Statistical formulae of the energy distribution among a globular protein structure ensemble

Takuyo Aita[1,2,3] and Yuzuru Husimi[2*]

[1] *Tsukuba Research Institute, Novartis Pharma K. K., Ohkubo 8,*
Tsukuba, 300-2611, Japan

[2] *Department of Functional Materials Science, Saitama University,*
Saitama 338-8570, Japan

[3] *Computational Biology Research Center,*
National Institute of Advanced Industrial Science and Technology
Tokyo 135-0064, Japan

[3] current address

[*] corresponding author

e-mail address: husimi@fms.saitama-u.ac.jp

Tel & Fax +81-48-858-3531

August 29, 2002

Abstract

In prediction of a protein main-chain structure into which a query sequence of amino acids folds, one evaluates the relative stability of a candidate structure against reference structures. We developed a statistical theory for calculating the energy distribution over a main-chain structure ensemble, only with an amino acid composition given as a single argument. Then, we obtained a statistical formulae of the ensemble mean $\langle E \rangle$ and ensemble variance $V[E]$ of the reference structural energies, as explicit functions of the amino acid composition. The mean $\langle E \rangle$ and the variance $V[E]$ calculated from the formulae were well or roughly consistent with those resulting from a gapless threading simulation. We can use the formulae not only to perform the high-through-put screening of sequences in the inverse folding problem, but also to handle the problem analytically.

1 Introduction

The calculation of energy distributions over a protein 3D structure ensemble with a given amino acid sequence has been well conducted in protein physics (Miyazawa & Jernigan,1999; Zou & Saven,2000; Choy & Forman-Kay,2001; Paci *et al.*,2001). In prediction of a protein 3D structure through the threading method, the fitness of a candidate structure is evaluated with the Z score of the structural energy against a reference energy distribution: $Z = -(E_{\text{candidate}} - \langle E_{\text{reference}} \rangle)/SD[E_{\text{reference}}]$, where the $\langle E_{\text{reference}} \rangle$ and $SD[E_{\text{reference}}]$ are the ensemble mean and ensemble standard deviation of the reference structural energies, respectively. The calculation of both $\langle E_{\text{reference}} \rangle$ and $SD[E_{\text{reference}}]$ has been conducted by using computer experiments in many cases (Casari & Sippl,1992; Babajide *et al.*,1997; Zou & Saven,2000). In the threading method, one mounts a query sequence onto each of the reference structures, and calculate the energy distributions over the whole structure ensemble.

We assumed that the energy distributions depend significantly on the amino acid composition rather than on the sequence. This assumption has been also suggested from the Random Energy Model (Bryngelson & Wolynes,1987; Pande *et al.*,1997; Mirny *et al.*,2000). Our aim is to obtain formulae that output the values of $\langle E_{\text{reference}}\rangle$ and $SD[E_{\text{reference}}]$ straightforwardly when a query composition of amino acids is given. It is of great advantage that the formulae can save a considerable amount of effort in computer experiments for calculating the $\langle E_{\text{reference}}\rangle$ and $SD[E_{\text{reference}}]$ through the threading procedures. As a result, we succeeded in deriving the formulae for several energy terms such as side-chain vs side-chain interaction or hydration. Focusing on a main-chain structure ensemble consisting of various globular proteins with a constant chain length, we exemplified the validity of the formulae by comparing the theoretical results with the results from a gapless threading simulation.

2 Formulating energy distributions for individual energy terms

Let α (or β) be an arbitrary amino acid residue among different λ types ($\alpha, \beta \in \{1, 2, 3, \cdots, \lambda\}$), where λ is the number of available amino acids and $\lambda = 20$ for naturally occuring amino acids. We consider that an arbitrary sequence with the chain length ν and with amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, 3, \cdots, \lambda\}; \sum_\alpha \nu_\alpha = \nu\}$ is mounted on a globular main-chain structure i ($i = 1, 2, 3, \cdots, N$) with the same chain length ν, and the rotamer states of all side-chains are optimized at a constant temperature. The ν_α is the occurence frequency of an amino acid α among ν sites ($\sum_\alpha \nu_\alpha = \nu$). There are no gaps along the main-chain in the mounting process. The theoretical distributions for individual energy terms of (1) side-chain vs side-chain interaction energy E_{ss}, (2) side-chain vs main-chain interaction energy E_{sm}, (3) hydration energy E_{hy} and (4) secondary structure energy E_{se} are described in the following subsections.

