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

Sanzo Miyazawa

miyazawa@smlab.sci.gunma-u.ac.jp

Faculty of Technology, Gunma University

Kiryu, Gunma 376-8515, Japan

presented at

The Annual meeting in 2004 of the Biophysical Society of Japan

(December 14, 2004)

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. 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 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 orietational 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. Ref: J. Chem. Phys. (2004) in press.

1. INTRODUCTION

Most attempts to develop coarse-grained potentials are paiwise isotropic potentials.

Attempts to develop other types of potentials are limited.

One of difficulties to develop such a potential of mean force is that a method of dividing space into many cells and counting samples observed in each cell requires too many samples.

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

    Their results indicated that the discrimination power of potentials could not be improved by taking account of Euler angle dependencies.
  - ⋆ Buchete et al. (2003) and (2004) .

    Only radial and polar angle dependencies were taken into account.

Purposes of the present work:

- To evaluate dependences on polar (θ, ϕ) and Euler (Θ, Φ, Ψ) angles and correlations between them in residue-residue orientations; the residue-residue orientations significantly depends on Euler angles.
- To assess the effectiveness of a potential of mean force for residue-residue orientation on fold recognition; the orientational potential can improve the recognition power for the native folds, and the total energy potential taken as a simple sum of contact, orientation, and (ϕ, ψ) potentials performs well to identify the native folds.

Differences from other works:

- 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 orientaions.
- 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) .
  - ⋆ Each expansion coefficient is separately corrected for the sample size depending on the resolution of each term.
  - ⋆ Higher order terms are ignored to remove artificial contributions from the small size of samples.

2. METHODS

Coarse-grained conformational energy

$$E^{conf} = E^l + E^s = E^c + E^r + E^s \quad (1)$$

where

- E^l long-range interaction energy,
- E^s short-range interaction energy,
- E^c long-range residue-residue contact energy including orientational energies,
- E^r long-range repulsive packing energy that is a function of the excess number of contacting residues,
- E^s short-range secondary structure energy that is a backbone (ϕ, ψ) statistical potential here.

Statistical potentials previously estimated are used for the potentials above except for the orientational potential that is reported here.

Contact potentials

$$E^c \;=\; \frac{1}{2}\sum_i \sum_{j\neq i} e^c(r_i, r_j) \tag{2}$$

The contact energy, $e^c(r_i, r_j)$, between the ith and jth residues is defined as

$$e^c(r_i, r_j) \;=\; \Delta^c(r_i, r_j)\,[\, e^c_{a_i a_j} + e^o{}_{a_i a_j}(r_i, r_j)\,] \tag{3}$$

where

$\Delta^c(r_i, r_j)$	a switching function measuring 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,
$e^c_{a_i a_j}$	the contact energy for residues of type a_i and a_j in contact,
$e^o{}_{a_i a_j}(r_i, r_j)$	the orientational energy between amino acids of type a_i and a_j,
$r_i, \quad r_j$	positions of ith and jth residues.

Residue-residue orientational potentials between contacting residues

$$e^{o}{}_{aa'} = \frac{1}{2}\left[\{-\log f_{aa'} + <\log f_{aa'}>\} + \{-\log f_{a'a} + <\log f_{a'a}>\}\right] \quad (4)$$

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,
θ, ϕ	polar angles to specify two degrees of directional freedom for the orientation,
Θ, Φ, Ψ	Euler angles to specify three degrees of rotational freedom for the orientation,
$< -\log f_{aa'} >$	orientational entropy as a reference state which is the uniform distribution.

How to estimate the distribution of residue-residue orientations.

Expansion in spherical harmonics functions:

$$f_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) = \sum_{l_p=0}\sum_{m_p=-l_p}^{l_p}\sum_{l_e=0}\sum_{m_e=-l_e}^{l_e}\sum_{k_e} c^{aa'}_{l_p m_p l_e m_e k_e}\, g_{l_p m_p l_e m_e k_e}(\theta,\phi,\Theta,\Phi,\Psi) \tag{5}$$

g is represented as

$$g_{l_p m_p l_e m_e k_e} \equiv Y^{m_p}_{l_p}(\cos\theta,\phi) Y^{m_e}_{l_e}(\cos\Theta,\Phi) R_{k_e}(\Psi) \tag{6}$$

$$Y^m_l(\cos\theta,\phi) = [\frac{(2l+1)(l-|m|)!}{2(l+|m|)!}]^{1/2} P^{|m|}_l(\cos\theta) R_m(\phi) \tag{7}$$

$$R_m(\phi) = \begin{cases} \frac{1}{\sqrt{\pi}}\sin(m\phi) & \text{for } m>0 \\ \frac{1}{\sqrt{2\pi}} & \text{for } m=0 \\ \frac{1}{\sqrt{\pi}}\cos(m\phi) & \text{for } m<0 \end{cases} \tag{8}$$

where

Y^m_l the normalized spherical harmonics function,

$P^{|m_p|}_{l_p}$ the associated Legendre function.

