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

Sanzo Miyazawa[1] and Robert L. Jernigan[2]

[1] miyazawa@smlab.sci.gunma-u.ac.jp, Graduate School of Engineering, Gunma University, Japan

[2]jernigan@iastate.edu, Iowa State University, U.S.A.

presented at

The 21st annual meeting of Protein Society in 2007 in Boston, MA

(July 21-25, 2007)

ABSTRACT

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 residues. 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. To evaluate the fully-anisotropic distributions of relative orientations as a function of polar and Euler angles, we choose a method 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 evaluated. It is shown by using decoy sets that the discrimination power of the orientational potential in fold recognition increases by taking account of the Euler angle dependencies and becomes comparable to that of a simple contact potential, and that the total energy potential taken as a simple sum of contact, orientation, and (ϕ, ψ) potentials performs well to identify the native folds.

1. INTRODUCTION

Hydrophobic interactions are essential for proteins to fold. However,

- All-atom MD simulations to explicitly evaluate solvent effects take too much CPU time.
- Current atomic potentials with implicit treatments of solvent effects do not perform better than simple coarse-grained potentials in recognition of native structures.

Attempts to develop coarse-grained potentials that can distinguish native folds from decoys.

- Pairwise contact/distance-dependent energies.

Since then, many statistical/knowledge-base potentials ($\equiv -\log$ odds + const) are devised.

- Potentials at an atomic level.
- Multibody potentials.
- Optimized potentials to identify native folds; an isotropic pairwise potential is not sufficient to identify all native structures and other interactions must be taken into account.
- ...

Current status of statistical potentials for fold recognition

Capability of pairwise isotropic potentials:

- Pairwise isotropic interactions are insufficient for proteins to fold into the stable native structures.

  (Mirny & Shakhnovich, *J. Mol. Biol.*, **264**, 1164, 1996)

- No pairwise isotropic potential can identify all native structures.

  (Toby & Elber, *Proteins*, **41**, 40, 2000)

Extension of pairwise isotropic potentials:

- Multi-body isotropic potentials:
  - ⋆ Munson & Singh (1997) ,
  - ⋆ Liwo et al. (2001) .
- Two-body anisotropic potentials:
  - ⋆ Onizuka et al. (2002) ,

    They concluded that the discrimination power of potentials could not be improved by taking account of Euler angle dependencies in addition to radial and polar angle dependencies.
  - ⋆ Buchete et al. (2003) and (2004) .

    Only radial distance and polar angle dependencies were taken into account.

Purposes of the present work: To extend the capability of a pairwise contact potential.

- To evaluate the fully anisotropic distribution of relative orientations between residues in contact as a function of polar (θ, ϕ) and Euler (Θ, Φ, Ψ) angles.

  How to overcome limitations on the number of contacting residue pairs in the database?

- To assess the effectiveness of residue-residue orientations for fold recognition by using a statistical potential.

Distinctive features in the present method:

- Orientational distributions are estimated in the expansion with spherical harmonics functions.
  - ⋆ Expansion coefficients are evaluated from observed distributions that are represented as sums of δ function; this method was first proposed by Onizuka et al. (2002) .
  - ⋆ Higher order terms are ignored to remove artificial contributions from the small size of samples.
  - ⋆ Each expansion coefficient is separately corrected for the sample size depending on the resolution of each term.
  - ⋆ A local coordinate system for each residue is defined based only on main chain atoms.
- Orientational energy for contacting residues is evaluated as a correction term for contact energy.
- A reference state for the orientational potential is the uniform rather than overall distribution for residue-residue orientations.

