# 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
Kiryu, Gunma 376-8515, Japan

presented at ”Structural Bioinformatics”
GSIC, Tokyo Institute of Technology
(April 5, 2006)

# 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.
- . . .

## What are statistical potentials ?

**statistical potential $\equiv -\log$ odds + const**

$$E(x) \equiv -\log \frac{P(x)}{P_0(x)} + \text{constant} \tag{1}$$

where

$P(x)$ is an observed probability distribution of quantity $x$ in protein native structures.

$P_0(x)$ is a reference probability distribution of quantity $x$

$\Longrightarrow$ **They represent the degree of native-like characteristics in protein conformations.**

$\Longrightarrow$ **A condition for statistical potentials to be independent:** $\frac{P(x,y)}{P_0(x,y)} = \frac{P(x)P(y)}{P_0(x)P_0(y)}$

# Can statistical potentials be regarded as physical potentials?

Basic assumptions for statistical potentials to be regarded as physical potentials:

1. In frequency distributions such as residue-residue contacts observed in a large number of protein crystal structures,
    - the effects of specific sequences and chain connectivities are averaged out, and therefore they can be regarded as equilibrium distributions under residue-residue interactions.

    $\Longrightarrow$ a statistical potential = a potential of mean force

    Objections: Thomas & Dill (1996); Ben-Naim (1997).

    Supports: Mirny & Shakhnovich (1996); Zhang (1998).

2. A potential of mean force in this many body system can reflect a physical interaction, and can be used to be approximately additive.

    Potential of mean force, $F(x)$: $\exp(-F(x)) \equiv \int \exp(-\beta E(x,\ldots,q_i,\ldots)) \prod_i dq_i$

    Objections: Ben-Naim (1997).

# Statistical potentials reflect physical interactions.

## An underlying principle: a principle of consistency/minimally frustration among variuos interactions

- The consistency among interactions is an effective way for proteins to increase structural stabilities. (Go, 1983)
- The energy landscape for natural proteins must be minimally frustrated for proteins to fold into single stable structures within a limited time. (Bryngelson & Wolynes, 1987)

  ### Supporting evidences:

  1. Secondary structures can be predicted to a certain extent based on short-range interactions only.
  2. Side-chain conformations can be predicted to a certain extent without taking account of sidechain-sidechain interactions. (Tanimura et al., 1994)
  3. Long- and short-range interactions in native structures are consistent/minimally frustrated in sequence space. (*Proteins*, **50**, 35-43, 2003)

## Observed distributions in a large number of structures can be approximately regarded as equilibrium distributions.

1. Observed $(\phi, \psi)$ distributions in protein native conformations resemble their Boltzmann distributions calculated from atomic potentials.

   (Pohl, *Nature New Biol.*, **234**, 277, 1971; Némethy & Scheraga, *Quart. Rev. Biophys.*, **10**, 239, 1977)
2. The contact frequencies of residues over proteins can be regarded with small relative errors ($< 10$ %) as contact frequencies at statistical equilibrium in residue mixtures. (*Proteins*, **34**, 49-68, 1999)

# 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 $(\theta, \phi)$ and Euler $(\Theta, \Phi, \Psi)$ 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 $\delta$ 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:

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

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

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

where

| | |
|---|---|
| $\vec{r}_i, \quad \vec{r}_j$ | positions of $i$th and $j$th 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 $i$th and $j$th 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

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

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 $a$, |
| $f_{a'a}(\theta', \phi', \Theta', \Phi', \Psi')$ | 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:

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

$g$ 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**

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

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

**from the observed density function:**

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

$$N_{aa'} = \sum_{\mu \in \{a\text{-}a'\}} w_\mu \tag{12}$$

where

| | |
|---|---|
| $(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu)$ | a set of angles observed for the contact $\mu$ |
| $w_\mu$ | a sampling weight for this contact $\mu$, |
| $\mu$ | 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.

