Linguistic proteomics: highly parallel supercomputing for searching peptide

N.Pucello, G.Giuliano, V.Rosato

ENEA, Centro Ricerche Casaccia, Via Anguillarese 301, 00100 ROMA

Using the high-performance, highly parallel supercomputing devices available at ENEA, as well as a dedicated FPGA device (Rosato et al.) we have generated peptide "words" of 2, 3,.. 6 letters and searched for their occurence in various proteomes, including yeast, Drosophila, and E. coli. This task is incresingly demanding, in computational terms, with increasing peptide lengths, since the number of possible peptides is 20^n, where n is the length of a peptide. The significance of the occurrence of a peptide has been weighted using the "Coefficient of representation" (Cr), calculated through the formula (observed frequency)/(expected frequency). The preliminary results will be discussed. They indicate that the peptide distributions in the various proteomes is non-random, and that the presence of an amino acid at a given protein position is able to influence the choice of its nearest neighbours up to position +/-5.