2P05

Empirical Energy Potentials and Reference States in Fold and Sequence Recognition[1].

Sanzo Miyazawa
Faculty of Technology, Gunma University

An empirical energy potential, which consists of three terms, i.e., an attractive inter-residue contact potential, a repulsive packing potential, and a secondary structure potential, has been estimated for simulation and threading from residue arrangement in protein structures. Attractive inter-residue contact energies for proteins are re-evaluated with the same assumptions and approximations used originally by us in 1985, but with a significantly larger set of protein crystal structures. An additional repulsive packing energy term, operative at higher densities to prevent overpacking, has also been estimated for all 20 amino acids as a function of the number of contacting residues based on their observed distributions. The two terms of opposite sign are intended to be used together to provide an estimation of tertiary structure energies of inter-residue interactions in simplified proteins without atomic details. For secondary structures of proteins, short range interactions those within five residues are also empirically evaluated as potentials of mean force from the observed frequencies of secondary structures in known protein structures. The effects of longer range interactions on secondary structures are taken into account only as a mean field, ignoring any correlations between the long-range and short-range interactions. The ensemble of secondary structures observed in known protein structures is assumed to be in equilibrium with respect to the short range interactions. The secondary structure potentials are approximated as additive contributions from neighboring residues along the sequence. A secondary conformation at each residue position in a protein is described by a tripeptide, including one nearest neighbor on each side. Intrinsic potentials of residues, potentials of backbone - backbone interactions, and of side chain - backbone interactions are evaluated separately. Interactions among side chains are neglected, because of the relatively limited number of protein structures.

To overcome the problem of how to utilize the many homologous proteins in the Protein Data Bank, a new scheme has been devised to assign different weights to each protein, based on similarities among amino acid sequences. 1168 protein structures containing 1661 subunit sequences are actually used here. After the sequence weights are applied, these correspond to an effective residue number of 54,356 and an effective number of residue-residue contacts of 113,914, or about 6 times more than were used in the old analysis. Remarkably the new attractive contact energies are nearly identical to the old ones except for those with Leu and the rarer amino acids Trp and Met. The largest change found for Leu is surprising. The estimates of hydrophobicity from the contact energies for nonpolar side chains agree well with the experimental values.

In an application of these potentials, the sequences of 88 structurally distinct proteins in the Protein Data Bank are threaded at all possible positions without gaps into 189 different folds of proteins whose sequences differ from each other by at least 35% sequence identity. Both the native structures and the native sequences for these sequences or structures, excluding exceptional proteins such as membrane proteins, are demonstrated to have the lowest alignment energies of residues. Including the secondary structure potentials significantly improve fold and sequence recognitions. We also demonstrate how to define a single reference state for these potentials that is appropriate for both fold and sequence recognitions, including both multimeric and monomeric proteins.

Reference: 1. *J. Mol. Biol.* (1996) **256**, 623-644.
Address: 376 桐生市天神町 1-5-1; Kiryu, Gunma 376, JAPAN
Tel: 0277-30-1940, E-mail: miyazawa@smlab.eg.gunma-u.ac.jp