2. METHODS

Contact potentials

Total contact energy:

$$E^c = \frac{1}{2}\sum_i \sum_{j \neq i} e^c(\vec{r}_i, \vec{r}_j) \tag{1}$$

The contact energy, $e^c(\vec{r}_i, \vec{r}_j)$, between the ith and jth residues is defined as

$$e^c(\vec{r}_i, \vec{r}_j) = \Delta^c(\vec{r}_i, \vec{r}_j)\,[\, e^c_{a_i a_j} + e^o{}_{a_i a_j}(\vec{r}_i, \vec{r}_j)\,] \tag{2}$$

where

$\vec{r}_i, \quad \vec{r}_j$	positions of ith and jth residues.
$\Delta^c(\vec{r}_i, \vec{r}_j)$	a switching function representing the degree of contact and sharply changing its value from one to zero around 6.5 Å as a function of the distance between the side-chain centers of ith and jth residues,
$e^c_{a_i a_j} = e^c_{rr} + \Delta e^c_{a_i a_j}$	an isotropic contact energy for residues of type a_i and a_j in contact,
$e^o{}_{a_i a_j}(\vec{r}_i, \vec{r}_j)$	an orientational energy between amino acids of type a_i and a_j,

Residue-residue orientational potentials between contacting residues

$$e^{o}{}_{aa'} = \frac{1}{2}\left[\{-\log f_{aa'} + <\log f_{aa'}>\} + \{-\log f_{a'a} + <\log f_{a'a}>\}\right] \quad (3)$$

where

$f_{aa'}(\theta, \phi, \Theta, \Phi, \Psi)$	a probability density function for a residue of type a'
	at the orientation $(\theta, \phi, \Theta, \Phi, \Psi)$ in relative to the residue of type a,
$f_{a'a}(\theta', \phi', \Theta', \Phi', \Psi')$	a probability density function for a residue of type a
	at the orientation $(\theta', \phi', \Theta', \Phi', \Psi')$ in relative to the residue of type a',
$\theta, \phi;\ \theta', \phi'$	polar angles to specify two degrees of directional freedom for the orientation,
$\Theta, \Phi, \Psi;\ \Theta', \Phi', \Psi'$	Euler angles to specify three degrees of rotational freedom for the orientation,
$< -\log f_{aa'} >$	orientational entropy as a reference state that is the uniform distribution.

How to estimate the distribution of residue-residue orientations.

Expansion in the products of spherical harmonics functions and trigonometric functions:

$$f_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) = \sum_{l_p=0}\sum_{m_p=-l_p}^{l_p}\sum_{l_e=0}\sum_{m_e=-l_e}^{l_e}\sum_{k_e} c^{aa'}_{l_p m_p l_e m_e k_e}\, g_{l_p m_p l_e m_e k_e}(\theta,\phi,\Theta,\Phi,\Psi) \tag{4}$$

$$g_{l_p m_p l_e m_e k_e} \equiv Y^{m_p}_{l_p}(\cos\theta,\phi) Y^{m_e}_{l_e}(\cos\Theta,\Phi) R_{k_e}(\Psi) \tag{5}$$

1. The coefficients are calculated from the observed polar and Euler angles, $(\theta_\mu,\phi_\mu,\Theta_\mu,\Phi_\mu,\Psi_\mu)$.

$$c^{aa'}_{l_p m_p l_e m_e k_e} = \sum_{\mu\in\{a\text{-}a'\}} w_\mu\, g_{l_p m_p l_e m_e k_e}(\theta_\mu,\phi_\mu,\Theta_\mu,\Phi_\mu,\Psi_\mu) / \sum_{\mu\in\{a\text{-}a'\}} w_\mu \tag{6}$$

2. Each coefficient is separately corrected for sample size by a pseudo count method.

3. Higher order terms ($O_{l_p m_p l_e m_e k_e} > O_{cutoff}$) are ignored.

   $O_{l_p m_p l_e m_e k_e}$ is defined to be the number of frequency modes lower than or equal to $(l_p, m_p, l_e, m_e, k_e)$.