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

$$\approx \frac{1}{1 + \beta^{aa'}_{l_p m_p l_e m_e k_e}} \left[ \frac{\beta^{aa'}_{l_p m_p l_e m_e k_e}}{2} (c^{ar}_{l_p m_p l_e m_e k_e} + c^{ra}_{l_p m_p l_e m_e k_e}) + \frac{1}{N_{aa'}} \sum_{\mu \in \{a\text{-}a'\}} w_\mu \, g_{l_p m_p l_e m_e k_e}(\theta_\mu, \phi_\mu, \Theta_\mu, \Phi_\mu, \Psi_\mu) \right] \tag{14}$$

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

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

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

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

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

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

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

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

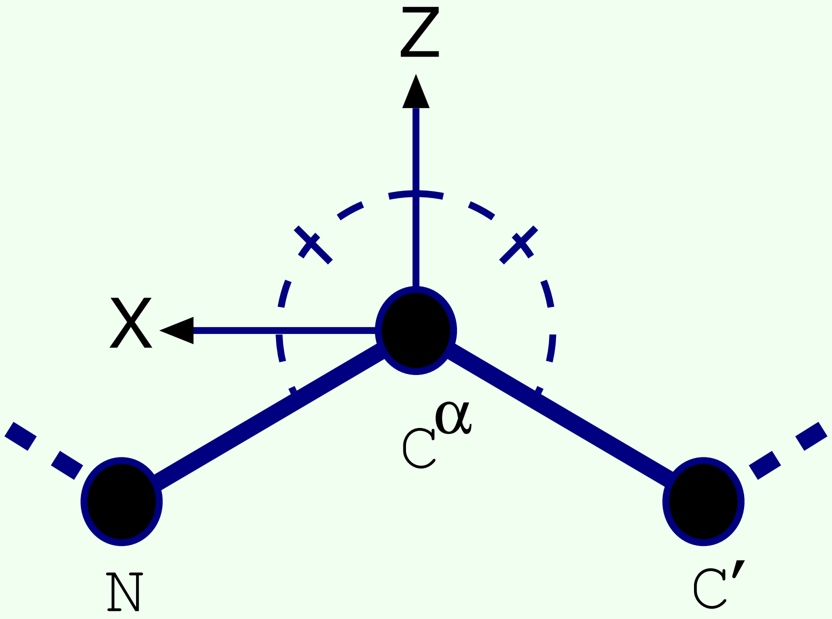
where $H(x)$ is the Heaviside step function, so that the summation above is over $l_p \le l_p^{max},\ l_e \le l_e^{max},\ |k_e| \le k_e^{max}$ and $O_{l_pm_pl_em_ek_e} \le O_{cutoff}$.

## 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

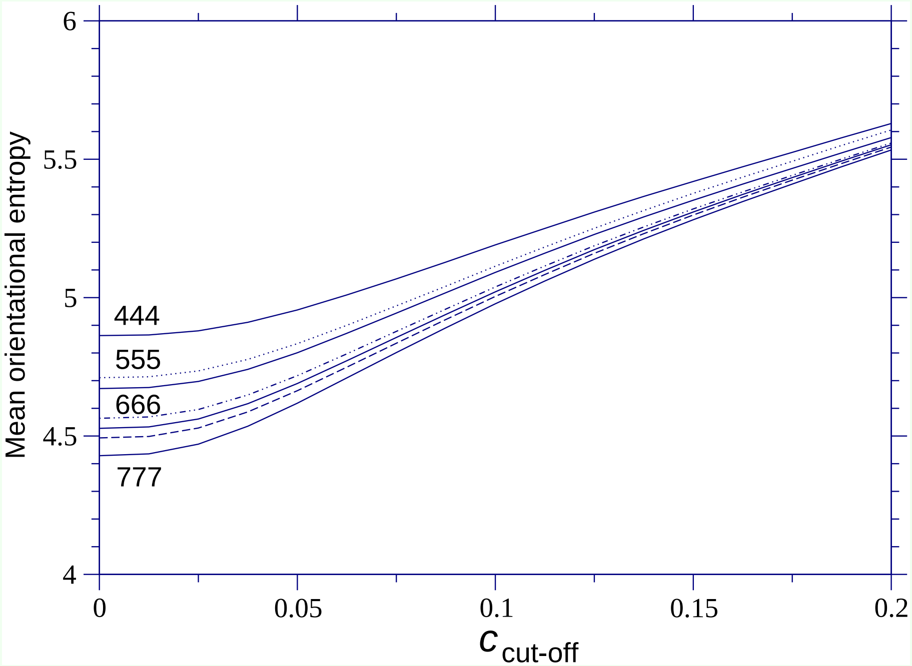
## A local coordinate system affixed to each residue is based only on main chain atoms for fold recognition.

