# How effective in fold recognition is a potential of mean force for relative orientation 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

L.H. Baker Center for Bioinformatics and Biological Statistics

Iowa State University, Ames, IA 50011-3020, U.S.A.

(August 3, 2004)

# 1. INTRODUCTION

**Many attempts to develop coarse-grained potentials:**

Most of them are pairwise isotropic potentials:

- distant-dependent/contact pairwise potentials
- pairwise potentials at the atomic/residue level

**Methods:**

- a potential of mean force from residue distributions observed in protein structures

  The Boltzmann distribution is assumed for residue distributions.

- optimization of potentials to identify native folds

  The optimization methods indicate the limitation of pairwise isotropic potentials on fold recognition.

**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, 2004.

    Only radial and polar angle dependencies were taken into account.

## Purposes of the present work

- To evaluate dependences on polar and Euler angles and correlations between them in residue-residue orientations.
- To assess the effectiveness of a potential of mean force for residue-residue orientation on fold recognition.

## Differences from other works

- a method of estimating the distribution of residue-residue orientations

  We use the method proposed by Onizuka et al. (2002) in which the observed distribution of residue-residue orientations is represented as a sum of $\delta$ functions each of which represents the observed location in an angular space, and then is estimated in the form of a series expansion with spherical harmonics functions by ignoring high frequency modes that are subjected to the sample size.

- the evaluation of a potential from the distribution of residue-residue orientations

# 2. METHODS

## Coarse-grained conformational energy

$$E^{\text{conf}} = E^l + E^s = E^c + E^r + E^s \tag{1}$$

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

# Contact potentials

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

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

$$e^c(\vec{r}_i, \vec{r}_j) \;=\; \Delta^c(\vec{r}_i, \vec{r}_j)\,[\, e^c_{a_i a_j} + e^o{}_{a_i a_j}(\vec{r}_i, \vec{r}_j)\,] \tag{3}$$

where

| | |
|---|---|
| $\Delta^c(\vec{r}_i, \vec{r}_j)$ | 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 $i$th and $j$th residues. |
| $e^c_{a_i a_j}$ | the contact energy for residues of type $a_i$ and $a_j$ in contact |
| $e^o{}_{a_i a_j}(\vec{r}_i, \vec{r}_j)$ | the orientational energy between amino acids of type $a_i$ and $a_j$ |
| $\vec{r}_i, \vec{r}_j$ | positions of $i$th and $j$th residues |

# Residue-residue orientational potentials

$$e^{o}{}_{aa'} \;=\; \frac{1}{2}\,[\,\{-\log f_{aa'} + <\log f_{aa'}>\} + \{-\log f_{a'a} + <\log f_{a'a}>\}\,] \tag{4}$$

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 $a$ |
| $< -\log f_{aa'} >$ | orientational entropy |

In principle,

$$f_{aa'}(\theta, \phi, \Theta, \Phi, \Psi) \;=\; f_{a'a}(\theta', \phi', \Theta', \Phi', \Psi') \tag{5}$$

$$(\;\Theta' = -\Theta, \qquad \Phi' \;=\; -\Psi, \qquad \Psi' = -\Phi, \qquad \cdots\;) \tag{6}$$

but in the present statistical estimation,

$$f_{aa'}(\theta, \phi, \Theta, \Phi, \Psi) \;\simeq\; f_{a'a}(\theta', \phi', \Theta', \Phi', \Psi') \tag{7}$$

$$\tag{8}$$

# How to estimate the distribution of residue-residue orientations

$$f_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) = \sum_{l_p=0}\sum_{m_p=-l_p}^{l_p}\sum_{l_e=0}\sum_{m_e=-l_e}^{l_e}\sum_{k_e} c^{aa'}_{l_p m_p l_e m_e k_e}\, g_{l_p m_p l_e m_e k_e}(\theta,\phi,\Theta,\Phi,\Psi) \tag{9}$$

$g$ is represented as

$$g_{l_p m_p l_e m_e k_e} \equiv Y^{m_p}_{l_p}(\cos\theta,\phi) Y^{m_e}_{l_e}(\cos\Theta,\Phi) R_{k_e}(\Psi) \tag{10}$$

$$Y^m_l(\cos\theta,\phi) = [\frac{(2l+1)(l-|m|)!}{2(l+|m|)!}]^{1/2} P^{|m|}_l(\cos\theta) R_m(\phi) \tag{11}$$

$$\frac{1}{\sqrt{\pi}}\sin(m\phi) \quad \text{for } m>0 \tag{12}$$

$$R_m(\phi) = \frac{1}{\sqrt{2\pi}} \quad \text{for } m=0 \tag{13}$$

$$\frac{1}{\sqrt{\pi}}\cos(m\phi) \quad \text{for } m<0 \tag{14}$$

where

- $Y^m_l$ the normalized spherical harmonics function
- $P^{|m_p|}_{l_p}$ the associated Legendre function; $P^0_{l_p}$ with $m_p=0$ is the Legendre polynomial.

