How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins?

Sanzo Miyazawa [1], Robert L. Jernigan[2]
miyazawa@smlab.sci.gunma-u.ac.jp, Faculty of Technology, Gunma University, Japan[1]
L. H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, USA[2]

"Statistical potentials", which are estimated from statistical preferences observed in protein crystal structures, are widely used to recognize and fold protein structures as well as to assess interactions between proteins. A pairwise isotropic contact potential between the 20 types of amino acids is a typical statistical potential. In order to extend the capability of the contact potential, we estimate the statistical distribution of relative orientations between contacting residues from a database of protein structures and evaluate the potential of mean force for relative orientations between contacting residue. A total contact energy for contacting residue pair is evaluated as a sum of the isotropic contact energy and the present orientational potential. Polar angles and Euler angles are used to specify two degrees of directional freedom and three degrees of rotational freedom for the orientation of one residue relative to another in contacting residues, respectively. A local coordinate system affixed to each residue based only on main chain atoms is defined for fold recognition. The 4435 protein domains defined in SCOP-1.61 were used with sampling weights determined on the basis of a sequence identity matrix between them; the effective number of contacting residue pairs used is equal to 1467302. The number of contacting residue pairs in the database will severely limit the resolution of the statistical distribution of relative orientations, if it is estimated by dividing space into cells and counting samples observed in each cell. To overcome such problems and to evaluate the fully-anisotropic distributions of relative orientations as a function of polar and Euler angles, we choose a method[1] in which the observed distribution is represented as a sum of δ functions each of which represents the observed orientation of a contacting residue, and is evaluated as a series expansion of spherical harmonics functions. The sample size limits the frequencies of modes whose expansion coefficients can be reliably estimated. High frequency modes are statistically less reliable than low frequency modes. Each expansion coefficient is separately corrected for the sample size according to suggestions from a Bayesian statistical analysis. As a result, many expansion terms can be utilized to evaluate orientational distributions. Also, unlike other orientational potentials, the uniform distribution is used for a reference distribution in evaluating a potential of mean force for each type of contacting residue pair from its orientational distribution, so that residue-residue orientations can be fully utilized to recognize protein structures. The zero energy level of the orientational potential, which is formulated as a logarithm of the probability density, is defined such that the expected value of orientational energy for the native folds is equal to zero for each type of contacting residue pair. It is shown by using decoy sets ("Decoys'R'Us") that the discrimination power of the orientational potential in fold recognition increases by taking account of both the polar and the Euler angle dependences, and becomes comparable to that of a simple contact potential. In the result, the total energy potential taken as a simple sum of contact, orientation, and backbone (ϕ, ψ) potentials identifies native structures better than any other method including a CHARMM potential. In addition, the results strongly indicate that all these energy terms complement each other and are needed to recognize native structures in a wide range of decoys from near native to denatured structures.

1. Onizuka, K., Noguchi, T., Akiyama, Y., Matsuda, H.: *Intelligent Systems*, **17**, 48-54, (2002).