The coefficients in the expansion of Eq. (5) can be calculated by

$$c^{aa'}_{l_p m_p l_e m_e k_e} = \int f_{aa'}\; g_{l_p m_p l_e m_e k_e}\; d\cos\theta\, d\phi\, d\cos\Theta\, d\Phi\, d\Psi \tag{9}$$

$$c^{aa'}_{00000} = \frac{1}{2(2\pi)^{3/2}} \tag{10}$$

from the observed density distribution:

$$f^{obs}_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) = \frac{1}{N_{aa'}} \sum_{\mu\in\{(aa')\}} w_\mu\, \delta(\cos\theta - \cos\theta_\mu)\, \delta(\phi-\phi_\mu)\, \delta(\cos\Theta - \cos\Theta_\mu)\, \delta(\Phi-\Phi_\mu)\, \delta(\Psi-\Psi_\mu) \tag{11}$$

$$N_{aa'} = \sum_{\mu\in\{(aa')\}} w_\mu \tag{12}$$

where

$(\theta_\mu,\phi_\mu,\Theta_\mu,\Phi_\mu,\Psi_\mu)$	a set of angles observed for the contact μ between residue types a and a',
w_μ	a weight for this contact μ,
μ	contacting residue pairs whose geometric centers of side chains are within $6.5\mathring{A}$,
$N_{aa'}$	the effective number of contacts (a,a').

The summations in the equations above are over all contacts of amino acid types a versus a'.

Each expansion coefficient is **separately** corrected for the sample size according to suggestions from a Bayesian statistical analysis.

$$c^{aa'}_{l_p m_p l_e m_e k_e} = \frac{1}{N_{aa'}} \sum_{\mu \in \{(aa')\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \tag{13}$$

$$\approx \frac{1}{1 + \beta^{aa'}_{l_p m_p l_e m_e k_e}} \left[\beta^{aa'}_{l_p m_p l_e m_e k_e} \, c^{ar}_{l_p m_p l_e m_e k_e} + \frac{1}{N_{aa'}} \sum_{\mu \in \{(aa')\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{14}$$

$$c^{ar}_{l_p m_p l_e m_e k_e} \approx \frac{1}{1 + \beta^{ar}_{l_p m_p l_e m_e k_e}} \left[\beta^{ar}_{l_p m_p l_e m_e k_e} \, c^{rr}_{l_p m_p l_e m_e k_e} + \frac{1}{N_{ar}} \sum_{\mu \in \{(ar)\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{15}$$

$$c^{rr}_{l_p m_p l_e m_e k_e} \approx \frac{1}{1 + \beta^{rr}_{00000}} \left[\beta^{rr}_{00000} \, c^{rr}_{00000} \, \delta_{0 l_p} \, \delta_{0 m_p} \, \delta_{0 l_e} \, \delta_{0 m_p} \, \delta_{0 k_e} + \frac{1}{N_{rr}} \sum_{\mu \in \{(rr)\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{16}$$

where r means any type of residues and $\beta^{aa'}_{l_p m_p l_e m_e k_e}$ is taken to be

$$\beta^{aa'}_{l_p m_p l_e m_e k_e} \equiv \frac{\beta \, O_{l_p m_p l_e m_e k_e}}{N_{aa'}} \tag{17}$$

$$O_{l_p m_p l_e m_e k_e} \equiv \text{(the number of frequency modes lower than or equal to } (l_p, m_p, l_e, m_e, k_e)\text{)}$$

$$= (l_p^2 + 2|m_p| + 1)(l_e^2 + 2|m_e| + 1)(2|k_e| + 1) \tag{18}$$

in order to reduce statistical errors resulting from small sample size; β in Eq. (17) is a parameter to be optimized.

Higher order terms are ignored to remove artificial contributions from the small size of samples, and also terms with the small values of coefficients are neglected to reduce the number of expansion terms.

$$f_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) \approx \sum_{l_p=0}^{l_p^{max}} \sum_{m_p=-l_p}^{l_p} \sum_{l_e=0}^{l_e^{max}} \sum_{m_e=-l_e}^{l_e} \sum_{k_e}^{k_e^{max}} H(O_{cutoff} - O_{l_p m_p l_e m_e k_e}) \\ H(|c^{aa'}_{l_p m_p l_e m_e k_e}| - c_{cutoff}\, c^{aa'}_{00000}) c^{aa'}_{l_p m_p l_e m_e k_e}\, g_{l_p m_p l_e m_e k_e}(\theta,\phi,\Theta,\Phi,\Psi) \quad (19)$$

where $H(x)$ is the Heaviside step function, so that the summation above is over $l_p \le l_p^{max},\ l_e \le l_e^{max},\ k_e \le k_e^{max}$ and $O_{l_p m_p l_e m_e k_e} \le O_{cutoff}$.

Repulsive potentials to prevent packing at overly high densities

$$E^r = \sum_i [\frac{1}{2} \sum_j \{e^{hc}(r_i, r_j) + e^{re}_{ij}\} + e^{rp}_i] \tag{20}$$

$$\tag{21}$$

where

Hard/soft core repulsive energy: $$e^{hc}(r_i, r_j) \equiv 10\ S_w(|r_i - r_j|, 2.2, 2.6) \tag{22}$$

Excess contact energy: $$e^{re}_{ij} = H(n^c_i - q^c_{a_i})\,[\,(\frac{q^c_{a_i}}{n^c_i} - 1)e^c(r_i, r_j)\,] \tag{23}$$

Repulsive packing energy: $$e^{rp}_i = H(n^c_i - q^c_{a_i})\,[\,-\log(\frac{N(a_i, n^c_i) + \epsilon}{N(a_i, q^c_{a_i}) + \epsilon})\,] \tag{24}$$

Total number of contacting residues: $$n^c_i = \sum_j \Delta^c(r_i, r_j) \tag{25}$$

where $S_w(x, a, b)$ is a switching function in $a \leq x \leq b$, H is the Heaviside step function, $q^c_{a_i}$ is the coordination number for the residue of type a_i, $N(a_i, n^c_i)$ is the observed number of occurencies for the type of residue a_i with the number of contacting residues n^c_i, and ϵ is a small value added to avoid the divergence of the logarithm function.