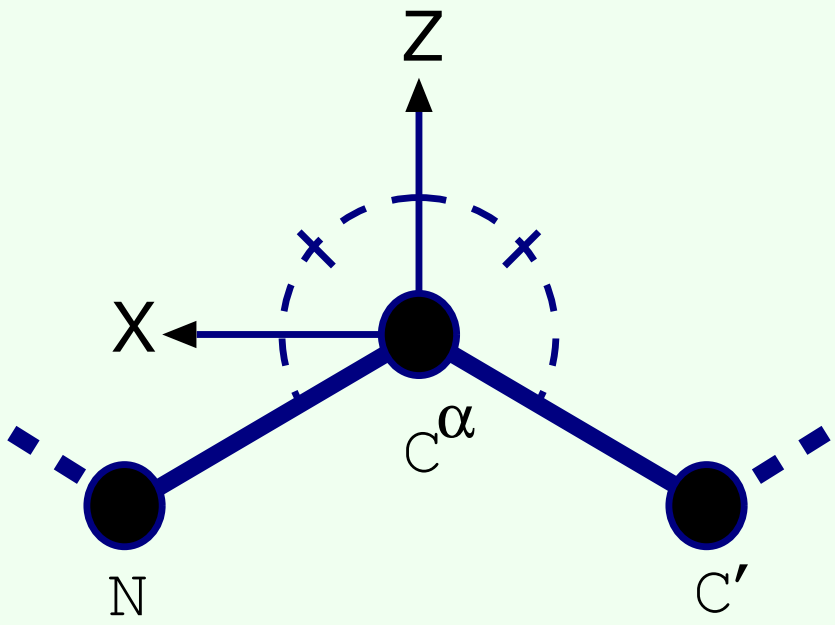
4. Terms with the small values of coefficients ($|c^{aa'}_{l_p m_p l_e m_e k_e}| < c_{cutoff}\, c^{aa'}_{00000}$) are neglected.

Datasets of protein structures used to estimate the orientational potentials

- Proteins that belong to class 1 to 5 in Release 1.61 of the SCOP have been used.
- Only structures better than 2.5 Å determined by X-ray are used.
- Species representatives of 4369 proteins are chosen by removing proteins included in the decoy set "Decoys'R'us".
- A sampling weight for each protein representative is calculated by the sampling method based on a sequence identity matrix between proteins; the effective numbers of sequences and contacts are 3506 and 1463806, respectively.

3. RESULTS

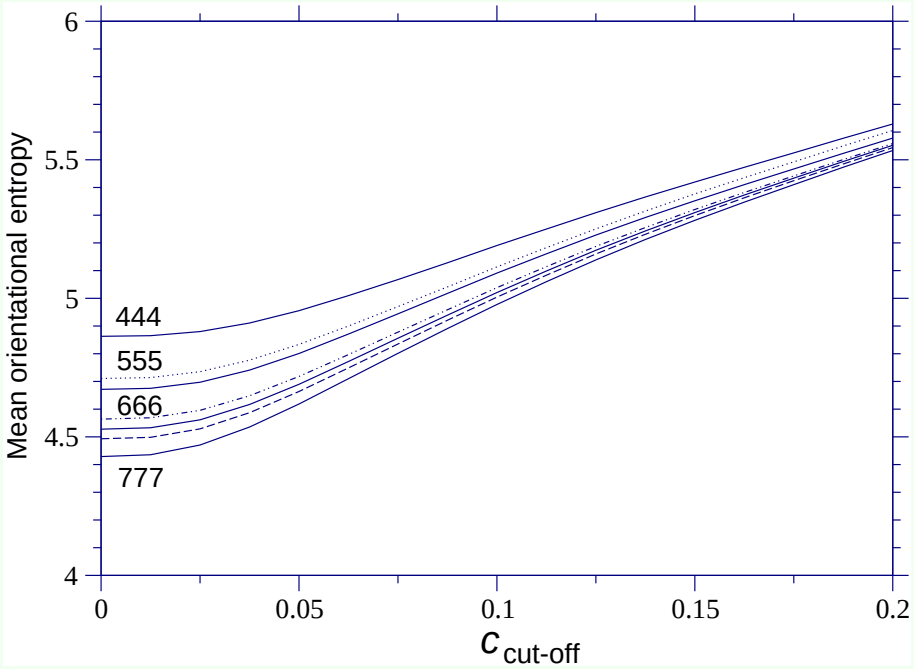
A local coordinate system affixed to each residue is based only on main chain atoms for fold recognition.

