Fast classification of protein three-dimensional structures

O.Carugo

Dip. Chimica Generale, Univ. di Pavia, via Taramelli 12, 27100 Pavia

The comparison of macromolecular three-dimensional (3D) models is of fundamental importance in molecular biology.

If these models are very similar, the root-mean-square-distance (rmsd) between equivalent atoms is usually taken as a measure of similarity. Despite its routine use, the rmsd must be standardized, since it depends on the size of the proteins. Intuitively, large proteins have a higher probability to differ than small proteins. A general, simple standardization procedure is derived from extensive simulations on 200 protein 3D structures. The new measure of similarity, rmsd_100 = rms / [-1.3+0.5ln (number of residues)], is the rmsd that would be observed for a pair of structures of 100 residues exhibiting the same degree of similarity.

If the protein 3D models are very different, the rmsd value has little significance and they must be compared with other approaches. The novel, fast procedure presented here is based on the contingency table analysis of the distributions of the Calpha(i)-Calpha(i+n) distances, with $3 < n < 30$. Lower n values monitor the secondary structure content, while higher n values reflect the protein topology. Such an approach, very simple and fast, automatically provides a probability of identity between two 3D models. Its efficiency compares well with that of more sofisticated and time consuming methods presently available.