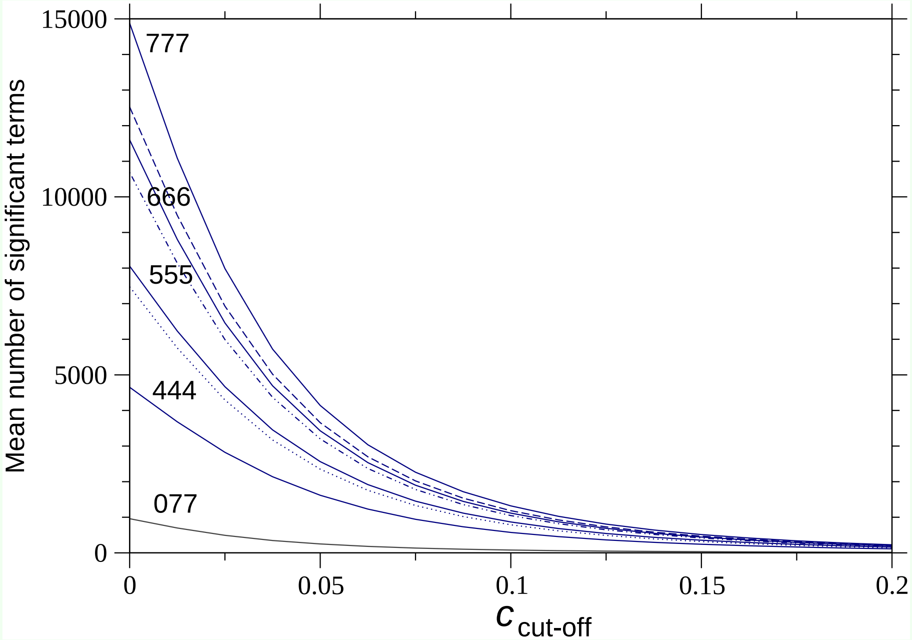


The origin $O$ of the local coordinate system is located at the $C^\alpha$ position of each residue. The $Y$ and $Z$ axes are ones formed by the vector product and the sum of the unit vectors from $N$ to $C^\alpha$ and from $C'$ to $C^\alpha$, respectively. The $X$ axis is taken to form a right-handed coordinate system. The relative direction and rotation of one residue to the other in contacting residues are represented by polar angles $(\theta, \phi)$ and Euler angles $(\Theta, \Phi, \Psi)$, respectively.

# Orientational distributions of contacting residues

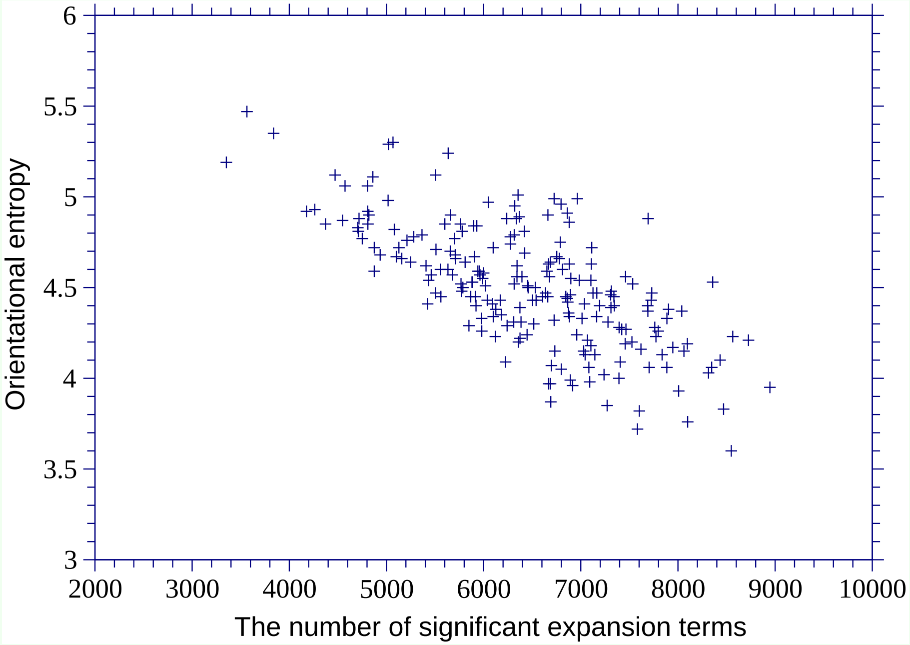
## Dependencies of orientational entropies and the number of significant expansion terms on parameters