Short range secondary structure potentials

It is estimated by the sum of dihedral angle dependent energies of a main-chain;

$$E^s = \sum_i e^s_{a_i}(\phi_i, \psi_i) \tag{26}$$

$$e^s_a(\phi, \psi) \equiv -\log(N_a(\phi, \psi)/N_a) + < \log(N_a(\phi, \psi)/N_a) > \tag{27}$$

$$< -\log(N_a(\phi, \psi)/N_a) > = \frac{-1}{N_a} \sum_{(\phi,\psi)} N_a(\phi, \psi) \log(N_a(\phi, \psi)/N_a) \tag{28}$$

where

$N_a(\phi, \psi)$ the number of amino acids of type a at (ϕ, ψ) observed in protein native structures,

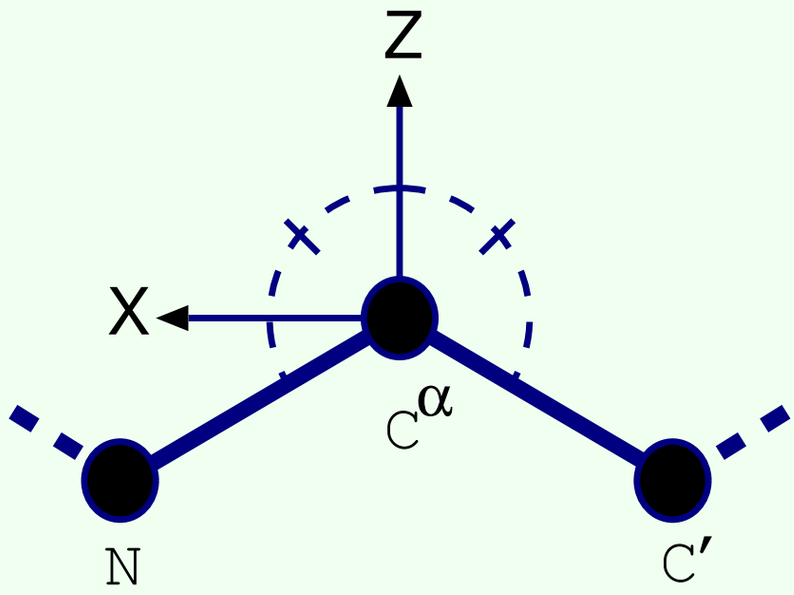
N_a the sum of $N_a(\phi, \psi)$ over the entire (ϕ, ψ) space.

Datasets of protein structures used to estimate the orientational potentials

- Proteins which belong to class 1 to 5 in Release 1.61 of the SCOP have been used.
- Only structures better than 2.5 Ådetermined 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

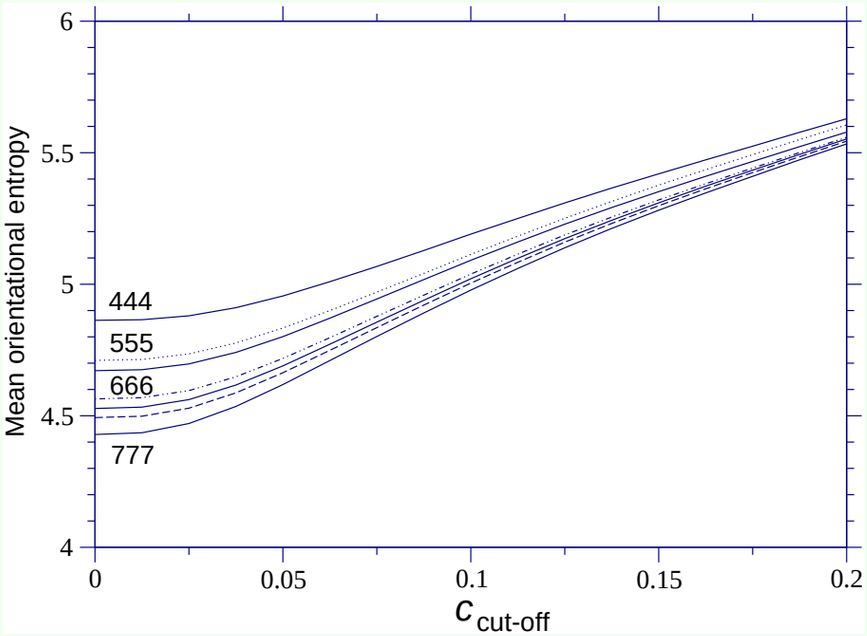
A local coordinate system affixed to each residue is based only on main chain atoms for fold recognition.