2.1 Side-chain vs side-chain interaction energy E_{ss}

We designate the spatial distance between the side-chain center of a residue α_j at the jth site and that of a residue $\beta_{j'}$ at the j'th site as the "distance-through-space" for the pair of α_j and $\beta_{j'}$ ($\alpha_j, \beta_{j'} \in \{1, 2, 3, \cdots, \lambda\}; j, j' \in \{1, 2, 3, \cdots, \nu\}$). We also designate the number of peptide bonds that lie between the residue α_j and residue $\beta_{j'}$, $|j - j'|$ as the

”distance-through-bond” for the pair of α_j and $\beta_{j'}$. If the distance-through-space is less than the cutoff distance R_{c}, we regard the pair of α_j and $\beta_{j'}$ as interacting with each other. The R_{c} takes 7.5-10$\overset{\circ}{\mathrm{A}}$ in usual cases. We adopt the following assumptions.

Assumption 1: Any pair taken from among the ν residues is possible to interact with each other with equal probability. In other words, the probability that the distance-through-space for a certain pair is less than R_{c} is independent of the distance-through-bond for the pair.

Assumption 2: For the pairs of interacting residues, there is no correlation between the distance-through-bond and the distance-through-space.

Based on these assumptions, the effect of the sequence (=the order of occuring amino acids from N terminus through C terminus) is negligible and the amino acid composition is a single argument (Pande *et al.*,1997; Miyazawa & Jernigan,1999).

In the main-chain structure i, let $n_{\alpha\beta}$ be the number of pairs of interacting residues α and β, and let n be the number of all pairs of interacting residues:

$$n = \sum_{\alpha\beta} n_{\alpha\beta},$$

where $\sum_{\alpha\beta}$ is the summation over all of the $\binom{\lambda+2-1}{2}$ pairs of residues. Let $e_{\alpha\beta}^{(k)}$ be an interaction energy contributed by the k-th pair of interacting residues α and β. The total interaction energy E of the structure i is

$$E = \sum_{\alpha\beta} \sum_{k=1}^{n_{\alpha\beta}} e_{\alpha\beta}^{(k)}.$$

According to **assumption 1** and **2**, we adopt another assumption as follows.

Assumption 3: The $e_{\alpha\beta}^{(k)}$ takes an independent random value that obeys the probability density $p_{\alpha\beta}(e)$, which is specific to the residue pair (α, β).

Let $\epsilon_{\alpha\beta}$ and $\sigma_{\alpha\beta}^2$ be the mean and variance of the density function $p_{\alpha\beta}(e)$, respectively. When $n_{\alpha\beta}$'s are fixed, the density function of energy E is given by the convolution of the $n_{\alpha\beta}$-fold convolution of $p_{\alpha\beta}(e)$:

$$\begin{aligned}
&(\overset{n_{11}}{*}\ p_{11}(E)) * (\overset{n_{12}}{*}\ p_{12}(E)) * \cdots * (\overset{n_{\alpha\beta}}{*}\ p_{\alpha\beta}(E)) * \cdots * (\overset{n_{\lambda\lambda}}{*}\ p_{\lambda\lambda}(E)) \\
\approx\ & \mathcal{N}(E|\sum_{\alpha\beta} \epsilon_{\alpha\beta}\ n_{\alpha\beta}, \sum_{\alpha\beta} \sigma_{\alpha\beta}^2\ n_{\alpha\beta}),
\end{aligned}$$

where $\mathcal{N}(E|mean, variance)$ represents a normal distribution of a variable E with a $mean$ and $variance$.

According to **Assumption 1**, the probability $q_{\alpha\beta}$ that an arbitrary pair from among n pairs of interacting sites is a pair of residues α and β is

$$q_{\alpha\beta} = \begin{cases} \binom{\nu_\alpha}{2} / \binom{\nu}{2}, & \text{if } \alpha = \beta \\ \nu_\alpha \, \nu_\beta / \binom{\nu}{2}, & \text{if } \alpha \neq \beta, \end{cases} \quad . \tag{1}$$

Therefore, the probability that a state $\{n_{11}, n_{12}, \cdots, n_{\alpha\beta}, \cdots, n_{(\lambda-1)\lambda}, n_{\lambda\lambda}\}$ is realized obeys the following polynomial distribution:

$$\mathcal{M}(\{n_{\alpha\beta}\}|\{q_{\alpha\beta}\}) \equiv \frac{n!}{n_{11}! n_{12}! \cdots n_{\alpha\beta}! \cdots n_{\lambda\lambda}!} q_{11}^{n_{11}} q_{12}^{n_{12}} \cdots q_{\alpha\beta}^{n_{\alpha\beta}} \cdots q_{\lambda\lambda}^{n_{\lambda\lambda}}.$$