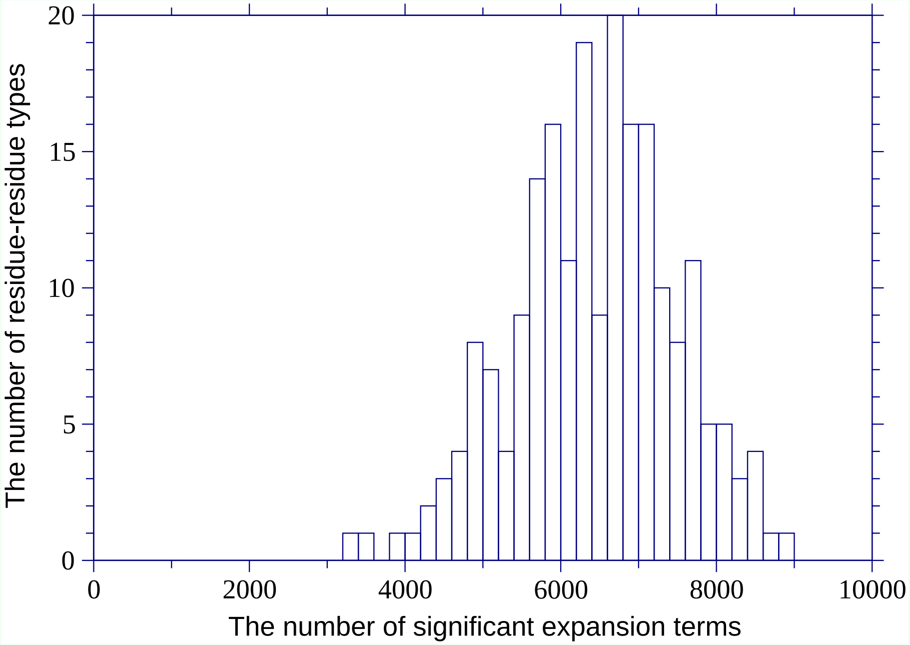




Triplets of digits near solid lines indicate the values of $(l_p^{max}, l_e^{max}, k_e^{max})$; for non-solid lines, $l_p^{max} = l_e^{max} = k_e^{max} = 6$ is used. The other parameters are: $\beta = 0.2$ for all lines, and $O_{cutoff} = O_{33333} = 1792$ for solid lines. The upper dotted line shows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

