Linguistic studies on thermophilic proteomes

V. Rosato, N. Pucello, L. Ferraro, G. Giuliano

ENEA - Via Anguillarese 301

We have applied a linguistic analysis to study the amino acid distribution in the proteomes of several organisms. This analysis shows that the presence of an amino acid at a given position positively influences the presence of identical amino acids at nearby positions (amino acid clustering). Clustering extends well beyond the closest neighboring sites and is particularly pronounced for some amino acids, like cysteine and tryptophan. Cysteine clusters preferentially in the form of CXXC structures ("knuckles"), often involved in metal coordination or disulfide bond formation. Differently from all other amino acids, cysteine clustering correlates with the growth temperature of the organism. This seems to be a general property of living organisms.

In a separate approach, we determined the frequency of the peptide "words" (2-peptides to 5-peptides) in different proteomes, and calculated a "representation coefficient", measuring their over- or under-representation with respect to the "expected" frequency. We plotted the representation coefficients of the same peptide in different bacteria vs the growth temperature of the organisms. On the basis of this analysis, a dictionary of "thermophilic" and "thermophobic" peptides has been constructed.