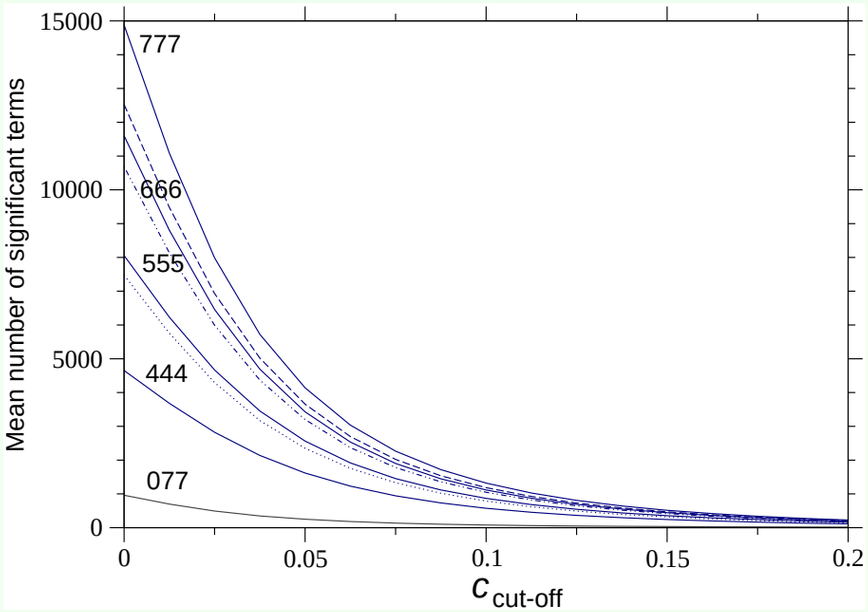


The origin O of the local coordinate system is located at the C^{α} position of each residue. The Y and Z axes are ones formed by the vector product and the sum of the unit vectors from N to C^{α} and from C' to C^{α}, respectively. The X axis is taken to form a right-handed coordinate system. The relative 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

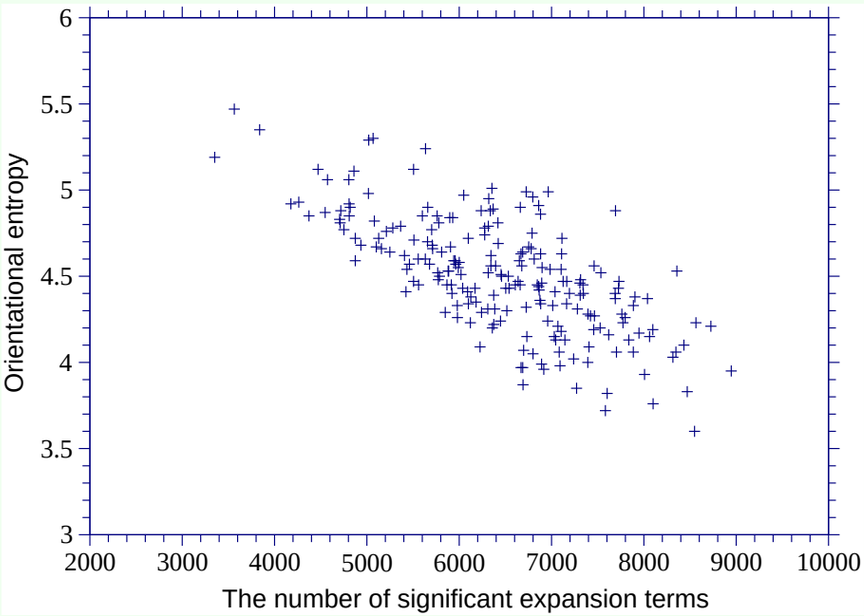
Dependencies of orientational entropies and the number of significant expansion terms on parameters

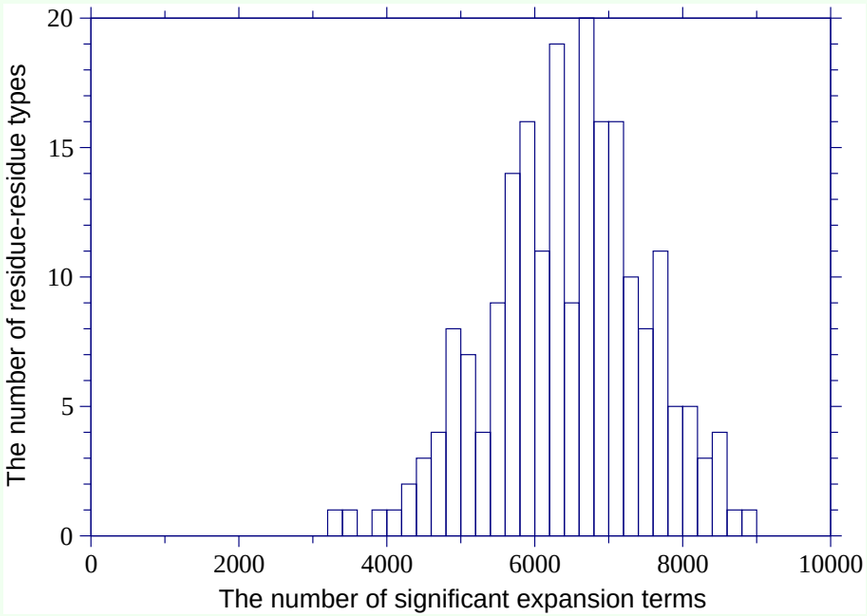




Triplets of digits near solid lines indicate the values of $(l_p^{max}, l_e^{max}, k_e^{max})$; for non-solid lines, $l_p^{max} = l_e^{max} = k_e^{max} = 6$ is used. The other parameters are: $\beta = 0.2$ for all lines, and $O_{cutoff} = O_{33333} = 1792$ for solid lines. The upper dotted line shows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