The coefficients in the expansion of Eq. (9) can be calculated by

$$c^{aa'}_{l_p m_p l_e m_e k_e} = \int f_{aa'}\; g_{l_p m_p l_e m_e k_e}\; d\cos\theta\, d\phi\, d\cos\Theta\, d\Phi\, d\Psi \tag{15}$$

$$c^{aa'}_{00000} = \frac{1}{2(2\pi)^{3/2}} \tag{16}$$

from the observed density distribution:

$$f^{\text{obs}}_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) = \frac{1}{N_{aa'}} \sum_{\mu\in\{(aa')\}} w_\mu\, \delta(\cos\theta - \cos\theta_\mu)\, \delta(\phi-\phi_\mu)\, \delta(\cos\Theta - \cos\Theta_\mu)\, \delta(\Phi-\Phi_\mu)\, \delta(\Psi-\Psi_\mu) \tag{17}$$

$$N_{aa'} = \sum_{\mu\in\{(aa')\}} w_\mu \tag{18}$$

where

| | |
|---|---|
| $(\theta_\mu,\phi_\mu,\Theta_\mu,\Phi_\mu,\Psi_\mu)$ | a set of angles observed for the contact $\mu$ between residue types $a$ and $a'$ |
| $w_\mu$ | a weight for this contact $\mu$ |
| contact $\mu$ | 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 done over all contacts of amino acid types $a$ versus $a'$.

# A correction for the small size of samples: pseudocounts are used according to the Bayesian statistical analysis

$$c^{aa'}_{l_p m_p l_e m_e k_e} = \frac{1}{N_{aa'}} \sum_{\mu \in \{(aa')\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \tag{19}$$

$$\overset{\text{est.}}{=} \frac{1}{1 + \beta^{aa'}_{l_p m_p l_e m_e k_e}} \left[ \beta^{aa'}_{l_p m_p l_e m_e k_e} \, c^{ar}_{l_p m_p l_e m_e k_e} + \frac{1}{N_{aa'}} \sum_{\mu \in \{(aa')\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{20}$$

$$c^{ar}_{l_p m_p l_e m_e k_e} \overset{\text{est.}}{=} \frac{1}{1 + \beta^{ar}_{l_p m_p l_e m_e k_e}} \left[ \beta^{ar}_{l_p m_p l_e m_e k_e} \, c^{rr}_{l_p m_p l_e m_e k_e} + \frac{1}{N_{ar}} \sum_{\mu \in \{(ar)\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{21}$$

$$c^{rr}_{l_p m_p l_e m_e k_e} \overset{\text{est.}}{=} \frac{1}{1 + \beta^{rr}_{00000}} \left[ \beta^{rr}_{00000} \, c^{rr}_{00000} \, \delta_{0 l_p} \, \delta_{0 m_p} \, \delta_{0 l_e} \, \delta_{0 m_p} \, \delta_{0 k_e} + \frac{1}{N_{rr}} \sum_{\mu \in \{(rr)\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{22}$$

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

$$\beta^{aa'}_{l_p m_p l_e m_e k_e} \equiv \frac{\beta \, O_{l_p m_p l_e m_e k_e}}{N_{aa'}} \tag{23}$$

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

$$= (l_p^2 + 2|m_p| + 1)(l_e^2 + 2|m_e| + 1)(2|k_e| + 1) \tag{24}$$

in order to reduce statistical errors resulting from small sample size; $\beta$ in Eq. (23) 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.**

$$f_{aa'}(\theta,\phi,\Theta,\Phi,\Psi) = \sum_{l_p=0}^{l_p^{\max}} \sum_{m_p=-l_p}^{l_p} \sum_{l_e=0}^{l_e^{\max}} \sum_{m_e=-l_e}^{l_e} \sum_{k_e}^{k_e^{\max}} H(O_{\text{cutoff}} + 1 - O_{l_p m_p l_e m_e k_e}) \\ H(c^{aa'}_{l_p m_p l_e m_e k_e} - c_{\text{cutoff}}\, c_{00000}) c^{aa'}_{l_p m_p l_e m_e k_e}\; g_{l_p m_p l_e m_e k_e}(\theta,\phi,\Theta,\Phi,\Psi) \tag{25}$$

where $H$ is the heaviside step function, and

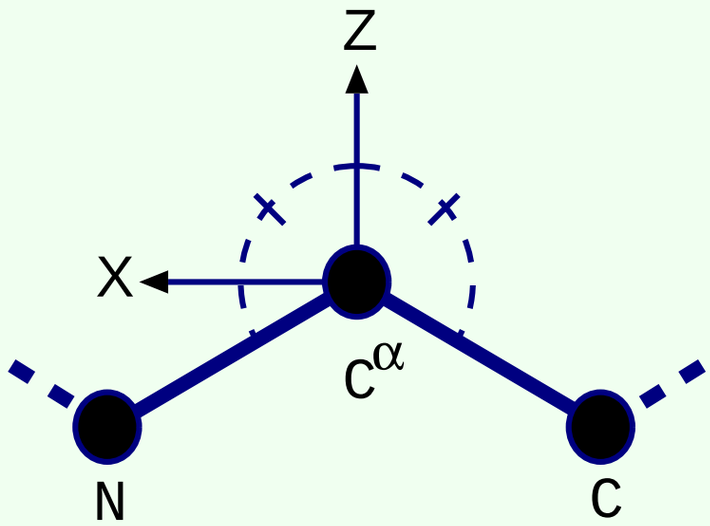
$$l_p \le l_p^{\max},\ l_e \le l_e^{\max},\ k_e \le k_e^{\max} \text{ and } O_{l_p m_p l_e m_e k_e} \le O_{\text{cutoff}} \tag{26}$$