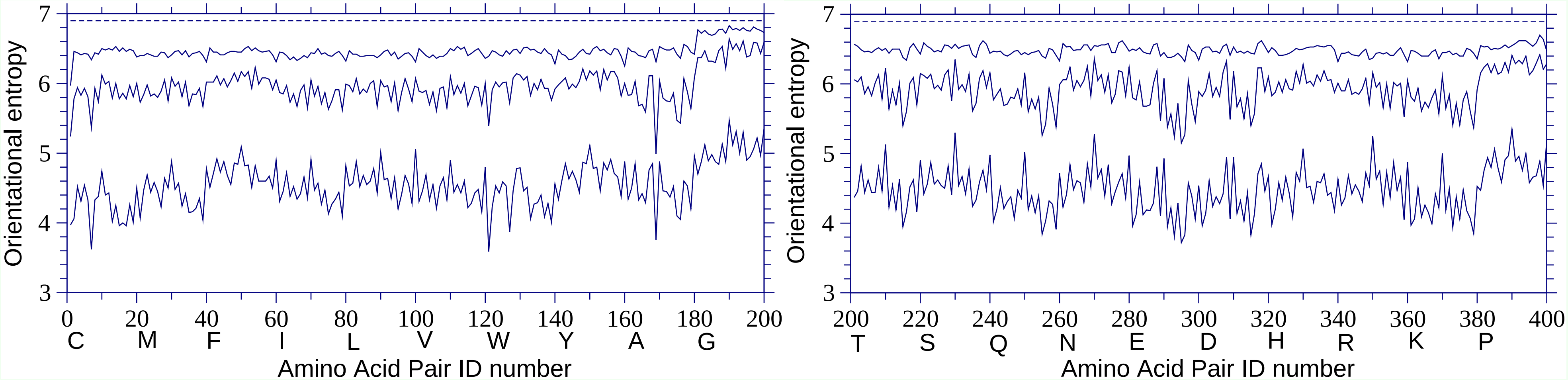




The orientational potentials are evaluated with $l_p^{max} = l_e^{max} = k_e^{max} = 6, O_{cutoff} = 1792, \beta = 0.2, c_{cutoff} = 0.025.$

# 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.**

$$P_e \equiv \text{the rank of the native fold in a energy scale} \;/\; \text{the number of decoys} \tag{20}$$

$$P_r \equiv \text{the rank of the lowest energy fold in the RMSD scale}/ \;\text{the number of decoys} \tag{21}$$

- **Z scores.**

$$Z_e \equiv \frac{E_{native} - \overline{E_{decoy}}}{\sigma_E} \tag{22}$$

$$Z_r \equiv Z_{rmsd} \equiv \frac{RMSD_{lowest} - \overline{RMSD_{decoy}}}{\sigma_{rmsd}} \tag{23}$$

# 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 Ig 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 Ig 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 |

# 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e^c_{rr}$ | | $\Delta e^c_{aa'}$ | | $e^o$ | | $e^r$ | | $e^s$ | # 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 |
| | | | | $e^o$ | $+$ | $e^r$ | | | 35 | -3.79 | -3.08 | -2.32 | -2.33 | -2.01 | -1.49 | 0.33 |
| | | $\Delta e^c$ | | | | | | | 36 | -4.12 | -3.20 | -2.56 | -2.12 | -2.37 | -1.63 | 0.33 |
| $e^c_{rr}$ | $+$ | $\Delta e^c$ | | | | | | | 36 | -4.05 | -3.29 | -2.68 | -2.32 | -2.61 | -1.86 | 0.32 |
| | | $\Delta e^c$ | $+$ | $e^o$ | | | | | 52 | -4.53 | -4.24 | -3.18 | -3.19 | -2.79 | -2.60 | 0.37 |
| $e^c_{rr}$ | $+$ | $\Delta e^c$ | $+$ | $e^o$ | | | | | 58 | -4.79 | -4.88 | -4.38 | -3.92 | -4.08 | -3.55 | 0.40 |
| $e^c_{rr}$ | $+$ | $\Delta e^c$ | $+$ | $e^o$ | $+$ | | | $e^s$ | 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 | -4.49 | -4.49 | -4.21 | -3.56 | -3.86 | -3.10 | 0.39 |

---

[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) For the 81 immunogloblin decoy sets

The true ground state for the contact potentials, $e^c_{rr}$ and $\Delta e^c_{ij}$, 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 |
| | | | $e^o$ | $+$ | $e^r$ | | | 62 | -3.35 | -3.23 | -2.15 | -2.85 | -2.27 | -2.61 | 0.36 |
| | | | $e^o$ | $+$ | $e^r$ | $+$ | $e^s$ | 68 | -3.38 | -3.46 | -3.29 | -3.03 | -3.44 | -2.71 | 0.37 |
| | | $\Delta e^c$ | | | | | | 6 | -1.55 | -1.38 | -0.52 | -0.65 | -0.51 | -0.47 | 0.38 |
| $e^c_{rr}$ | $+$ | $\Delta e^c$ | | | | | | 0 | -0.40 | -1.33 | 0.54 | -0.46 | 0.44 | -0.49 | 0.35 |

---

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

# 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 |
| | $\Delta e^c$ + | $e^o$ | | | 52 | -4.53 | -4.24 | -3.18 | -3.19 | -2.79 | -2.60 | 0.37 |
| $e^c_{rr}$ + | $\Delta e^c$ + | $e^o$ | | | 58 | -4.79 | -4.88 | -4.38 | -3.92 | -4.08 | -3.55 | 0.40 |
| $e^c_{rr}$ + | $\Delta e^c$ + | $e^o$ + | | $e^s$ | 61 | -4.63 | -4.63 | -4.45 | -3.68 | -4.11 | -3.41 | 0.39 |