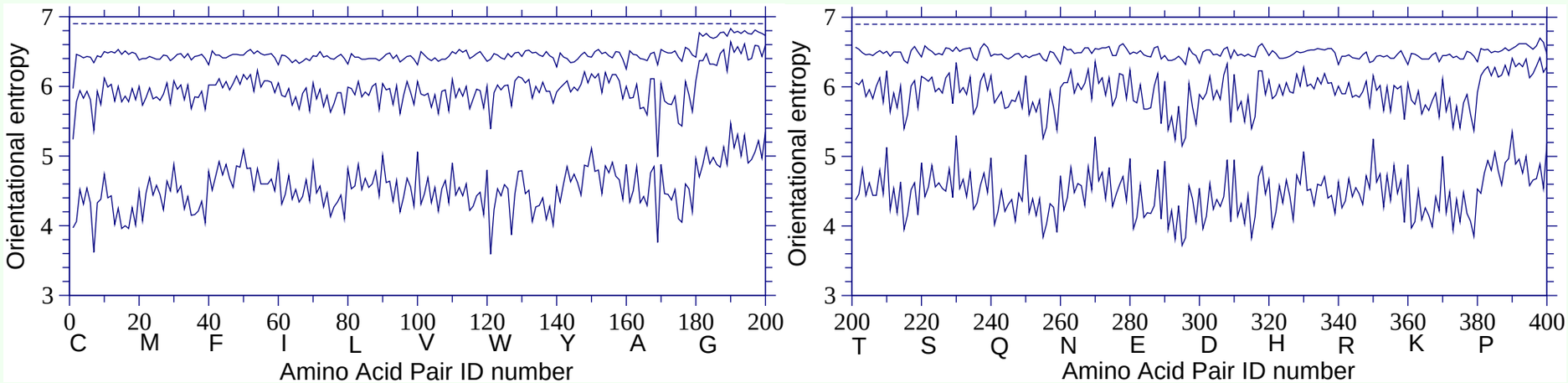




The orientational potentials are evaluated with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025.$

Distributions of residue orientations significantly depend on Euler 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 angles 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 angles 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 to 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 for performance:

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

$$P_e \equiv \text{the rank of the native fold in a energy scale / the number of decoys} \quad (29)$$

$$P_r \equiv \text{the rank of the lowest energy fold in the RMSD scale/ the number of decoys} \quad (30)$$

- **Z scores.**

$$Z_e \equiv \frac{E_{native} - \overline{E_{decoy}}}{\sigma_E} \quad (31)$$

$$Z_r \equiv Z_{rmsd} \equiv \frac{RMSD_{lowest} - \overline{RMSD_{decoy}}}{\sigma_{rmsd}} \quad (32)$$

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

(A) Dependencies on polar angles

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	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
4	0.0	23	-2.79	-2.09	-1.41	29	-2.66	-1.88	-1.45
5	0.0	31	-3.35	-2.57	-1.84	31	-2.68	-1.96	-1.46
6	0.0	27	-3.23	-2.55	-1.77	34	-2.69	-2.19	-1.45
7	0.0	30	-3.45	-2.60	-1.98	45	-2.93	-2.52	-1.57
8	0.0	28	-3.37	-2.59	-1.91	38	-2.73	-2.24	-1.48
9	0.0	25	-3.38	-2.43	-1.92	32	-2.66	-2.06	-1.54
10	0.0	27	-3.32	-2.55	-1.83	37	-2.55	-2.13	-1.52
11	0.0	28	-3.44	-2.67	-1.94	39	-2.68	-2.16	-1.71
12	0.0	25	-3.29	-2.45	-1.78	41	-2.70	-2.29	-1.76
13	0.0	30	-3.39	-2.73	-1.80	39	-2.80	-2.19	-1.83
14	0.0	31	-3.42	-2.89	-1.84	46	-2.87	-2.48	-1.91

(B) Dependencies on Euler angles

l_e^{max}		$l_p^{max}=0$, $\beta=0.2$, $O_{cutoff}=\infty$							
k_e^{max}	c_{cutoff}	79 monomeric decoy sets				81 lg decoy sets			
		#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
4	0.0	25	-3.18	-2.68	-1.78	33	-2.63	-2.26	-1.31
	0.025	25	-3.14	-2.71	-1.75	33	-2.61	-2.31	-1.29
5	0.0	25	-3.26	-2.79	-1.77	44	-2.85	-2.55	-1.65
	0.025	26	-3.23	-2.80	-1.74	44	-2.84	-2.58	-1.61
6	0.0	26	-3.25	-2.79	-1.83	47	-3.04	-2.78	-1.84
	0.025	24	-3.20	-2.57	-1.81	45	-3.00	-2.79	-1.77
7	0.0	30	-3.31	-2.84	-1.88	52	-3.03	-2.94	-1.82
	0.025	28	-3.24	-2.70	-1.83	52	-3.02	-2.92	-1.73