## 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

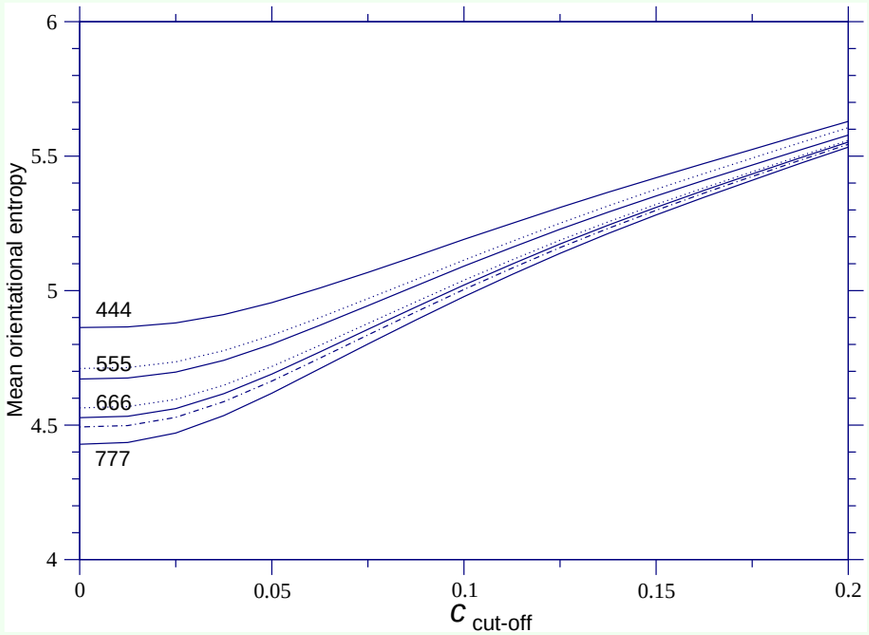
## A local reference frame for each residue

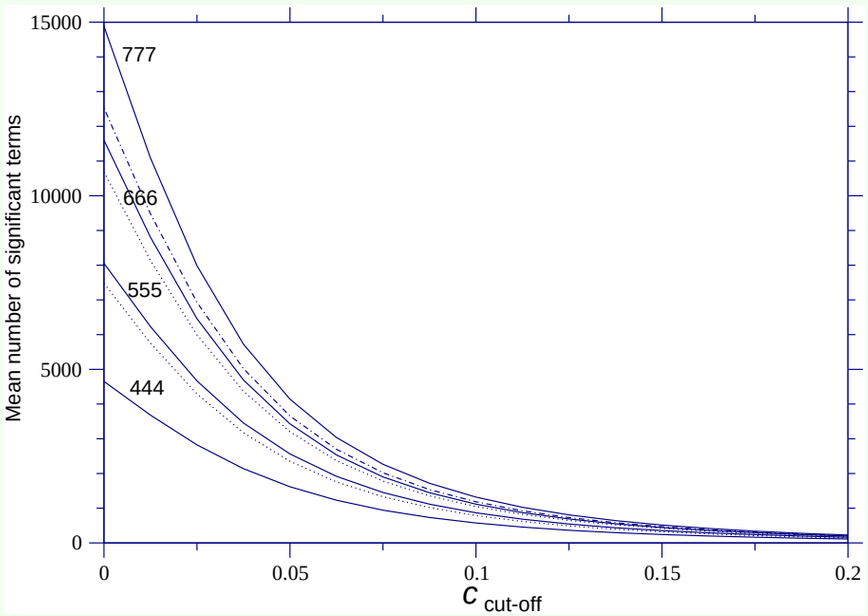



The definitions of a local coordinate system fixed at each residue. The origin $O$ of the local coordinate system is located at the $C_\alpha$ position of each residue. The $Y$ and $Z$ axes are ones formed by the vector product or the sum of the unit vectors from $N$ to $C_\alpha$ and from $C$ to $C_\alpha$, respectively. The $X$ axis is taken to form a right-handed coordinate system. Variables that describe the relative directions and rotations of contacting residues are polar angles $(\theta, \phi)$ and Euler angles $(\Theta, \Phi, \Psi)$, respectively.

# Orientational distributions of contacting residues

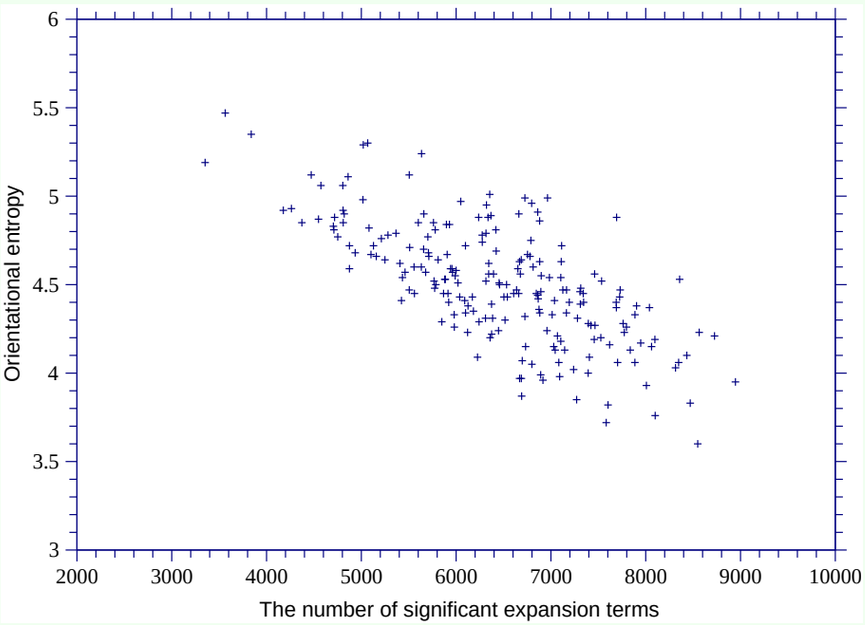
## Dependencies of orientational entropies and the number of significant expansion terms on parameters

