cellular components and resolved by 2D-gel electrophoresis. Protein bands are excised from gel and digested with a protease, which cleaves the protein sequence at specific peptide bonds depending on the amino-acid sequence. The resultant peptides are then analyzed by MS, often in MALDI-TOF configuration, obtaining the acquired spectrum. 2. The enzymatic digestion is reproduced in silico for each known protein to create the search database, in which are stored the theoretical peptide masses associated to each protein. A protein database is chosen (e.g. SwissProt) and processed to obtain a new suitable database. Some missing information, such as molecular weight (MW) and isoelectric point (pI) are also added. To simulate the proteolytic digestion by the selected enzyme, a routine embedding the complex cleavage rules is implemented. The generated peptides also account for the presence of missed cleavages (MCs) and post-translational modifications (PTMs), i.e. chemical modifications of specific amino-acids affecting the MW. PTMs occur in cells or are induced by the preparation procedure. PTMs can be fixed or variable: in the first case, a PTM is present at each occurrence of the respective amino-acid, while in the second one the modification may or may not be present. 3. A list of peptide masses is extracted from the acquired spectrum. The contaminant masses that have to be removed are the skin keratins and the peptides produced by the autolysis of the protease used for enzymatic digestion. The resultant mass list is compared with the theoretical masses stored in the generated database. This comparison has to be made considering a tolerance window in Da or ppm around the experimental mass to take into account the intrinsic measurement error typical of the mass analyzer in use. The output of the comparison is a list of candidate proteins corresponding to the protein entries in the database which have at least one peptide corresponding to those of the experimental peak list. Each protein in the candidate list is evaluated with a scoring function to rank the results obtained. In MsPI are implemented the three scoring methods proposed by Samuelsson et al. (2004). Each score is based on different probabilistic hypotheses and one of them is implemented in a software tool available on-line (Piums). To limit the number of false positives (proteins wrongly included in the candidates list) in the ranked list a statistical validation of the results is also implemented in MsPI, through the construction of a randomly generated protein database.

Results

MsPI was tested on a dataset of 10 human proteins with the following parameters: up to two MCs, fixed carbamidomethyl modification, variable methionine oxidation modification (up to two for peptide) and two mass tolerances: 0.3 Da and 100 ppm. The results of MsPI were compared with those obtained by Piums and by another software tool, Mascot. The performance of MsPI is better than that of Piums. In fact, Piums allows to include in the candidates lists the proteins really present in the sample 4 times over 10, whereas MsPI 9 times over 10. Also Mascot correctly includes in the candidates list the proteins really present in the sample 9 times over

10, but MsPI has less false positives. Some of the false positive proteins included by MsPI have a bigger MW than the real proteins and then they can be easily removed from the list on the basis of MW determined through the electrophoresis. The number of false positives decreases to 5 for MsPI and to 6 for Mascot when the mass tolerance is fixed as relative and decreases to 3 for MsPI and to 4 for Mascot in the other case, highlighting that in this dataset MsPI minimizes the number of false positives.

Contact : alessandra.tiengo@unipv.it