## (B) For the 81 immunogloblin decoy sets

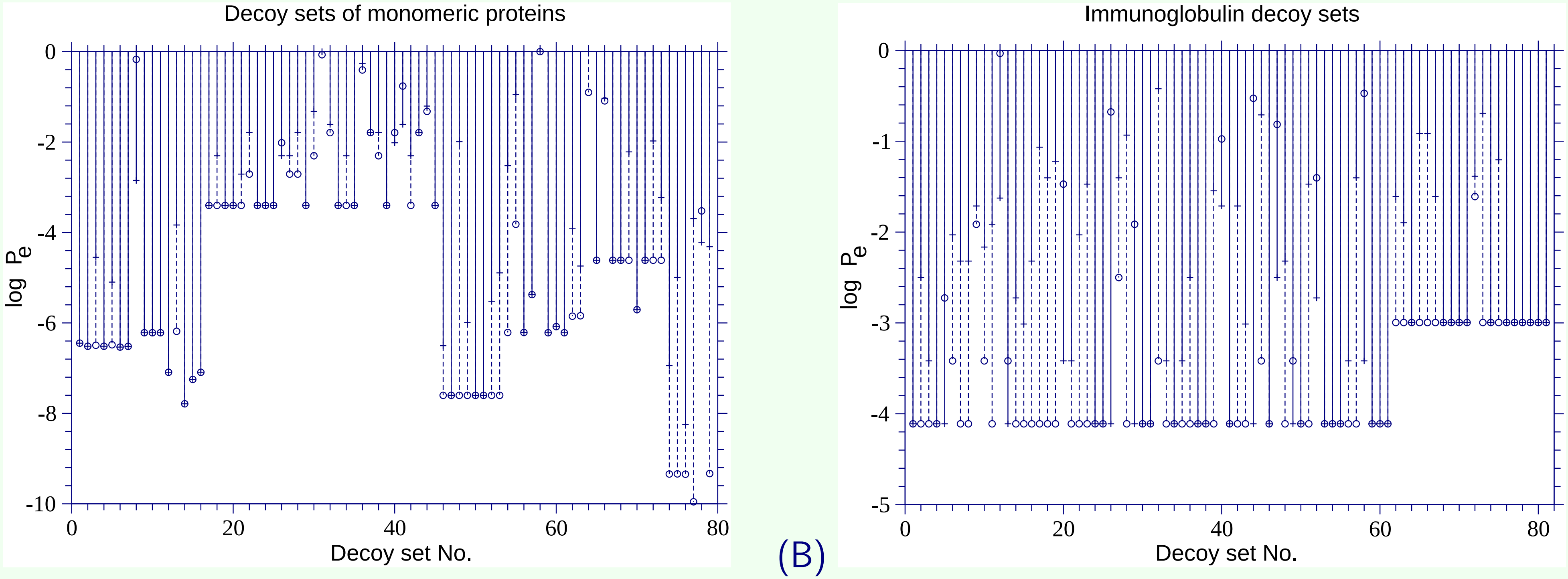
The true ground state for the contact potentials, $e^c_{rr}$ and $\Delta e^c_{ij}$, requires all of the chains to be present.

| 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 |
| | $\Delta e^c$ | | | | 6 | -1.55 | -1.38 | -0.52 | -0.65 | -0.51 | -0.47 | 0.38 |
| $e^c_{rr}$ + | $\Delta e^c$ | | | | 0 | -0.40 | -1.33 | 0.54 | -0.46 | 0.44 | -0.49 | 0.35 |

---

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

# 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] | ∼ 7/7 | ∼ -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 | # tops | mean | mean | mean | |
|---|---|---|---|---|---|
| Potentials | /# total | $\overline{\log P_e}$ | $\overline{Z_e}$ | $\overline{R}^1$ | |
| 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<br>Potentials | # tops<br>/# total | mean<br>$\overline{\log P_e}$ | mean<br>$\overline{Z_e}$ | mean<br>$\overline{R}^1$ | |
|---|---|---|---|---|---|
| 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. 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.

