# Self-consistent Estimation of Inter-residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues

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

[1]Gunma University, Faculty of Technology
[2]National Institutes of Health, NCI, LECB

(July 10, 1999)

# Abstract

Pairwise contact energies for 20 types of residues are estimated self-consistently from the actual observed frequencies of contacts with regression coefficients that are obtained by comparing "input" and predicted values with the Bethe approximation for the equilibrium mixtures of residues interacting. This is premised on the fact that correlations between the "input" and the predicted values are sufficiently high although the regression coefficients themselves can depend to some extent on protein structures as well as interaction strengths.

Residues in each native protein structure are shuffled to generate the equilibrium mixture of residues. The relative hydrophobic energies $\Delta e_{ir} (\equiv e_{ir} - e_{rr})$ and the intrinsic pairwise energies $\delta e_{ij} (\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$ are predicted from the numbers of contacts accumulated over proteins with the Bethe approximation; $e_{rr}$ is a collapse energy and cannot be evaluated for this system. Residue coordination numbers are optimized to obtain the best correlation between "input" and predicted values for the relative hydrophobic energies. Regression coefficients between the "input" and predicted values for both $\Delta e_{ir}$ and $\delta e_{ij}$ are calculated and used to obtain better estimates of relative contact energies $\Delta e_{ij} (\equiv e_{ij} - e_{rr})$.

The contact energies self-consistently estimated this way indicate that the relative hydrophobic energies predicted with the Bethe approximation should be reduced by a factor of about 0.3 and the intrinsic pairwise energies by a factor of about 0.6. This equilibrium mixture approximation of residues for proteins is supported at least to the extent that the observed distribution of contacts can be approximated with a small relative error of only about 0.08 as an equilibrium mixture of residues, if many proteins were employed to collect more than $20,000$ contacts. Including repulsive packing interactions and secondary structure interactions further reduces the relative errors. These new contact energies are demonstrated by threading to have improved their ability to discriminate native structures from other non-native folds.

Reference: *Proteins*, 34:49-68 (1999)

# Questions considered here

1. How accurately can the Bethe approximation (quasi-chemical approximation) reproduce “input” contact energies from equilibrium distributions of residue mixtures?

2. How well can the actual observed distributions of contacts in protein native structures be approximated as equilibrium mixtures of unconnected residues?

## How to answer these questions

1. The reproducibility of “input” contact energies by the Bethe approximation is analyzed by comparing “input” contact energies with those extracted from equilibrium mixtures of unconnected residues in proteins.

   - Equilibrium mixtures of residues are generated in a Monte Carlo simulation by shuffling residues in each protein structure.
   - The “input” contact energies assumed are estimated from the equilibrium distributions of contacts with the Bethe approximation.
   - The “input” and predicted contact energies are compared; regression and correlation coefficients are calculated.
   - The regression coefficients obtained from the equilibrium distributions of contacts accumulated over proteins are used to estimate contact energies from the actual observed frequencies of contacts in the proteins. Because these regression coefficients depend on the strength of interaction energies as well as protein structures, the self-consistent estimates of contact energies are interactively calculated.

2. The adequacy of the equilibrium mixture approximation of residues for inter-residue contacts in proteins is examined by comparing the equilibrium and actual observed frequencies of contacts.

In this way, the self-consistency of extracted potentials is tested, and then contact energies are estimated based on these analyses.

# Questions considered here

1. How accurately can the Bethe approximation (quasi-chemical approximation) reproduce “input” contact energies from equilibrium distributions of residue mixtures?

2. How well can the actual observed distributions of contacts in protein native structures be approximated as equilibrium mixtures of unconnected residues?

## How to answer these questions

1. The reproducibility of “input” contact energies by the Bethe approximation is analyzed by comparing “input” contact energies with those extracted from equilibrium mixtures of unconnected residues in proteins.

2. The adequacy of the equilibrium mixture approximation of residues for inter-residue contacts in proteins is examined by comparing the equilibrium and actual observed frequencies of contacts.

