Sequence Bias in PDB Proteins: Comparison of Dipeptidyl Fragment Counts vs. the Residue Composition

Felder, C.[1], **Einav, U.**[1], **Segal, D.**[1], **Sussman, J.**[1], **Silman, I.**[2], **Beckmann, J.**[3] and **Yakir, B.**[4]

[1] Dept. of Structural Biology, [2] Dept. of Neurobiology, [3] Dept. of Cellular Genetics,
Weizmann Institute of Science ; [4]Dept. of Biological Statistics, Hebrew University of Jerusalem

Examination of sequence frequencies of dipeptidyl units of PDB proteins reveals a bias toward certain sequences relative to what would be expected from a random assembly from the residue composition. A database of the frequency counts of all possible 400 dipeptidyl fragment sequences in PDB proteins was constructed, using the PDB_select list of Hobohm and Sander at 90% homology to eliminate redundant entries. A parallel database of sequence composition was also made. From these data we calculated the observed probability of each dipeptidyl sequence, eg. the raw count divided by the total number of dipeptide fragments in all proteins, against what would be expected from a random combination of residues based on the residue composition. We noted a clear bias in favor of certain dipeptides, such as CH, MM, HP, YW, YC and QQ; and against other dipeptides, such as LW, MW, EC, EP, ES, CV and GP. The ratio of observed over expected probability ranges from 0.7 to 1.5, with an average near 1.0 and std. dev. 1.12. The results suggest that certain combinations of residues may be preferred to facilitate proper folding and function of the proteins.