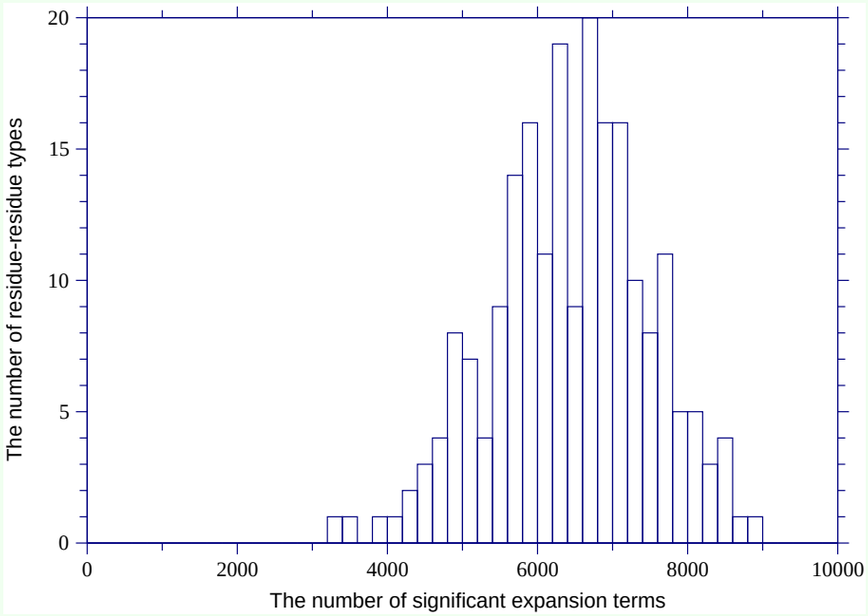



Triplets of digits near solid lines indicate the values of $(l_p^{\max}, l_e^{\max}, k_e^{\max})$; for the dotted lines, $l_p^{\max} = l_e^{\max} = k_e^{\max} = 6$ is used. The other parameters are: $\beta = 0.2$ for all lines, and $O_{\text{cutoff}} = O_{33333} = 1792$ for solid lines. $O_{\text{cutoff}} = 960, 1584,$ and $2025$, for dotted lines.

## 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.

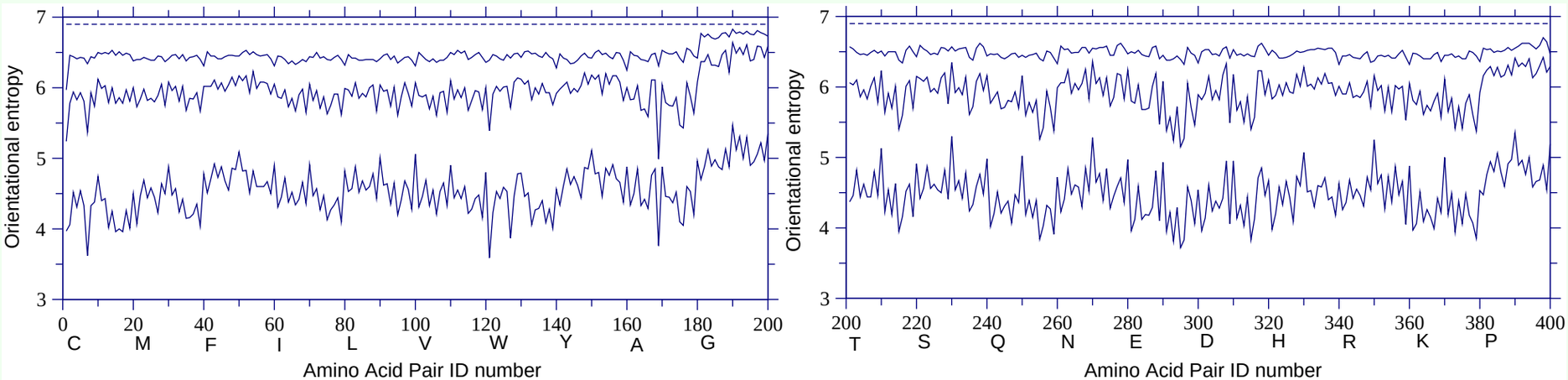



The orientational potentials are evaluated with $l_p^{\max} = l_e^{\max} = k_e^{\max} = 6, O_{\text{cutoff}} = 1792, \beta = 0.2, c_{\text{cutoff}} = 0.025.$

# Distributions of residue orientations significantly depend on Euler angles

## Orientational entropies for three types of distributions




The broken line shows the entropy, $6.900$, for a uniform distribution. The lowest solid line shows the distribution with polar and Euler angle dependencies, $l_p^{\max} = l_e^{\max} = k_e^{\max} = 6$; $c_{\text{cutoff}} = 0.025$. The highest solid line shows the distribution with $l_p^{\max} = 6, l_e^{\max} = k_e^{\max} = 0$ that depends on polar angles only; $c_{\text{cutoff}} = 0.0$. The middle solid line shows the distribution that depends on polar angles with $l_p^{\max} = 6$, and on Euler angles with $l_e^{\max} = k_e^{\max} = 6$, but ignores any correlation between polar and Euler angles; $c_{\text{cutoff}} = 0.0$. The values of other parameters are $O_{\text{cutoff}} = 1792$ and $\beta = 0.2$. The abscissa indicates the amino acid pair identification number; amino acid types are numbered in the order of amino acids written along the abscissa.