In this way, the self-consistency of extracted potentials is tested, and then contact energies are estimated based on these analyses.

# Methods

## Constraints on the number of contacts, $n_{ij}$

Equilibrium to be attained by shuffling residues within each protein is considered.

$$\begin{aligned}
\sum_{i=1}\sum_{j=1} n_{ij} &= n_{rr} && (1)\\
\sum_{i=1} n_{i0} &= n_{r0} && (2)\\
\sum_{j=0} n_{ij} &= \frac{q_i n_i}{2} && (3)
\end{aligned}$$

$n_{ij}$ : the number of contacts between $i$ and $j$ type of residues in a protein; $n_{ij} = n_{ji}$
the type 0 means solvent.

$q_i$, $n_i$ : coordination number for an $i$ type of residue and the number of $i$ type residues

## What interaction energies can be estimated for this system?

- Configuration partition function:

$$Z = const \sum_{\{n_{ij}\}} \frac{n_{r0}!n_{0r}!n_{rr}!}{\prod_{i=1} n_{i0}! \prod_{j=1} n_{0j}! \prod_{i=1}\prod_{j=1} n_{ij}!} \exp(-\sum_{i=1}\sum_{j=1} e_{ij} n_{ij}) \qquad (4)$$

$$e_{ij} = e_{ji} \equiv E_{ij} + E_{00} - E_{i0} - E_{0j} \qquad (5)$$

$E_{ij}$ : absolute contact energy between $i$ and $j$ type of residues.

- Relative contact energy $\Delta e_{ij}$ can be estimated.

$$\begin{aligned}
\Delta e_{ij} \equiv e_{ij} - e_{rr} &= \Delta e_{ir} + \Delta e_{rj} + \delta e_{ij} && (6)\\
\Delta e_{ir} \equiv e_{ir} - e_{rr} &= \log[\ \overline{n_{i0}}/[\frac{\overline{n_{ir} n_{r0}}}{n_{rr}}]\ ] && (7)\\
\delta e_{ij} \equiv e_{ij} + e_{rr} - e_{ir} - e_{rj} &= -\log[\ \overline{n_{ij}}/[\frac{\overline{n_{ir} n_{rj}}}{n_{rr}}]\ ] && (8)
\end{aligned}$$

- Collapse energy $e_{rr}$ cannot be evaluated because structures are fixed.

$$\begin{aligned}
\exp(-e_{rr}) &\equiv [\frac{\sum_{i=1}\sum_{j=1} \overline{n_{ij}} \exp(e_{ij})}{n_{rr}}]^{-1} && (9)\\
&= \frac{\sum_{i=1}\sum_{j=1} \overline{n_{i0} n_{0j}} \exp(-e_{ij})}{n_{r0} n_{0r}} && (10)
\end{aligned}$$

- Hydrophobic energies $\Delta e_{ir}$ and intrinsic pairwise energies $\delta e_{ij}$ are estimated by

$$
\begin{aligned}
\exp(\Delta e_{ir}) &= \frac{N_{i0}}{C_{i0}} && (11)\\
\exp(-\delta e_{ij}) &= \frac{N_{ij}}{C_{ij}} && (12)
\end{aligned}
$$

$N_{ij}$: the number of $i - j$ contacts accumulated over proteins

$C_{ij}$: the expected value of $N_{ij}$ in random mixing with fixed $N_{i0}$ or $N_{r0}$

**Procedure:**

1. - Equilibrium mixtures of residues are generated in a Monte Carlo simulation by shuffling residues in each protein structure.
   - The "input" contact energies assumed are estimated from the equilibrium distributions of contacts with the Bethe approximation.
   - The "input" and predicted contact energies are compared; regression and correlation coefficients are calculated.
   - The regression coefficients obtained from the equilibrium distributions of contacts accumulated over proteins are used to estimate contact energies from the actual observed frequencies of contacts in the proteins.

     Because these regression coefficients depend on the strength of interaction energies as well as protein structures, the self-consistent estimates of contact energies are interactively calculated.
2. The adequacy of the equilibrium mixture approximation of residues for inter-residue contacts in proteins is examined by comparing the equilibrium and actual observed frequencies of contacts.