We assume that the probability distribution of n obeys a normal distribution $\mathcal{N}(n|\overline{n}, \Delta n^2)$. Then, the probability density of energy E of the structure i is given by

$$\sum_n \mathcal{N}(n|\overline{n}, \Delta n^2) \sum_{n_{11}} \sum_{n_{12}} \cdots \sum_{n_{\lambda\lambda}} \mathcal{N}(E| \sum_{\alpha\beta} \epsilon_{\alpha\beta} n_{\alpha\beta}, \sum_{\alpha\beta} \sigma_{\alpha\beta}^2 n_{\alpha\beta}) \; \mathcal{M}(\{n_{\alpha\beta}\}|\{q_{\alpha\beta}\}).$$

The probability density of energy E over the structure ensemble ($i = 1, 2, 3, \cdots, N$) is given by

$$\frac{1}{N} \sum_{i=1}^{N} \sum_n \mathcal{N}(n|\overline{n}^{(i)}, {\Delta n^{(i)}}^2) \sum_{n_{11}} \sum_{n_{12}} \cdots \sum_{n_{\lambda\lambda}} \mathcal{N}(E| \sum_{\alpha\beta} \epsilon_{\alpha\beta}^{(i)} n_{\alpha\beta}, \sum_{\alpha\beta} {\sigma_{\alpha\beta}^{(i)}}^2 n_{\alpha\beta}) \; \mathcal{M}(\{n_{\alpha\beta}\}|\{q_{\alpha\beta}\}),$$

where the superscript (i) represents "for the structure i" in this paper. The mean $\langle E \rangle$, $\langle E^2 \rangle$ and variance $V[E]$ of energy E over the structure ensemble are given, respectively, as follows:

$$\begin{aligned} \langle E \rangle &= \frac{1}{N} \sum_{i=1}^{N} \left(\overline{n}^{(i)} \sum_{\alpha\beta} \epsilon_{\alpha\beta}^{(i)} \, q_{\alpha\beta} \right) \\ &= \sum_{\alpha\beta} q_{\alpha\beta} \, \langle \overline{n} \epsilon_{\alpha\beta} \rangle \end{aligned} \tag{2}$$

$$\langle E^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \left(\overline{n}^{(i)} \sum_{\alpha\beta} ({\sigma_{\alpha\beta}^{(i)}}^2 + {\epsilon_{\alpha\beta}^{(i)}}^2) q_{\alpha\beta} + ({\Delta n^{(i)}}^2 - \overline{n}^{(i)} + {\overline{n}^{(i)}}^2) (\sum_{\alpha\beta} \epsilon_{\alpha\beta}^{(i)} q_{\alpha\beta})^2 \right) \tag{3}$$

$$\begin{aligned} V[E] &= \langle E^2 \rangle - \langle E \rangle^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(\overline{n}^{(i)} \sum_{\alpha\beta} ({\sigma_{\alpha\beta}^{(i)}}^2 + {\epsilon_{\alpha\beta}^{(i)}}^2) q_{\alpha\beta} + ({\Delta n^{(i)}}^2 - \overline{n}^{(i)}) (\sum_{\alpha\beta} \epsilon_{\alpha\beta}^{(i)} q_{\alpha\beta})^2 + \left(\sum_{\alpha\beta} q_{\alpha\beta} (\overline{n}^{(i)} \epsilon_{\alpha\beta}^{(i)} - \langle \overline{n} \epsilon_{\alpha\beta} \rangle) \right)^2 \right), \end{aligned} \tag{4}$$

where

$$\langle \overline{n}\epsilon_{\alpha\beta} \rangle = \frac{1}{N}\sum_{i=1}^{N} \overline{n}^{(i)}\epsilon_{\alpha\beta}^{(i)}.$$

The $\langle * \rangle$ and $V[*]$ represent the mean and variance of a quantity $*$ over the whole structure ensemble, respectively, in this paper.

As a result, once parameters $\{\epsilon_{\alpha\beta}^{(i)}, \sigma_{\alpha\beta}^{(i)}, \overline{n}^{(i)}, \Delta n^{(i)}\}$ for each structure i are determined in a preliminary computer simulation, we can straightforwardly calculate the mean $\langle E \rangle$ and variance $V[E]$ of the side-chain vs side-chain interaction energy E over the structure ensemble with a given amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, \cdots, \lambda\}\}$.

