Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes

Oxana V. GALZITSKAYA, Michail Yu. LOBANOV

Institute of Protein Research RAS, Pushchino, 142290, Russia, ogalzit@vega.protres.ru

We have constructed the clustered Protein Data Bank and obtained clusters of chains of different identity inside each cluster. We have compiled the largest database of disordered patterns (141) from the clustered PDB where identity between chains inside of a cluster is larger or equal to 75% by using simple rules of selection. The results of these analyses would help to further our understanding of the physicochemical and structural determinants of intrinsically disordered regions that serve as molecular recognition elements. We have analyzed the occurrence of the selected patterns in 97 eukaryotic and in 26 bacterial proteomes. The disordered patterns appear more often in eukaryotic than in bacterial proteomes. The matrix of correlation coefficients between numbers of proteins where a disordered pattern from the library of 141 disordered patterns appears at least once in 9 kingdoms of eukaryota and 5 phyla of bacteria have been calculated. As a rule, the correlation coefficients are higher inside of the considered kingdom than between them. The patterns with the frequent occurrence in proteomes have low complexity (PPPPP, GGGGG, EEEED, HHHH, KKKKK, SSTSS, QQQQQP), and the type of patterns vary across different proteomes. The largest fraction of homorepeats of 6 residues belongs to Amoebozoa proteomes (D. discoideum), 46%. Moreover, the longest uninterrupted repeats belong to S306 from D. discoideum (Amoebozoa). Homorepeats of some amino acids occur more frequently than others and the type of homorepeats vary across different proteomes. For example, E6 appears most frequent for all considered proteomes for Chordata, Q6 for Arthropoda, S6 for Nematoda. The averaged occurrence of multiple long runs of 6 amino acids in a decreasing order for 97 eukaryotic proteomes is as follows: Q6, S6, A6, G6, N6, E6, P6, T6, D6, K6, L6, H6, R6, F6, V6, I6, Y6, C6, M6, W6, and for 26 bacterial proteomes it is A6, G6,

P6, and the others occur seldom. This suggests that such short similar motifs are responsible for common functions for nonhomologous, unrelated proteins from different organisms. A new method (IsUnstruct) based on the Ising model for prediction of disordered residues from protein sequence alone has been developed. The general idea is new and has the distinct advantage over various machine learning methods. For this method we have used the potentials derived from the clustered Protein Data Bank where there are clusters of chains of different identity inside each cluster. For the first time we have added in our method the library of disordered patterns (141) constructed from the clustered PDB. The IsUnstruct has been compared with other available methods and found to perform well.

This work was supported by the programs "Molecular and cellular biology" and “Fundamental sciences – medicine”, by the Russian Foundation for Basic Research