Bethe近似による隣接エネルギーの再現性に関する、実際のタンパク質構造および格子タンパク質を用いての予備的計算からの示唆。

- 隣接エネルギーを過大に評価。残基間隣接数の平衡分布と既知のタンパク質で見い出される隣接数との比較から示唆される。
- 入力値と予測値との相関は $\Delta e_{ir}$ は 0.9以上 および $\delta e_{ij}$ は 0.8以上。
- 回帰係数は、相互作用の強弱やタンパク構造に強く依存する。

Table IV. Correlations between "input" and predicted values for relative partition energies ($\Delta e_{ir}$) and intrinsic inter-residue energies ($\delta e_{ij}$) from the equilibrium distributions of amino acid mixtures in the most compact configurations on lattices. (A) is for simple cubic lattices and (B) for face-centered cubic lattices; in these calculations, the value of the coordination number is fixed to the actual value, 6 or 12. The self-consistently estimated values of contact energies in Method-B are used as "input" nearest neighbor interactions. Relative temperature $T_{\text{rel}}$ is taken to be one.

A. Simple cubic lattice

| #residues | $n_{r0}/(q_r n_r/2)$ | $\Delta e_{ir}(\equiv e_{ir} - e_{rr})$ | | $\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$ | |
|---|---|---|---|---|---|
| | | correlation coefficient | regression coefficient | correlation coefficient | regression coefficient |
| 64 | 0.25 | 0.991 | 1.13 | 0.995 | 0.97 |
| 125 | 0.20 | 0.987 | 0.95 | 0.992 | 0.89 |
| 216 | 0.17 | 0.985 | 0.85 | 0.990 | 0.86 |
| 343 | 0.14 | 0.984 | 0.80 | 0.991 | 0.84 |
| 512 | 0.12 | 0.983 | 0.75 | 0.990 | 0.83 |
| 729 | 0.11 | 0.982 | 0.72 | 0.990 | 0.84 |

B. Face-centered cubic lattice

| #residues | $n_{r0}/(q_r n_r/2)$ | $\Delta e_{ir}(\equiv e_{ir} - e_{rr})$ | | $\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$ | |
|---|---|---|---|---|---|
| | | correlation coefficient | regression coefficient | correlation coefficient | regression coefficient |
| 63 | 0.37 | 0.993 | 0.38 | 0.971 | 0.84 |
| 108 | 0.31 | 0.992 | 0.35 | 0.942 | 0.65 |
| 172 | 0.27 | 0.994 | 0.32 | 0.914 | 0.54 |
| 256 | 0.23 | 0.993 | 0.29 | 0.890 | 0.47 |
| 365 | 0.21 | 0.993 | 0.28 | 0.870 | 0.43 |
| 500 | 0.19 | 0.993 | 0.26 | 0.853 | 0.39 |
| 666 | 0.17 | 0.994 | 0.25 | 0.838 | 0.37 |

**Iterative procedure to self-consistently estimate contact energies**

1. Assume the relative contact energies $\Delta e_{ij}^{\text{obs}}$ calculated with the Bethe approximation from the actual observed frequencies $N_{ij}^{\text{obs}}$ of contacts in proteins as "input" energies $\Delta e_{ij}^{\text{input}}$.

2. Perform a Monte Carlo simulation for each protein in which residues in a protein are assumed to interact with each other with the pairwise contact energies and shuffled to obtain an equilibrium distribution of contacts.

3. Calculate predicted contact energies $\Delta e_{ij}^{\text{pred}}$ by using the Bethe approximation from the total equilibrium distribution $N_{ij}^{\text{equil}}$ of contacts.