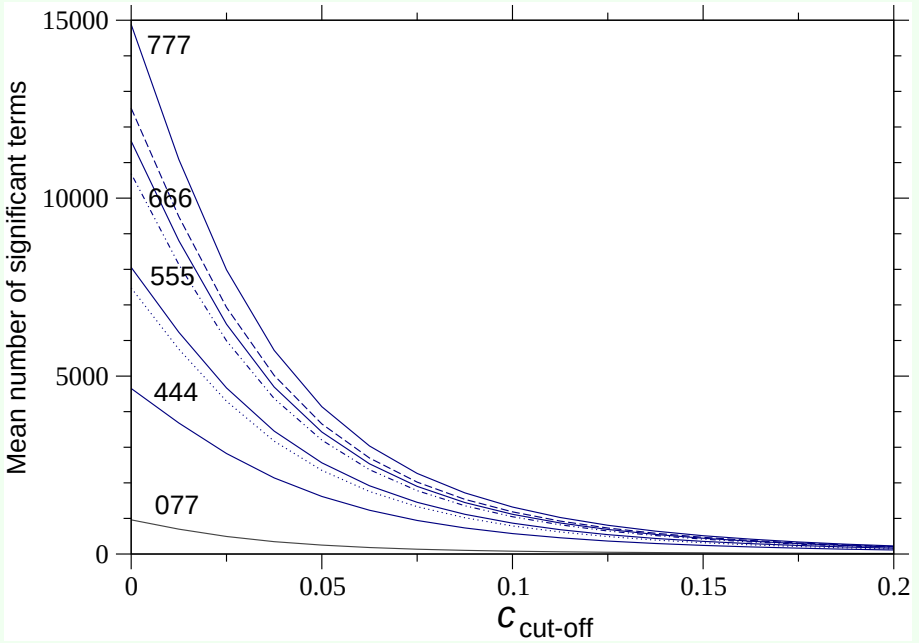


The origin O of the local coordinate system is located at the C^{α} position of each residue. The Y and Z axes are ones formed by the vector product and the sum of the unit vectors from N to C^{α} and from C' to C^{α}, respectively. The X axis is taken to form a right-handed coordinate system. The relative direction and rotation of one residue to the other in contacting residues are represented by polar angles (θ, ϕ) and Euler angles (Θ, Φ, Ψ), respectively.

Orientational distributions of contacting residues

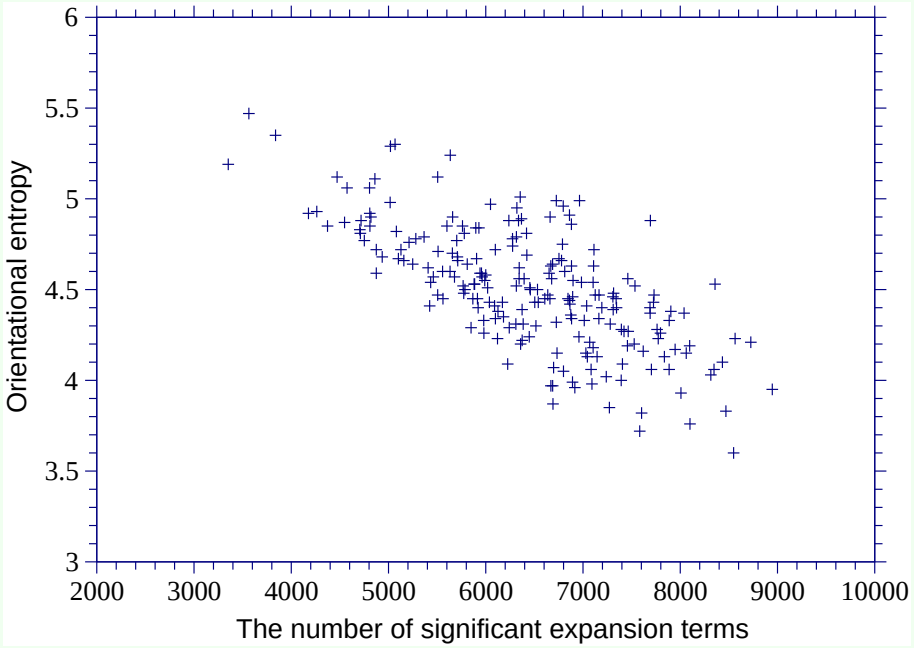
Dependencies of orientational entropies and the number of significant expansion terms on parameters