## Recognition power for native structures

The performance of the orientational 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.

# Evaluation of the performance of potential functions in fold recognition

## Measures

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

$$P_e \equiv \text{the rank of the native fold in a energy scale } / \text{ the number of decoys} \quad (34)$$

$$P_r \equiv \text{the rank of the lowest energy fold in the RMSD scale/ the number of decoys} \quad (35)$$

- **Z scores**

$$Z_e \equiv \frac{E_{\text{native}} - \overline{E_{\text{decoy}}}}{\sigma_E} \quad (36)$$

$$Z_r \equiv Z_{rmsd} \equiv \frac{\text{RMSD}_{\text{lowest}} - \overline{\text{RMSD}_{\text{decoy}}}}{\sigma_{rmsd}} \quad (37)$$

# How important are the Euler angle dependencies of relative residue orientations in fold recognition ?

## (A) Dependencies on polar angles

| | | $l_e^{\max} = k_e^{\max} = 0$, $\beta = 0.2$, $O_{\text{cutoff}} = \infty$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $l_p^{\max}$ | $c_{\text{cutoff}}$ | 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 | 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 | -1.99 | 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 | -1.81 | 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

| $l_e^{\max}$ | | $l_p^{\max} = 0$, $\beta = 0.2$, $O_{\text{cutoff}} = \infty$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k_e^{\max}$ | $c_{\text{cutoff}}$ | 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 | -1.76 | 33 | -2.61 | -2.31 | -1.29 |
| | 0.050 | 24 | -3.13 | -2.74 | -1.73 | 24 | -2.45 | -1.94 | -1.27 |
| 5 | 0.0 | 26 | -3.27 | -2.80 | -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 |
| | 0.050 | 24 | -3.16 | -2.70 | -1.70 | 34 | -2.65 | -2.24 | -1.49 |
| 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 |
| | 0.050 | 21 | -3.13 | -2.41 | -1.76 | 42 | -2.88 | -2.61 | -1.68 |
| 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 |
| | 0.050 | 25 | -3.18 | -2.51 | -1.75 | 47 | -2.88 | -2.77 | -1.60 |

## (A) Dependencies on $l^{\max}$ and cutoff $O_{\text{cutoff}}$

$l_e^{\max} = k_e^{\max} = l_p^{\max}$, $\beta = 0.2$, $c_{\text{cutoff}} = 0.025$

| $l_p^{\max}$ | $O_{\text{cutoff}}$ | 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 | 960 | 34 | -3.72 | -3.24 | -2.18 | 47 | -2.97 | -2.81 | -1.59 |
| | 1584 | 37 | -3.75 | -3.29 | -2.20 | 46 | -3.00 | -2.70 | -1.64 |
| | 1792 | 36 | -3.77 | -3.27 | -2.21 | 47 | -3.01 | -2.79 | -1.67 |
| 5 | 637 | 35 | -3.78 | -3.29 | -2.23 | 53 | -3.10 | -2.93 | -1.74 |
| | 960 | 36 | -3.82 | -3.38 | -2.27 | 56 | -3.18 | -3.02 | -1.81 |
| | 1584 | 37 | -3.85 | -3.25 | -2.32 | 55 | -3.21 | -2.91 | -1.93 |
| | 1792 | 38 | -3.87 | -3.22 | -2.33 | 55 | -3.23 | -2.92 | -1.96 |
| | 2025 | 37 | -3.87 | -3.27 | -2.36 | 56 | -3.23 | -2.96 | -1.95 |
| 6 | 637 | 37 | -3.86 | -3.31 | -2.32 | 55 | -3.15 | -3.04 | -1.83 |
| | 960 | 37 | -3.83 | -3.33 | -2.32 | 60 | -3.24 | -3.23 | -1.92 |
| | 1584 | 38 | -3.87 | -3.34 | -2.38 | 58 | -3.23 | -3.08 | -1.98 |
| | 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 | 637 | 37 | -3.90 | -3.31 | -2.37 | 51 | -3.12 | -2.90 | -1.85 |
| | 960 | 36 | -3.85 | -3.22 | -2.34 | 57 | -3.22 | -3.11 | -1.93 |
| | 1584 | 36 | -3.88 | -3.35 | -2.38 | 54 | -3.14 | -3.00 | -1.96 |
| | 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 $c_{\text{cutoff}}$

$l_e^{\max} = k_e^{\max} = l_p^{\max}$, $\beta = 0.2$, $O_{\text{cutoff}} = 960$

| $l_p^{\max}$ | $c_{\text{cutoff}}$ | 79 monomeric decoy sets | | | | 81 lg decoy sets | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | #tops | $\overline{\log P_e}$ | $\overline{\log P_r}$ | $\overline{Z_e}$ | #tops | $\overline{\log P_e}$ | $\overline{\log P_r}$ | $\overline{Z_e}$ |
| 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 |
| | 0.050 | 35 | -3.79 | -3.18 | -2.24 | 55 | -3.15 | -2.96 | -1.74 |
| 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 |
| | 0.050 | 33 | -3.77 | -2.99 | -2.27 | 56 | -3.19 | -3.12 | -1.82 |
| 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 |
| | 0.050 | 33 | -3.76 | -2.98 | -2.29 | 58 | -3.21 | -3.21 | -1.85 |