4. Calculate regression coefficients of the "input" versus "predicted" values for both types of energies.

$$
\begin{aligned}
\Delta e_{ir}^{\text{input}} &\sim \alpha \cdot \Delta e_{ir}^{\text{pred}} + \text{constant} && (13) \\
\delta e_{ij}^{\text{input}} &\sim \beta \cdot \delta e_{ij}^{\text{pred}} + \text{constant} && (14) \\
\Delta e_{ij}^{\text{input}} &\sim \eta \cdot [\alpha \cdot \Delta e_{ir}^{\text{pred}} + \alpha \cdot \Delta e_{rj}^{\text{pred}} + \beta \cdot \delta e_{ij}^{\text{pred}}] + \text{const} && (15)
\end{aligned}
$$

5. Calculate better estimates of real contact energies by using the new estimates of the regression coefficients obtained in the simulation together with the predicted energies with the Bethe approximation from the numbers of contacts actually observed in protein structures.

$$
\begin{aligned}
\Delta \varepsilon_{ir} &= \eta \cdot \alpha \cdot \Delta e_{ir}^{\text{obs}} && (16) \\
\delta \varepsilon_{ij} &= \eta \cdot \beta \cdot \delta e_{ij}^{\text{obs}} && (17) \\
\Delta \varepsilon_{ij} &= \Delta \varepsilon_{ir} + \Delta \varepsilon_{rj} + \delta \varepsilon_{ij} && (18)
\end{aligned}
$$

6. If the regression coefficients do not indicate a sufficiently good match with those values used to estimate contact energies for this iteration, repeat steps from 2 through 6 again by updating the "input" energies. Otherwise the procedure is completed and yields newly estimated energies.

## Proteins used

A sampling weight of each protein is determined on the basis of sequence identities between proteins.

Table 1. Summary of protein structures used in the present analysis.

| | |
|---|---|
| number of protein structures[a] | 1168 |
| number of protein subunit structures | 1661 |
| number of protein families[b] | 424 |
| effective number of proteins $(\sum_i w_i)$[c] | 251 |

[a] Structures whose resolutions were higher than 2.5$\AA$ and which were determined by X ray analyses and larger than 50 residues.

[b] A set of proteins with less than 95% sequence identity between any pair.

[c] Refer to Miyazawa & Jernigan (J. Mol. Biol., 256:623-644, 1996).

# Results

## Comparison of “input” energies $\Delta e_{ir}^{\text{input}}$ and energies $\Delta e_{ir}^{\text{pred}}$ predicted with the Bethe approximation for hydrophobic energies.

Method-A; Coordination numbers $(q_i)$ are fixed at the original values.
Interaction energies consist of contact energies only.

- The regression coefficient is 0.17.
- The correlation coefficient is 0.97.

**Comparison of "input" energies $\delta e_{ij}^{\rm input}$ and energies $\delta e_{ij}^{\rm pred}$ predicted with the Bethe approximation for intrinsic pairwise energies.**

Method-D; Coordination numbers ($q_i$) are also self-consistently optimized.
Secondary structure and tertiary structure energies are included.

- The regression coefficient is 0.61.
- The correlation coefficient is 0.94.

**Comparison of "input" energies $\Delta e_{ij}^{\text{input}}$ and energies $\Delta e_{ij}^{\text{est}}$ estimated from the regression coefficients between "input" values and values calculated with the Bethe approximation, for contact energies.**

Method-D; Coordination numbers $(q_i)$ are also self-consistently optimized.
Secondary structure and tertiary structure energies are included.

- The correlation coefficient is 0.99.

**Dependences of the regression coefficients of "input" versus predicted values, $\alpha$ for the hydrophobic energies $\Delta e_{ir}$, on the surface-volume ratios, $n_{r0}/(q_r n_r/2)$, of monomeric proteins.**

**Dependences of the regression coefficients of "input" versus predicted values, $\beta$ for the intrinsic pairwise energies $\delta e_{ij}$, on the surface-volume ratios, $n_{r0}/(q_r n_r/2)$, of monomeric proteins.**

**Estimation of real energies from energies calculated in the Bethe Approximation with the regression coefficients between "input" values and these predicted values.**