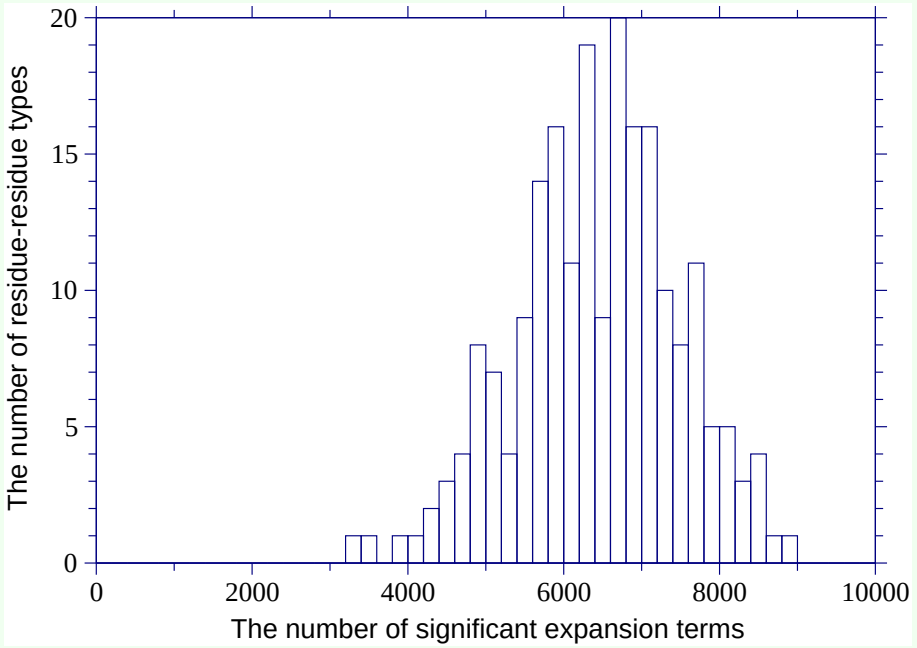




Triplets of digits near solid lines indicate the values of $(l_p^{max}, l_e^{max}, k_e^{max})$; for non-solid lines, $l_p^{max} = l_e^{max} = k_e^{max} = 6$ is used. The other parameters are: $\beta = 0.2$ for all lines, and $O_{cutoff} = O_{33333} = 1792$ for solid lines. The upper dotted line shows the case of $O_{cutoff} = O_{00777} = 960$, the lower dotted line is for $O_{cutoff} = O_{11555} = 1584$, and the dotted broken line is for $O_{cutoff} = O_{22444} = 2025$.

Correlation between the number of significant expansion terms and orientational entropy, and histograms of the numbers of significant expansion terms for the 210 types of residue pairs.

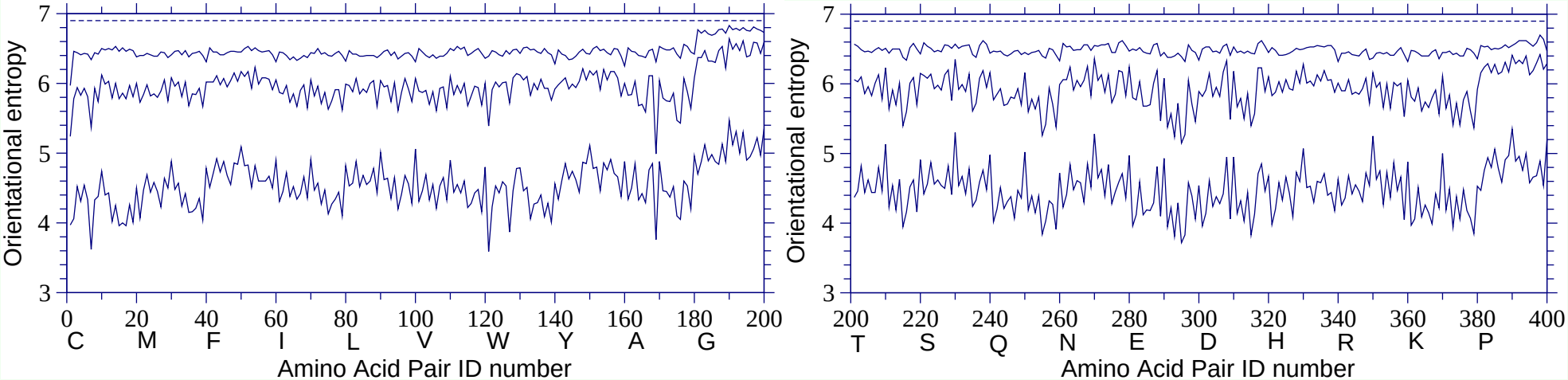




The orientational potentials are evaluated with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025$.

