ABSTRACT

The human proteome consists of more than 20.000 proteins. We analyzed the entire human proteome in order to identify the distribution of the shortest aminoacid sequences appearing solely in one protein, defined as core unique peptides. Furthermore, we defined the composite unique peptides that consist of overlapping core unique peptides.

We developed an algorithmic approach to populate a database containing the core and the composite unique peptides. We analyzed the whole human proteome from the perspective of core unique peptides and composite unique peptides, investigating the distribution of peptide length, aminoacid composition, starting position within the protein, density and coverage. We also analyzed the proteins that do not contain any unique peptides in order to understand their functionality.

More than $7x10^6$ core unique peptides have been identified forming $6.8x10^4$ composite unique peptides. The majority of the core unique peptides (~ 72%) consist of 6 aminoacids and approximately 20% consists of 7 aminoacids. The majority of composite unique peptides has a sequence length of 11-12 aminoacids while they comprise by 5-6 core unique peptides while most of them (~ 30%) appear at the protein's first aminoacids. Specific protein groups have been analyzed through these perspectives and results will be presented.

The results of the present study will be very useful for the identification of proteins by mass spectrometry and the application of selective reaction monitoring on complex protein mixtures.