(A) Dependencies on l^{max} and cutoff O_{cutoff}

l_p^{max}	O_{cutoff}	$l_e^{max} = k_e^{max} = l_p^{max}$, $\beta = 0.2$, $c_{cutoff} = 0.025$							
		79 monomeric decoy sets				81 lg decoy sets			
		#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
4	960	34	-3.72	-3.24	-2.18	47	-2.97	-2.81	-1.59
	1792	36	-3.77	-3.27	-2.21	47	-3.01	-2.79	-1.67
5	960	36	-3.82	-3.38	-2.27	56	-3.18	-3.02	-1.81
	1792	38	-3.87	-3.22	-2.33	55	-3.23	-2.92	-1.96
6	960	37	-3.83	-3.33	-2.32	60	-3.24	-3.23	-1.92
	1792	37	-3.88	-3.22	-2.38	59	-3.27	-3.11	-2.00
	2025	38	-3.85	-3.25	-2.36	56	-3.21	-3.05	-1.99
7	64	27	-3.53	-2.95	-1.93	30	-2.63	-2.04	-1.46
	960	36	-3.85	-3.22	-2.34	57	-3.22	-3.11	-1.93
	1792	38	-3.91	-3.31	-2.42	53	-3.20	-2.94	-2.02
	2025	37	-3.87	-3.29	-2.40	54	-3.20	-3.02	-2.04

(B) Dependencies on cutoff c_{cutoff}

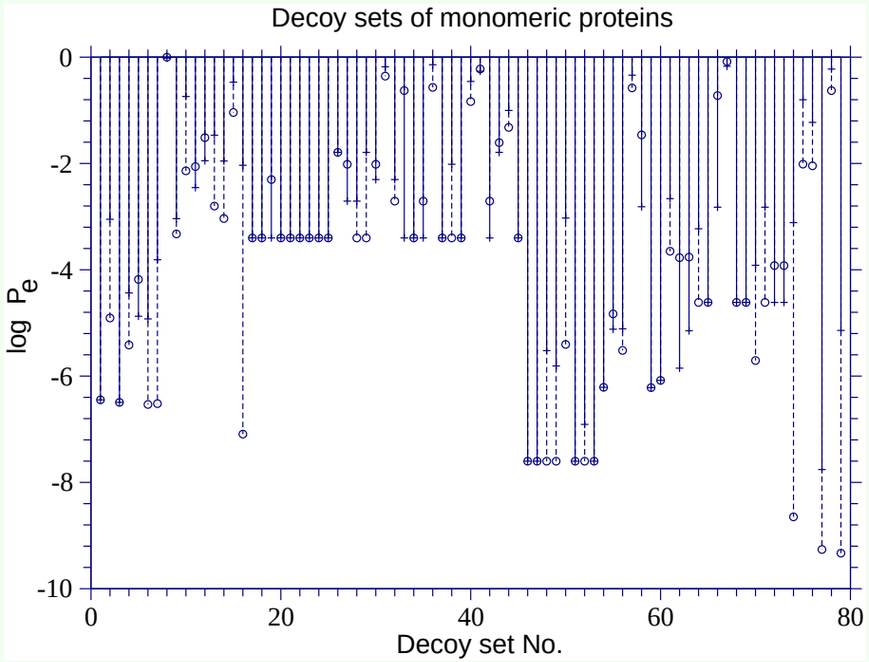
l_p^{max}	c_{cutoff}	$l_e^{max} = k_e^{max} = l_p^{max}$, $\beta = 0.2$, $O_{cutoff} = 960$							
		79 monomeric decoy sets				81 lg decoy sets			
		#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
5	0.0	35	-3.81	-3.33	-2.27	55	-3.17	-2.96	-1.83
	0.025	36	-3.82	-3.38	-2.27	56	-3.18	-3.02	-1.81
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
7	0.0	34	-3.82	-3.11	-2.33	59	-3.25	-3.17	-1.96
	0.025	36	-3.85	-3.22	-2.34	57	-3.22	-3.11	-1.93
l_p^{max}	c_{cutoff}	$l_e^{max} = k_e^{max} = l_p^{max}$, $\beta = 0.2$, $O_{cutoff} = 1792$							
5	0.0	38	-3.88	-3.30	-2.34	56	-3.23	-2.93	-1.96
	0.025	38	-3.87	-3.22	-2.33	55	-3.23	-2.92	-1.96
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
7	0.0	39	-3.92	-3.27	-2.43	55	-3.20	-3.05	-2.05
	0.025	38	-3.91	-3.31	-2.42	53	-3.20	-2.94	-2.02