Distributions of residue orientations significantly depend on directional and rotational angles.

Orientational entropies for three types of distributions



The broken line: A uniform distribution.

The highest solid lIne: Only polar angle dependencies are taken into accont; $l_p^{max} = 6, l_e^{max} = k_e^{max} = 0.$

The lowest solid lIne: Polar and Euler angle dependencies are taken into accont; $l_p^{max} = l_e^{max} = k_e^{max} = 6.$

The middle solid lIne: No correlations between polar and Euler angle dependencies are taken into accont;

$l_p^{max} = 6, l_e^{max} = k_e^{max} = 0$ and $l_p^{max} = 0, l_e^{max} = k_e^{max} = 6.$

Recognition power for native structures

The performance of the potentials to identify native folds is evaluated by using the decoy database, "Decoys'R'Us" (Samudrala and Levitt, 1999).

Decoy families are categorized into two classes, because the true ground state of multimeric proteins requires all of the chains to be present.

1. **Monomeric protein decoy sets; 79 decoy sets in 8 decoy families.**

   These decoy sets are for monomeric proteins with a few exceptions such as tetrameric hemoglobins.

2. **Immunoglobulin decoy sets; 81 decoy sets in 2 decoy families.**

   Each of these decoy structures consists of a single chain of a multimer.

Native structures included in these decoys are removed from a protein data set that is used to evaluate orientational potentials.

Measures of performance:

- **The number of top ranks in the energy scale or in the RMSD scale.**
- **Logarithms of rank probabilities.**

$$P_e \equiv \text{the rank of the native fold in a energy scale } / \text{ the number of decoys} \tag{7}$$

$$P_r \equiv \text{the rank of the lowest energy fold in the RMSD scale/ the number of decoys} \tag{8}$$

- **Z scores.**

$$Z_e \equiv \frac{E_{native} - \overline{E_{decoy}}}{\sigma_E} \tag{9}$$

$$Z_r \equiv Z_{rmsd} \equiv \frac{RMSD_{lowest} - \overline{RMSD_{decoy}}}{\sigma_{rmsd}} \tag{10}$$

Recognition power for native folds is increased by taking account of Euler angle dependencies.

(A) Dependences only on polar angles are taken into account.

l_p^{max}	c_{cutoff}	$l_e^{max} = k_e^{max} = 0$, $\beta = 0.2$, $O_{cutoff} = \infty$							
		79 monomeric decoy sets				81 lg decoy sets			
		#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
7	0.0	30	-3.45	-2.60	-1.98	45	-2.93	-2.52	-1.57
14	0.0	31	-3.42	-2.89	-1.84	46	-2.87	-2.48	-1.91

(B) Dependences on both polar and Euler angles are taken into account.

l_p^{max}	c_{cutoff}	$l_e^{max} = k_e^{max} = l_p^{max}$, $\beta = 0.2$, $O_{cutoff} = 960$							
		79 monomeric decoy sets				81 lg decoy sets			
		#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
6	0.0	34	-3.80	-3.24	-2.32	60	-3.26	-3.25	-1.95
	0.025	37	-3.83	-3.33	-2.32	60	-3.24	-3.23	-1.92
l_p^{max}	c_{cutoff}	$l_e^{max} = k_e^{max} = l_p^{max}$, $\beta = 0.2$, $O_{cutoff} = 1792$							
6	0.0	37	-3.87	-3.35	-2.40	60	-3.28	-3.14	-2.01
	0.025	37	-3.88	-3.22	-2.38	59	-3.27	-3.11	-2.00

All energy components complement each other and are necessary for fold recognition.