# Coarse-grained conformational energy

$$E^{conf} = E^{l} + E^{s} = E^{c} + E^{r} + E^{s} \tag{24}$$

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 $(\phi, \psi)$ statistical potential here. |

Statistical potentials previously estimated are used for the potentials above except for the orientational potential that is reported here. (*J. Mol. Biol.*, **256**, 623-644, 1996; *Proteins*, **34**, 49-68, 1999; *Proteins*, **36**, 347-356, 1999)

# Contact energies

The contact energy, $e^c_{aa'}$, between residues of type $a$ and $a'$ were estimated in the Bethe approximation as

$$e^c_{aa'} \equiv e^c_{rr} + \Delta e^c_{aa'} \tag{25}$$

$$\Delta e^c_{aa'} \equiv \Delta e^c_{ar} + \Delta e^c_{ra'} + \delta e^c_{aa'} \tag{26}$$

$$\text{Collapse energy:} \quad e^c_{rr} \qquad \text{cannot be estimated from contact frequencies.} \tag{27}$$

$$r\text{-}0 + 0\text{-}r \longrightarrow r\text{-}r + 0\text{-}0$$

$$\text{Hydrophobic partition energy:} \quad \exp(\Delta e^c_{ar}) = n_{a0} \,/\, [\, \frac{n_{ar} n_{r0}}{n_{rr}} \,] \tag{28}$$

$$a\text{-}r + r\text{-}0 \longrightarrow a\text{-}0 + r\text{-}r$$

$$\text{Intrinsic contact energy:} \quad \exp(-\delta e^c_{aa'}) = n_{aa'} \,/\, [\, \frac{n_{ar} n_{ra'}}{n_{rr}} \,] \tag{29}$$

$$a\text{-}r + r\text{-}a' \longrightarrow a\text{-}a' + r\text{-}r$$

where

$n^c_{aa'} + n^c_{a'a}$ the number of contacts between residues of type $a$ and $a'$ in native structures.

The index "0" means the water and "r" means any type of residue.

# Repulsive potentials to prevent packing at overly high densities

$$E^r = \sum_i [\frac{1}{2} \sum_j \{e^{hc}(\vec{r}_i, \vec{r}_j) + e^{re}_{ij}\} + e^{rp}_i] \tag{30}$$

$$\tag{31}$$

where

$$\text{Hard/soft core repulsive energy:} \quad e^{hc}(\vec{r}_i, \vec{r}_j) \equiv 10\ S_w(|\vec{r}_i - \vec{r}_j|, 2.2, 2.6) \tag{32}$$

$$\text{Excess contact energy:} \quad e^{re}_{ij} = H(n^c_i - q^c_{a_i})\,[\,(\frac{q^c_{a_i}}{n^c_i} - 1)e^c(\vec{r}_i, \vec{r}_j)\,] \tag{33}$$

$$\text{Repulsive packing energy:} \quad e^{rp}_i = H(n^c_i - q^c_{a_i})\,[\,-\log(\frac{N(a_i, n^c_i) + \epsilon}{N(a_i, q^c_{a_i}) + \epsilon})\,] \tag{34}$$

$$\text{Total number of contacting residues:} \quad n^c_i = \sum_j \Delta^c(\vec{r}_i, \vec{r}_j) \tag{35}$$

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 occurrences for the type of residue $a_i$ with the number of contacting residues $n^c_i$, and $\epsilon$ 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;

$$E^s = \sum_i e^s_{a_i}(\phi_i, \psi_i) \tag{36}$$

$$e^s_a(\phi, \psi) \equiv -\log(N_a(\phi, \psi)/N_a) + <\log(N_a(\phi, \psi)/N_a)> \tag{37}$$

$$< -\log(N_a(\phi, \psi)/N_a) >= \frac{-1}{N_a} \sum_{(\phi,\psi)} N_a(\phi, \psi) \log(N_a(\phi, \psi)/N_a) \tag{38}$$

where

| | |
|---|---|
| $N_a(\phi, \psi)$ | the number of amino acids of type $a$ at $(\phi, \psi)$ observed in protein native structures, |
| $N_a$ | the sum of $N_a(\phi, \psi)$ over the entire $(\phi, \psi)$ space. |

# 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.