(C) Dependencies on a parameter for small sample correction, β

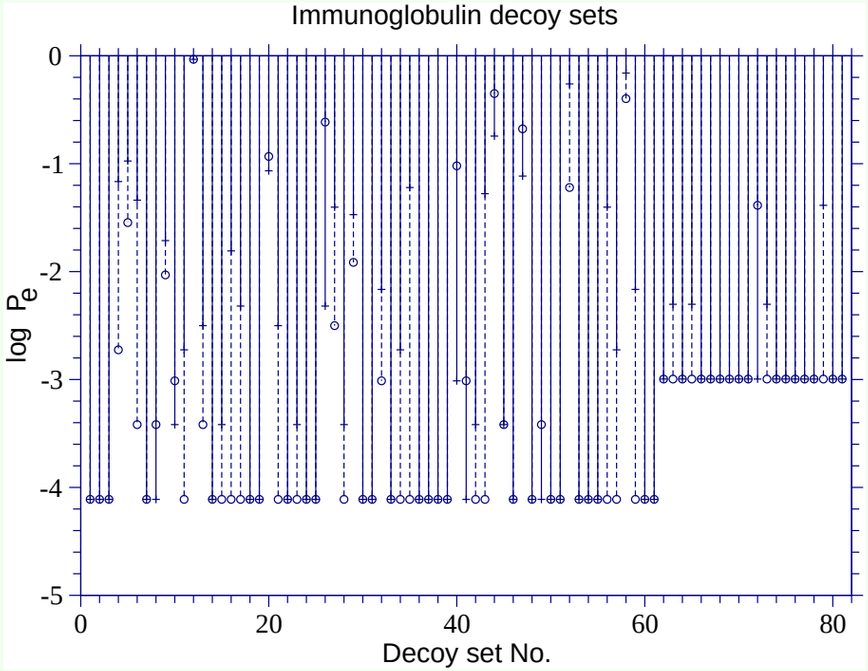
O_{cutoff}	β	$l_p^{max} = l_e^{max} = k_e^{max} = 6$, $c_{cutoff} = 0.025$							
		79 monomeric decoy sets				81 lg decoy sets			
		#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$	#tops	$\overline{\log P_e}$	$\overline{\log P_r}$	$\overline{Z_e}$
960	0.1	35	-3.82	-3.26	-2.32	60	-3.25	-3.23	-1.93
	0.2	37	-3.83	-3.33	-2.32	60	-3.24	-3.23	-1.92
	1	34	-3.78	-3.23	-2.28	58	-3.22	-3.19	-1.89
1792	0.1	36	-3.86	-3.15	-2.39	59	-3.27	-3.11	-2.00
	0.2	37	-3.88	-3.22	-2.38	59	-3.27	-3.11	-2.00
	1	36	-3.85	-3.18	-2.34	57	-3.24	-3.05	-1.97

The effects of Euler angle dependencies in the orientational potentials on the performance for fold recognition.

(A)



(B)



The dotted lines and open circles show the improvements of performance for each decoy set by taking account of Euler angle dependencies.

The potential function used here consists of orientational potentials e^o only. Cross marks and solid lines show the case for the orientational potential with $l_p^{max} = 7, l_e^{max} = k_e^{max} = 0, O_{cutoff} = \infty, c_{cutoff} = 0.025$. Open circles and broken lines show the case for the orientational potential with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, c_{cutoff} = 0.025$.

Performance of each potential component in fold recognition

All energy components 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}		Δe^c_{ij}		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}	$\overline{R^2}$
				e^o					37	-3.88	-3.22	-2.38	-2.49	-2.09	-1.65	0.33
				e^o	+	e^r			35	-3.79	-3.08	-2.32	-2.33	-2.01	-1.49	0.33
				e^o	+			e^s	53	-4.00	-3.99	-2.96	-3.13	-3.22	-2.59	0.35
				e^o	+	e^r	+	e^s	53	-3.98	-3.99	-2.93	-3.13	-3.16	-2.59	0.34
		Δe^c							36	-4.12	-3.20	-2.56	-2.12	-2.37	-1.63	0.33
		Δe^c	+			e^r			41	-3.90	-3.12	-2.23	-2.03	-2.04	-1.74	0.32
		Δe^c	+	e^o					52	-4.53	-4.24	-3.18	-3.19	-2.79	-2.60	0.37
		Δe^c	+	e^o	+	e^r			52	-4.38	-4.04	-2.95	-3.01	-2.54	-2.50	0.37
		Δe^c	+	e^o	+			e^s	58	-4.25	-4.30	-3.51	-3.38	-3.48	-3.04	0.37
		Δe^c	+	e^o	+	e^r	+	e^s	57	-4.15	-4.24	-3.35	-3.35	-3.17	-2.80	0.37
e^c_{rr}	+	Δe^c							36	-4.05	-3.29	-2.68	-2.32	-2.61	-1.86	0.32
e^c_{rr}	+	Δe^c	+			e^r			38	-4.18	-3.50	-2.53	-2.50	-2.49	-2.14	0.32
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^r			57	-4.73	-4.69	-4.13	-3.74	-3.76	-3.41	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
e^c_{rr}	+	Δe^c	+	e^o	+	e^r	+	e^s	59	-4.49	-4.49	-4.21	-3.56	-3.86	-3.10	0.39

(B) For the 81 immunogloblin decoy sets

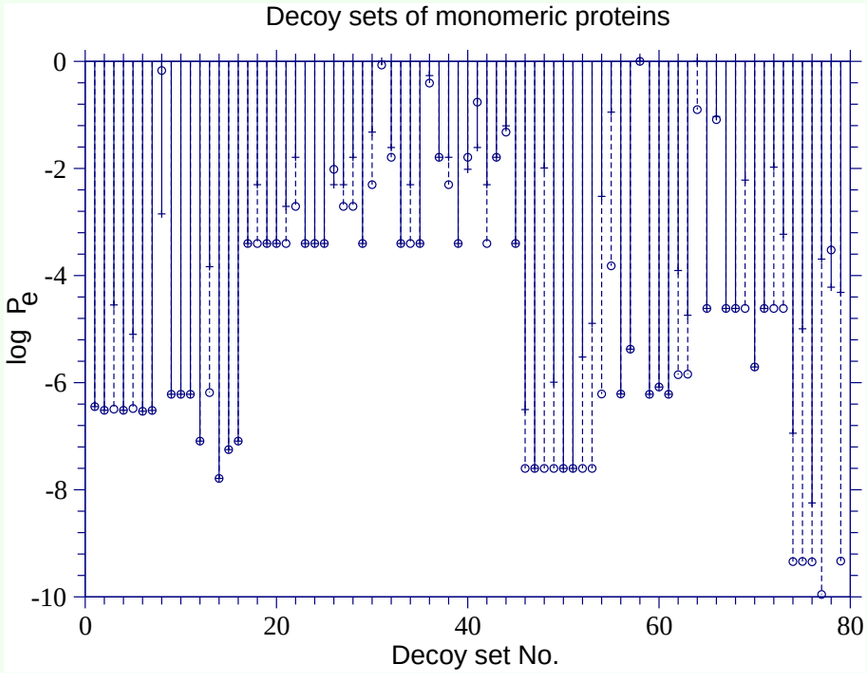
The true ground state for the contact potentials, e^c_{rr} and Δe^c_{ij}, requires all of the chains to be present.