(A) For the 79 monomeric decoy sets

Potentials[1]					# top ranks	mean	mean	mean	mean	median	median	mean
e^c_{rr}	$\Delta e^c_{aa'}$	e^o	e^r	e^s	# total = 79	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	$\overline{Z_{rmsd}}$	Z_e	Z_{rmsd}	$\overline{R}$
		e^o			37	-3.88	-3.22	-2.38	-2.49	-2.09	-1.65	0.33
	Δe^c	+ e^o			52	-4.53	-4.24	-3.18	-3.19	-2.79	-2.60	0.37
e^c_{rr} +	Δe^c	+ e^o			58	-4.79	-4.88	-4.38	-3.92	-4.08	-3.55	0.40
e^c_{rr} +	Δe^c	+ e^o +		e^s	61	-4.63	-4.63	-4.45	-3.68	-4.11	-3.41	0.39

(B) For the 81 immunoglobulin decoy sets

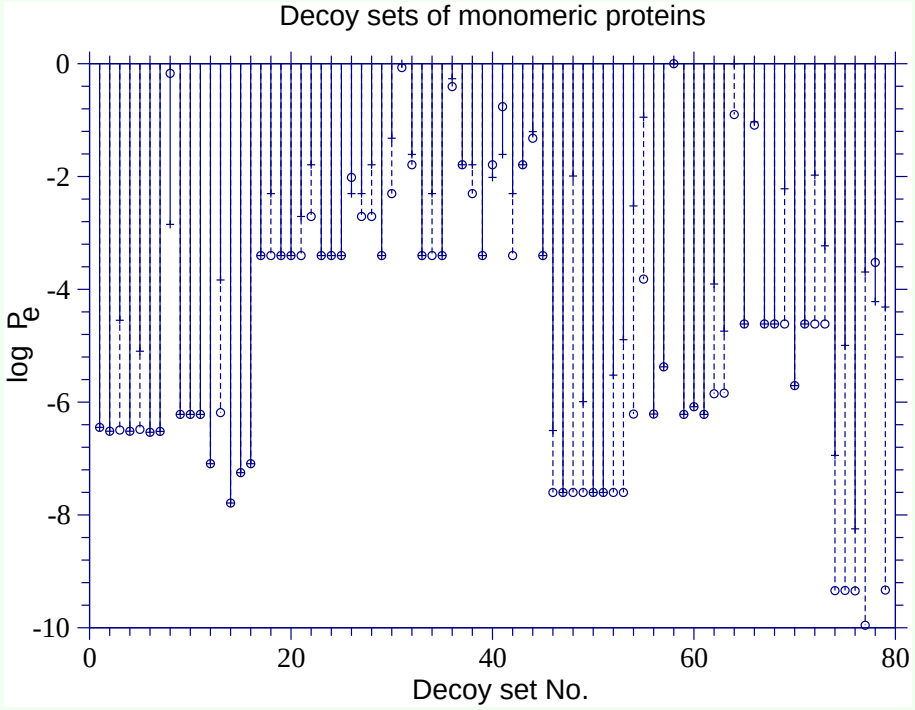
The true ground state for the contact potentials, e^c_{rr} and Δe^c_{ij}, requires all of the chains to be present.

Potentials[1]					# top ranks	mean	mean	mean	mean	median	median	mean
e^c_{rr}	$\Delta e^c_{aa'}$	e^o	e^r	e^s	# total = 81	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	$\overline{Z_{rmsd}}$	Z_e	Z_{rmsd}	$\overline{R}$
		e^o			59	-3.27	-3.11	-2.00	-2.74	-2.03	-2.55	0.38
		e^o +	e^r +	e^s	68	-3.38	-3.46	-3.29	-3.03	-3.44	-2.71	0.37
	Δe^c				6	-1.55	-1.38	-0.52	-0.65	-0.51	-0.47	0.38
e^c_{rr} +	Δe^c				0	-0.40	-1.33	0.54	-0.46	0.44	-0.49	0.35