- The correlation coefficients between "input" and predicted values with the Bethe approximation for hydrophobic energies $\Delta e_{ir}$ and for intrinsic pairwise energies $\delta e_{ij}$ are better than 0.95. The new estimates, $\epsilon_{ij}$, of contact energies are calculated by

$$
\begin{aligned}
\Delta \epsilon_{ir} &= 0.26 \cdot \Delta e_{ir}^{\text{obs}} && (19) \\
\delta \epsilon_{ij} &= 0.57 \cdot \delta e_{ij}^{\text{obs}} && (20)
\end{aligned}
$$

- This new estimate of contact energies is more reasonable than the calculated values with the Bethe approximation.
  - ○ $\Delta \epsilon_{cys,cys}$ is more stable than $\Delta e_{ij}^{\text{obs}}$.
  - ○ $\Delta e_{ij}^{\text{obs}} < 0$ but $\Delta \epsilon_{ij} > 0$ for residue pairs between (Glu, Asp, Arg, Lys) and (Met, Phe, Ile, Leu, Val).

### Can the distribution of contacts be approximated as the equilibrium mixture of unconnected residues?

Comparison of equilibrium distributions of inter-residue contacts with actual observed distributions in proteins.

- The relative errors of the distributions decrease with the total number of contacts accumulated over proteins in a power dependence close to the value 1/2 expected for random sampling errors, but these relative errors attain a limit, about 0.08 at about $20,000$ contacts.

$$\begin{aligned} \frac{|\Delta N_{ij}|}{|N_{ij}|} &\propto \ N_{rr}^{-0.43} \quad \text{for} \quad N_{rr} < 10^4 & (21) \\ &\sim \ 0.08 \quad \text{for} \quad N_{rr} > 10^4 & (22) \end{aligned}$$

- The power dependence of such relative errors on the total number of contacts in each protein is much smaller than the value expected for random sampling errors; it may be due to the effect of chain connectivity.

$$\frac{|\Delta n_{ij}|}{|n_{ij}|} \propto n_{rr}^{-0.28} \tag{23}$$

**New contact energies increase the capability of fold recognition.**

Comparison of the effects of contact energies on the discrimination of native structures from other non-native folds with a given sequence.

Both ordinate and abscissa show z-scores that are defined as the total energy scores per residue of proteins in standard deviation units from the mean in the energy distribution of random threadings; the estimate of contact energies with the Bethe approximation is used on the abscissa and the present estimate with Method-D in Table II is used on the ordinate.

# Conclusion

- The equilibrium mixture approximation of residues for proteins is supported at least to the extent that the observed frequencies of residue pairs in contact can be approximated with a relative error of about 0.08, if many proteins are employed to collect more than $20,000$ contacts.

- Correlations between “input” and values predicted with the Bethe approximation for both types of energies are so high that we can statistically estimate the “input” energies from those predicted values. The contact energies self-consistently estimated indicate that the hydrophobic energies $\Delta e_{ir}$ predicted with the Bethe approximation may be reduced by a factor of about 0.3 and the intrinsic pairwise energies $\delta e_{ij}$ by a factor of about 0.6, decreasing the contribution of the hydrophobic energies.

- The new estimate for contact energies, in which the proportion of hydrophobic energies is much less than in the original one, is more reasonable, and increases the capability for discriminating native structures from other non-native folds.

Conformational energy: secondary structure + tertiary structure energy

$$E^{conf} \equiv E^{sec} + E^{tert} \tag{24}$$

Tertiary structure energy: contact energy + repulsive packing energy

$$E_p^{tert} = E_p^c + E_p^r \tag{25}$$

### The effects of repulsive packing energy

Method-B; contact energy only.

Method-C; contact energy + repulsive packing energy

In Method-B and C, coordination numbers ($q_i$) are self-consistently optimized.

### The effects of secondary structure energy

Method-C; contact energy + repulsive packing energy

Method-D: secondary structure + tertiary structure energy

In Method-C and D, coordination numbers $(q_i)$ are self-consistently optimized.

**The effects of Secondary structure energy**

Method-C; contact energy + repulsive packing energy

Method-D: secondary structure + tertiary structure energy

In Method-C and D, coordination numbers ($q_i$) are self-consistently optimized.