$l_e^{\max} = k_e^{\max} = l_p^{\max}$, $\beta = 0.2$, $O_{\text{cutoff}} = 1792$

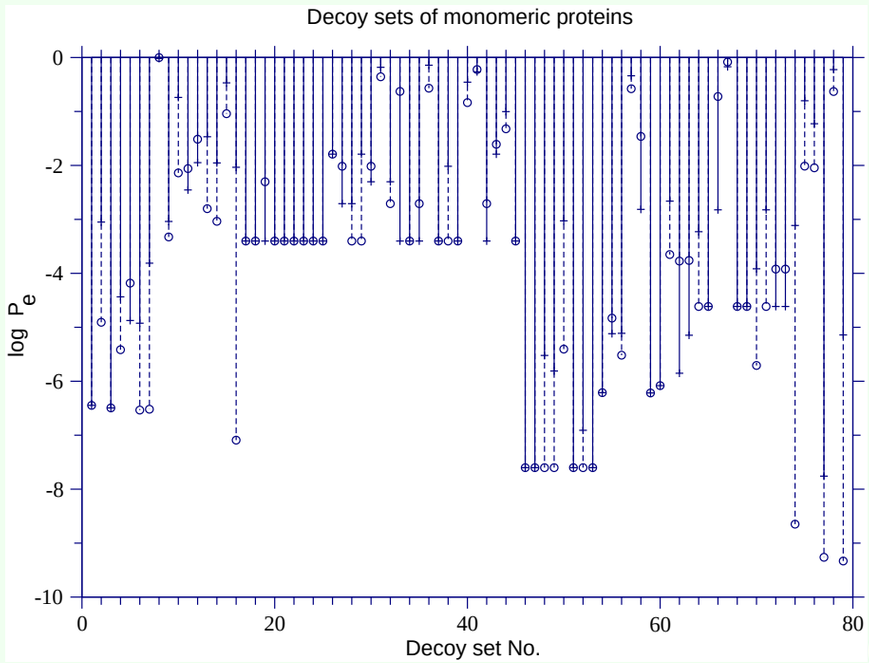
| $l_p^{\max}$ | $c_{\text{cutoff}}$ | 79 monomeric decoy sets | | | | 81 lg decoy sets | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | #tops | $\overline{\log P_e}$ | $\overline{\log P_r}$ | $\overline{Z_e}$ | #tops | $\overline{\log P_e}$ | $\overline{\log P_r}$ | $\overline{Z_e}$ |
| 5 | 0.0 | 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 |
| | 0.050 | 36 | -3.84 | -3.28 | -2.29 | 52 | -3.16 | -2.90 | -1.86 |
| 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 |
| | 0.050 | 35 | -3.85 | -3.19 | -2.36 | 56 | -3.21 | -3.09 | -1.89 |
| 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, $\beta$

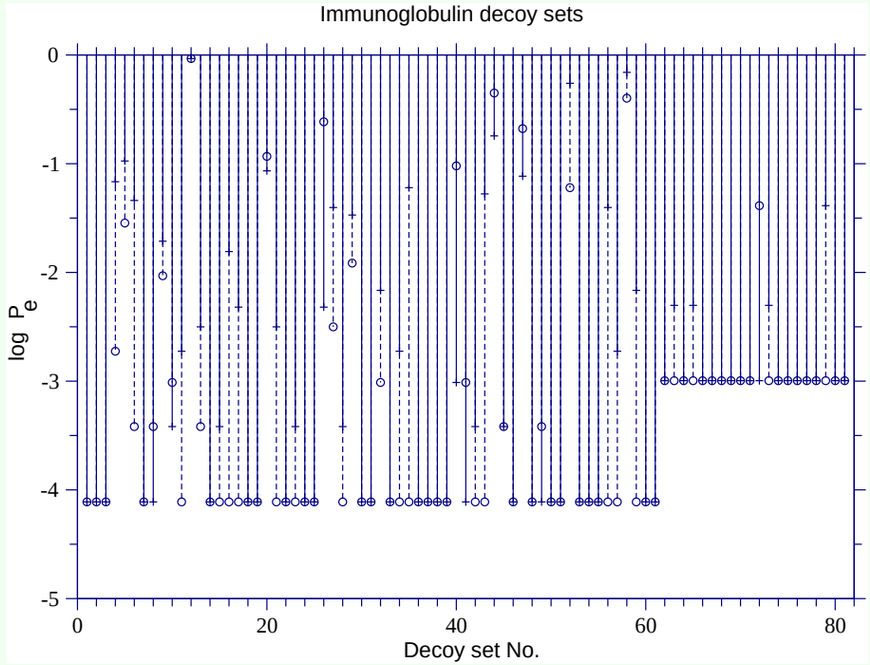
| $O_{\text{cutoff}}$ | $\beta$ | $l_p^{\max} = l_e^{\max} = k_e^{\max} = 6$, $c_{\text{cutoff}} = 0.025$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 79 monomeric decoy sets | | | | 81 lg decoy sets | | | |
| | | #tops | $\overline{\log P_e}$ | $\overline{\log P_r}$ | $\overline{Z_e}$ | #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 |
| | 0.5 | 35 | -3.80 | -3.24 | -2.30 | 58 | -3.23 | -3.20 | -1.91 |
| | 1 | 34 | -3.78 | -3.23 | -2.28 | 58 | -3.22 | -3.19 | -1.89 |
| 1584 | 0.1 | 38 | -3.87 | -3.34 | -2.38 | 58 | -3.23 | -3.08 | -1.97 |
| | 0.2 | 38 | -3.87 | -3.34 | -2.38 | 58 | -3.23 | -3.08 | -1.98 |
| | 0.5 | 38 | -3.86 | -3.32 | -2.36 | 59 | -3.24 | -3.14 | -1.97 |
| | 1 | 34 | -3.82 | -3.23 | -2.34 | 56 | -3.20 | -3.02 | -1.95 |
| 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 |
| | 0.5 | 36 | -3.86 | -3.22 | -2.37 | 60 | -3.27 | -3.15 | -1.99 |
| | 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 potential function used here consists of orientational potentials $e^o$ only. Cross marks and solid lines show the case for the orientational potential with $l_p^{\max} = 7, l_e^{\max} = k_e^{\max} = 0, O_{\text{cutoff}} = \infty, c_{\text{cutoff}} = 0.025$. Open circles and broken lines show the case for the orientational potential with $l_p^{\max} = l_e^{\max} = k_e^{\max} = 6, O_{\text{cutoff}} = 1792, c_{\text{cutoff}} = 0.025$.