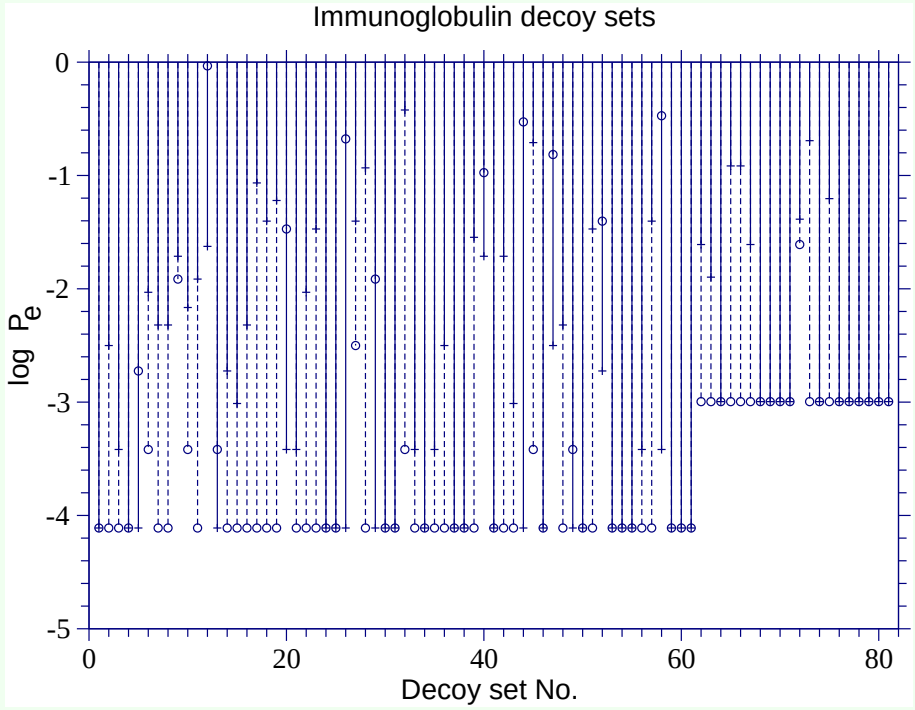
[a] The orientational energies used above are calculated with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025$.

[b] e^c_{rr}: collapse energy, $\Delta e^c_{aa'}$: contact energy between amino acids of type a and a', e^r: repulsive packing energy, e^s: secondary structure energy

The orientational potentials improve the performance for fold recognition in most decoy sets.



(A)



(B)

The dotted lines and open circles show the improvements of performance for each decoy set by the orientational potential.

(A) The potentials for monomeric protein decoy sets consist of $e^c_{rr} + \Delta e^c$ for cross marks and solid lines, and $e^c_{rr} + \Delta e^c + e^o$ for open circles and broken lines. (B) The potentials for immunoglobulin decoy sets consist of $\Delta e^c + e^r$ for cross marks and solid lines, and $e^o + e^r$ for open circles and broken lines. The orientational energies are evaluated with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025$.

Comparison of the performance of fold recognition between potentials

The present method outperforms other potentials including a CHARMM-based potential for most of the decoy families.

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

Decoy ID range, Decoy family Potentials	# tops /# total	mean $\overline{\log P_e}$	mean $\overline{Z_e}$	mean $\overline{R^1}$	
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

Decoy ID range, Decoy family Potentials	# tops /# total	mean $\overline{\log P_e}$	mean $\overline{Z_e}$	mean $\overline{R}$[1]	
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 mode
$(e^o + e^r + e^s)^2$	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

[a] R is the correlation coefficient of rank order between the energies and RMSDs of decoys in a decoy set.

[b] The present model; the orientational energies were calculated with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025$.

[c] Taken from Reference.

[d] The distance-dependent angular potential named "3C326" in Reference

[e] Generalized Born, Coulomb, non-polar solvation and van der Waals energy terms are included.

4. CONCLUSION

- The present results indicate that the present scheme of the corrections and cutoffs for expansion terms allows us to estimate orientational distributions in relatively high resolution.
- The residue-residue orientations significantly depends on Euler angles as well as polar angles, and the present orientational potentials have proved its effectiveness on fold recognition.
- The present potential function performs well in comparison with other scoring functions. The discrimination for the native structure is successful for 61 of 79 monomeric decoy sets and for 68 of 81 immunoglobulin decoy sets. Also, the mean Z-score Z_e in the energy scale which is equal to -4.45 for monomeric decoy sets and -3.29 for immunoglobulin decoy sets is statistically significant.
- All 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.