Potentials[1]								# top ranks	mean	mean	mean	mean	median	median	mean
e^c_{rr}		Δe^c_{ij}		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}	$\overline{R^2}$
				e^o					59	-3.27	-3.11	-2.00	-2.74	-2.03	-2.55	0.38
				e^o	+	e^r			62	-3.35	-3.23	-2.15	-2.85	-2.27	-2.61	0.36
				e^o	+			e^s	67	-3.36	-3.42	-3.14	-3.00	-3.27	-2.69	0.39
				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
		Δe^c	+			e^r			36	-2.78	-2.29	-1.02	-1.70	-0.95	-1.15	0.29
		Δe^c	+	e^o					57	-3.20	-3.09	-1.57	-2.70	-1.55	-2.53	0.44
		Δe^c	+	e^o	+	e^r			63	-3.39	-3.35	-1.82	-2.95	-1.79	-2.67	0.40
		Δe^c	+	e^o	+			e^s	68	-3.36	-3.50	-2.53	-3.09	-2.44	-2.69	0.43
		Δe^c	+	e^o	+	e^r	+	e^s	69	-3.39	-3.52	-2.81	-3.09	-2.81	-2.71	0.40
e^c_{rr}	+	Δe^c							0	-0.40	-1.33	0.54	-0.46	0.44	-0.49	0.35
e^c_{rr}	+	Δe^c	+			e^r			0	-0.44	-1.29	0.35	-0.50	0.24	-0.49	0.32
e^c_{rr}	+	Δe^c	+	e^o					19	-2.11	-2.08	-0.86	-1.26	-0.89	-0.79	0.50
e^c_{rr}	+	Δe^c	+	e^o	+	e^r			44	-2.82	-2.81	-1.20	-2.22	-1.25	-2.13	0.48
e^c_{rr}	+	Δe^c	+	e^o	+			e^s	55	-3.00	-3.10	-1.83	-2.63	-1.94	-2.53	0.49
e^c_{rr}	+	Δe^c	+	e^o	+	e^r	+	e^s	61	-3.24	-3.31	-2.25	-2.82	-2.34	-2.61	0.46

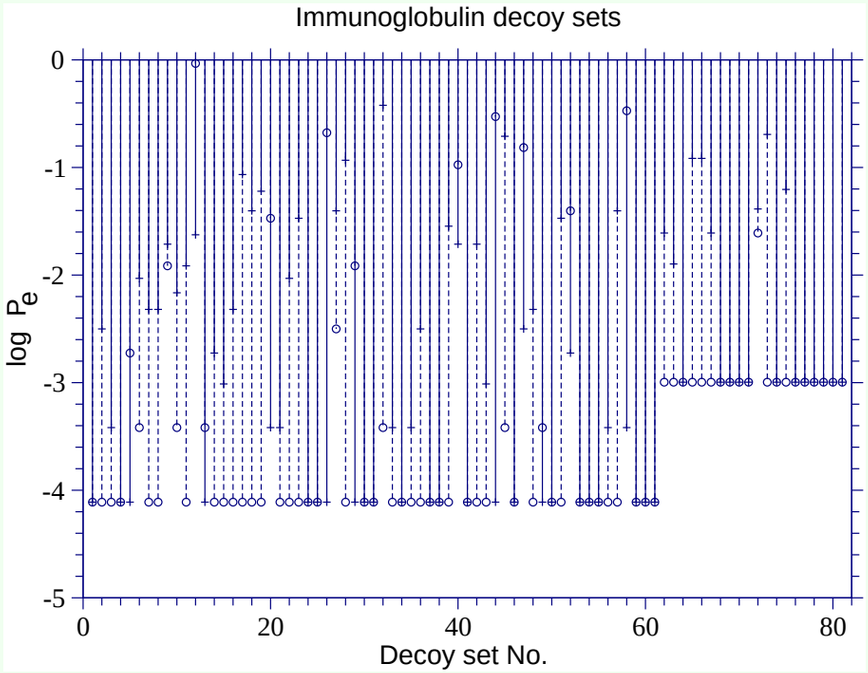
[a] The orientational energies used above are calculated with $l^{max}_p = l^{max}_e = k^{max}_e = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025$.

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 performance among potential functions for fold recognition

The present method outperforms the 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 modeli
$(e^o + e^r + e^s)^2$	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. DISCUSSION

- 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 results indicate that the present scheme of the corrections and cutoffs for expansion terms and for expansion coefficients allows us to estimate orientational distributions in relatively high resolution.
- 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.

Reference: J. Chem. Phys. (2004) in press.