# How important are relative orientations between residues in fold recognition ?

## (A) For 79 monomeric decoy sets

| Potentials[1] | | | | | | | | # top ranks | mean | mean | mean | mean | median | median | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e^c_{rr}$ | | $\Delta e^c_{ij}$ | | $e^o$ | | $e^r$ | | $e^s$ | # total = 79 | $\overline{\log P_e{}^2}$ | $\overline{\log P_r{}^2}$ | $\overline{Z_e}^3$ | $\overline{Z_{rmsd}}^3$ | $Z_e{}^3$ | $Z_{rmsd}{}^3$ | $\overline{R}^4$ |
| | | | | $e^o$ | | | | | 37 | -3.88 | -3.22 | -2.38 | -2.49 | -2.09 | -1.65 | 0.33 |
| | | | | $e^o$ | + | $e^r$ | | | 35 | -3.78 | -3.09 | -2.31 | -2.35 | -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 | -3.97 | -3.99 | -2.93 | -3.13 | -3.16 | -2.59 | 0.34 |
| | | $\Delta e^c$ | | | | | | | 36 | -4.12 | -3.20 | -2.56 | -2.12 | -2.37 | -1.63 | 0.33 |
| | | $\Delta e^c$ | + | | | $e^r$ | | | 41 | -3.88 | -3.17 | -2.22 | -2.09 | -2.04 | -1.75 | 0.32 |
| | | $\Delta e^c$ | + | $e^o$ | | | | | 52 | -4.53 | -4.24 | -3.18 | -3.19 | -2.79 | -2.60 | 0.37 |
| | | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | | | 52 | -4.37 | -4.04 | -2.94 | -3.01 | -2.54 | -2.50 | 0.37 |
| | | $\Delta e^c$ | + | $e^o$ | + | | | $e^s$ | 58 | -4.25 | -4.30 | -3.51 | -3.38 | -3.49 | -3.04 | 0.37 |
| | | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | + | $e^s$ | 57 | -4.15 | -4.20 | -3.35 | -3.32 | -3.17 | -2.80 | 0.37 |
| $e^c_{rr}$ | + | $\Delta e^c$ | | | | | | | 36 | -4.05 | -3.29 | -2.68 | -2.32 | -2.61 | -1.86 | 0.32 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | | | $e^r$ | | | 38 | -4.15 | -3.47 | -2.52 | -2.49 | -2.48 | -2.09 | 0.32 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | | | | | 58 | -4.79 | -4.88 | -4.38 | -3.92 | -4.08 | -3.55 | 0.40 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | | | 57 | -4.73 | -4.69 | -4.13 | -3.74 | -3.76 | -3.41 | 0.40 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | + | | | $e^s$ | 61 | -4.63 | -4.63 | -4.45 | -3.68 | -4.11 | -3.41 | 0.39 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | + | $e^s$ | 59 | -4.48 | -4.49 | -4.21 | -3.56 | -3.86 | -3.10 | 0.39 |

