How effective for fold recognition are relative orientations between contacting residues in proteins?

Sanzo Miyazawa
miyazawa@smlab.sci.gunma-u.ac.jp
Faculty of Technology, Gunma University, Japan

Hydrophobic interactions are essential for proteins to fold. However, All-atom MD simulations to explicitly evaluate solvent effects take too much time. On the other hand, current atomic potentials with implicit treatments of solvent effects do not perform better than simple coarse-grained potentials in recognition of the native structures. Thus, many attempts to develop coarse-grained potentials that can distinguish the native folds from non-native structures have been made. Most of them are statistical potentials that are defined as ($-\log$ odds + const) to reproduce statistical distributions of residues in protein crystal structures; they include the various types of potentials such as atomic or residue potentials, contact or distance dependent potentials, and pairwise or multibody potentials. Also attempts of optimizing pairwise potentials to identify the native folds have been made, indicating that simple isotropic residue-residue interactions are insufficient for proteins to fold the stable native structures and other interactions must be taken into account. Here we try to extend the capability of a pairwise contact potential by taking account of anistropic effects in residue-residue interactions, although short-range interactions are important as indicated by a success of the fragment assembly method in de novo protein structure predictions. We have devised a statistical potential for relative orientations between contacting residues from their distributions in protein structures, and examined how effective relative orientations between residues are to identify the native folds.

A total contact energy for a contacting residue pair is evaluated as a sum of the isotropic contact energy and the 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 proposed by Onizuka et al. (*Intelligent Systems*, **17**, 48-54, 2002) 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, that is, by the pseudo counts method. As a result, many expansion terms can be utilized to evaluate orientational distributions. Also, unlike other orientational potentials, the uniform distribution rather than the overall distribution for all types of amino acid pairs is used for a reference distribution in evaluating a statistical potential for each type of contacting residue pair from its orientational distribution, so that residue-residue orientations can be fully evaluated 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.

The orientational distributions evaluated indicate that correlations between polar and Euler angle dependences significantly contribute to the orientational entropies of contacting residues. As a result, 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. The capabilities of the potentials to recognize protein native structures were assessed by using decoy sets called "Decoys'R'Us" (Samudrala and Levitt, 2000; http://dd.stanford.edu/), and compared with those of other potentials including the CHARMM potential with a generalized Born, Coulomb, non polar solvation and van der Waals energy terms. It is shown that 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. In addition, the results strongly indicate that all these energy terms complement each other and are needed to recognize the native structures in a wide range of decoys from near native to denatured structures.

Comparison of performance between potentials in fold recognition

Decoy ID range, Decoy family Potentials	# tops /# total	mean $\overline{\log P_e}$	mean $\overline{Z_e}$	mean $\overline{R}^1$	
1-7 "4state_reduced": 7 decoy sets					4-state off-lattice model
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	7/7	-6.50	-4.44	0.66	the present potential
Fain et al. (2002)	1/7	-4.45	-2.3	0.52	optimal Chebyshev-expanded potential
Toby and Elber (2000)	3/6	-5.42	-3.14		optimized distance-dependent potential
Samudrala and Moult (1998) [3]	6/7	-6.06	-2.67	0.67	atomic contact potential
Onizuka et al. (2002) [4]	7/7	-6.50	-3.41		orientational potential
Dominy and Brooks (2002) [5]	$\sim$ 7/7	$\sim$ -6.5	-3.4	0.55	CHARMM with GB+Coul+NPSolv+vdW
8-11 "fisa": 4 decoy sets					fragment insertion simulated annealing
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	2/4	-4.04	-2.55	0.26	the present potential
Toby and Elber (2000)	2/3		-3.34		optimized distance-dependent potential
Onizuka et al. (2002) [4]	1/3		-1.38		orientational potential
12-16 "fisa_casp3": 5 decoy sets					predicted by the Baker group for CASP3
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	2/5	-5.38	-3.61	0.16	the present potential
Toby and Elber (2000)	1/3		-3.94		optimized distance-dependent potential
Onizuka et al. (2002) [4]	1/3		-2.01		orientational potential
17-45 "hg_structal": 29 decoy sets					29 globins by comparative modeling
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	22/29	-2.76	-2.62	0.72	the present potential
Dominy and Brooks (2002) [5]	19/29		-2.0	0.69	CHARMM with GB+Coul+NPSolv+vdW
46-53 "lattice_ssfit": 8 decoy sets					8 small proteins generated by ab initio methods
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	8/8	-7.60	-11.12	-0.01	the present potential
Fain et al. (2002)	8/8	-7.60	-6.84		optimal Chebyshev-expanded potential
Toby and Elber (2000)	4/6	-6.89	-4.10		optimized distance-dependent potential
Samudrala and Moult (1998) [3]	8/8	-7.60	-6.46		atomic contact potential
Onizuka et al. (2002) [4]	6/6	-7.60	-6.22		orientational potential
54-63 "lmds": 10 decoy sets					10 small proteins in diverse classes
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	8/10	-4.89	-5.34	0.14	the present potential
Fain et al. (2002)	3/9	-4.55	-2.83		optimal Chebyshev-expanded potential
Toby and Elber (2000)	4/7	-5.32	-3.27		optimized distance-dependent potential
Samudrala and Moult (1998) [3]	3/9	-3.04	-0.58		atomic contact potential
Onizuka et al. (2002) [4]	5/7	-5.00	-3.67		orientational potential
64-73 "lmds_v2": 10 decoy sets					2nd version of the local minima decoy sets, "lmds"
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	8/10	-3.85	-5.03	0.18	the present potential
Fain et al. (2002)	1/2	-4.81	-3.15		optimal Chebyshev-expanded potential
Samudrala and Moult (1998) [3]	1/2	-4.47	-3.05		atomic contact potential
74-79 "semfold": 6 decoy sets					6 proteins
$(e^c_{rr} + \Delta e^c + e^o + e^s)^2$	4/6	-8.13	-3.86	0.08	the present potential
1-61 "ig_structal": 61 dcoy sets					61 immunoglobulin domains by comparative modeling
$(e^o + e^r + e^s)^2$	49/61	-3.55	-2.96	0.36	the present potential
62-81 "ig_structal_hires": 20 decoy sets					high resolution subset of "ig_structal"
$(e^o + e^r + e^s)^2$	19/20	-2.86	-4.31	0.43	the present potential

[2] e^c_{rr}: collapse; Δe^c: pairwise contact; e^o: orientational; e^r repulsive; e^s: (ϕ, ψ) energies

Reference: Miyazawa, S. and Jernigan, R. L., J. Chem. Phys. 122, 024901, 2005.