## (B) For 81 immunogloblin decoy sets

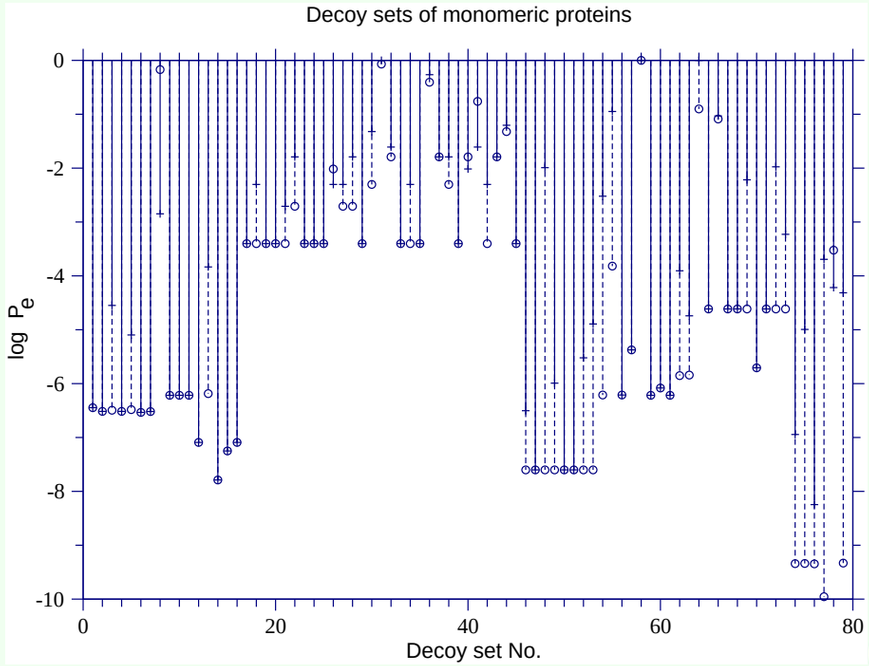
| Potentials[1] | | | | | | | | # top ranks | mean | mean | mean | mean | median | median | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e^c_{rr}$ | | $\Delta e^c_{ij}$ | | $e^o$ | | $e^r$ | | $e^s$ | # total = 81 | $\overline{\log P_e}$ | $\overline{\log P_r}$ | $\overline{Z_e}$ | $\overline{Z_{rmsd}}$ | ${Z_e}^3$ | $Z_{rmsd}$ | $\overline{R}^2$ |
| | | | | $e^o$ | | | | | 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 | -2.26 | -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 | -3.43 | -2.71 | 0.37 |
| | | $\Delta e^c$ | | | | | | | 6 | -1.55 | -1.38 | -0.52 | -0.65 | -0.51 | -0.47 | 0.38 |
| | | $\Delta e^c$ | + | | | $e^r$ | | | 35 | -2.76 | -2.22 | -1.01 | -1.66 | -0.96 | -0.94 | 0.29 |
| | | $\Delta e^c$ | + | $e^o$ | | | | | 57 | -3.20 | -3.09 | -1.57 | -2.70 | -1.55 | -2.53 | 0.44 |
| | | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | | | 63 | -3.39 | -3.35 | -1.82 | -2.95 | -1.79 | -2.67 | 0.40 |
| | | $\Delta e^c$ | + | $e^o$ | + | | | $e^s$ | 68 | -3.36 | -3.50 | -2.53 | -3.09 | -2.44 | -2.69 | 0.43 |
| | | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | + | $e^s$ | 69 | -3.39 | -3.52 | -2.81 | -3.09 | -2.81 | -2.71 | 0.40 |
| $e^c_{rr}$ | + | $\Delta e^c$ | | | | | | | 0 | -0.40 | -1.33 | 0.54 | -0.46 | 0.44 | -0.49 | 0.35 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | | | $e^r$ | | | 0 | -0.43 | -1.29 | 0.36 | -0.50 | 0.25 | -0.49 | 0.32 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | | | | | 19 | -2.11 | -2.08 | -0.86 | -1.26 | -0.89 | -0.79 | 0.50 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | | | 44 | -2.82 | -2.81 | -1.20 | -2.22 | -1.24 | -2.13 | 0.48 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | + | | | $e^s$ | 55 | -3.00 | -3.10 | -1.83 | -2.63 | -1.94 | -2.53 | 0.49 |
| $e^c_{rr}$ | + | $\Delta e^c$ | + | $e^o$ | + | $e^r$ | + | $e^s$ | 61 | -3.24 | -3.31 | -2.25 | -2.82 | -2.34 | -2.61 | 0.46 |

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

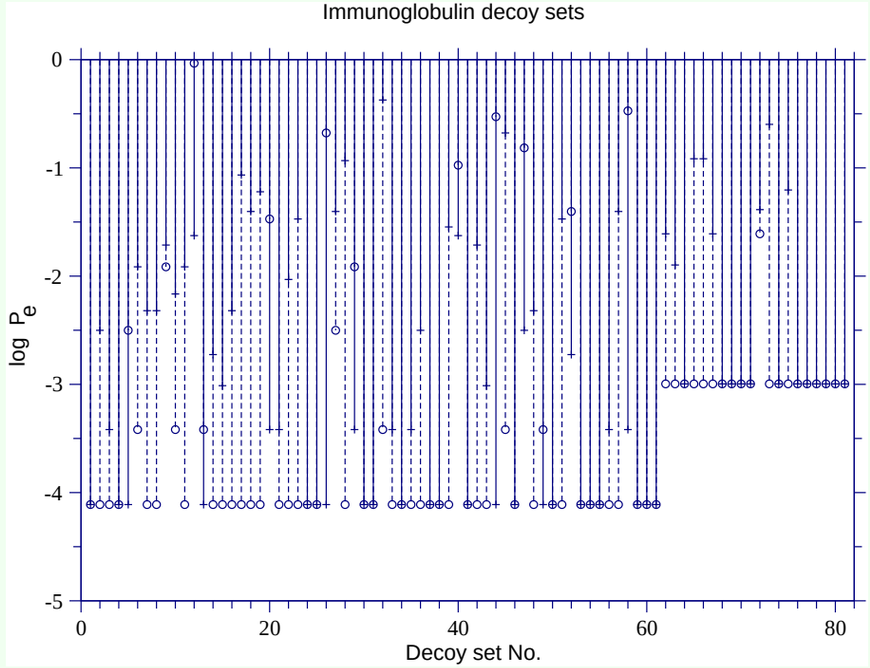
[b] $R$ is the correlation coefficient of rank order between the energies and RMSDs of decoys in a decoy set.

# The effects of the orientational potentials on performance for fold recognition.

(A)




(B)




The effects of the orientational potentials on performance for fold recognition. (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 $l_p^{\max} = l_e^{\max} = k_e^{\max} = 6, O_{\text{cutoff}} = 1792, \beta = 0.2, c_{\text{cutoff}} = 0.025$,

# 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.