2.2 Side-chain vs main-chain interaction energy $E_{\rm sm}$

If the distance between the side-chain center of a residue α_j at the jth site ($\alpha_j \in \{1, 2, 3, \cdots, \lambda\}$) and the coordinate of a main-chain atom $\beta_{j'}$ ($\beta_{j'} = N, C, O, C_\alpha, C_\beta$) is less than $R_{\rm c}$, we regard the pair of α_j and $\beta_{j'}$ as interacting with each other. We adopt the same assumption as those in the previous section, regarding the pairs of interacting sites. Theoretical distribution of side-chain vs main-chain interaction energy is derived in the same way as that in the previous section, by regarding β as a main-chain atom ($\beta = N, C, O, C_\alpha, C_\beta$) and substituting eqn.(1) with

$$q_{\alpha\beta} = \frac{\nu_\alpha}{5\nu}. \tag{5}$$

2.3 Hydration energy $E_{\rm hy}$

Let $e_\alpha^{(k)}$ be a hydration energy contributed by the k-th residue among ν_α sites occupied by α. The total hydration energy E of structure i is

$$E = \sum_\alpha \sum_{k=1}^{\nu_\alpha} e_\alpha^{(k)},$$

where $\sum_\alpha$ is the summation over all residues ($\alpha \in \{1, 2, 3, \cdots, \lambda\}$). We adopt the following assumption.

Assumption 4: The $e_\alpha^{(k)}$ takes an independent random value that obeys the probability density $p_\alpha(e)$, which is specific to the amino acid residue α.

Let ϵ_α and σ_α^2 be the mean and variance of the density function $p_\alpha(e)$, respectively. The density function of energy E is given by the convolution of the ν_α-fold convolution of $p_\alpha(e)$:

$$\begin{aligned} & (\overset{\nu_1}{*}\, p_1(E)) * (\overset{\nu_2}{*}\, p_2(E)) * \cdots * (\overset{\nu_\alpha}{*}\, p_\alpha(E)) * \cdots * (\overset{\nu_\lambda}{*}\, p_\lambda(E)) \\ \approx\ & \mathcal{N}(E|\sum_\alpha \epsilon_\alpha\ \nu_\alpha, \sum_\alpha \sigma_\alpha^2\ \nu_\alpha). \end{aligned}$$

Then, the mean $\langle E\rangle$, $\langle E^2\rangle$ and variance $V[E]$ of the hydration energy E over the structure ensemble are given, respectively, as follows:

$$\langle E\rangle = \frac{1}{N}\sum_{i=1}^{N}\sum_\alpha \epsilon_\alpha^{(i)}\ \nu_\alpha \tag{6}$$

$$\langle E^2\rangle = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_\alpha \sigma_\alpha^{(i)^2}\ \nu_\alpha + (\sum_\alpha \epsilon_\alpha^{(i)}\ \nu_\alpha)^2\right) \tag{7}$$

$$V[E] = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_\alpha \sigma_\alpha^{(i)^2}\ \nu_\alpha + \left(\sum_\alpha \nu_\alpha(\epsilon_\alpha^{(i)} - \langle\epsilon_\alpha\rangle)\right)^2\right). \tag{8}$$

As a result, once parameters $\{\epsilon_\alpha^{(i)}, \sigma_\alpha^{(i)} | \alpha \in \{1, 2, \cdots, \lambda\}\}$ for each structure i are determined in a preliminary computer simulation, we can straightforwardly calculate the mean $\langle E\rangle$ and variance $V[E]$ of the hydration energy E over the structure ensemble with a given amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, \cdots, \lambda\}\}$.

2.4 Secondary structure energy $E_{\rm se}$

We consider that in the main-chain structure i ($i = 1, 2, 3, \cdots, N$), a secondary structure s (s $= 1, 2, \cdots$) such as the α-helix or β-strand, occurs with a probability $q_{\rm s}$ ($\sum_{\rm s} q_{\rm s} = 1$) at each site independently.

In a structure i, let $n_{\alpha{\rm s}}$ be the number of amino acid residues α which are located on the secondary structure s:

$$\nu_\alpha = \sum_{\rm s} n_{\alpha{\rm s}},$$

where $\sum_{\rm s}$ is the summation over all secondary structures (s $= 1, 2, \cdots$). Let $e_{\alpha{\rm s}}^{(k)}$ be a secondary structure energy contributed by the k-th pair $(\alpha, {\rm s})$ of α and s. The total secondary structure energy E of the structure i is

$$E = \sum_\alpha \sum_{\rm s} \sum_{k=1}^{n_{\alpha{\rm s}}} e_{\alpha{\rm s}}^{(k)}.$$

We adopt the last assumption as follows.

Assumption 5: The $e_{\alpha\mathrm{s}}^{(k)}$ takes an independent random value that obeys the probability density $p_{\alpha\mathrm{s}}(e)$, which is specific to the pair (α, s).

Let $\epsilon_{\alpha\mathrm{s}}$ and $\sigma_{\alpha\mathrm{s}}^2$ be the mean and variance of the density function $p_{\alpha\mathrm{s}}(e)$, respectively. By using the similar scheme with the section 2.1, the density function of energy E is given by

$$\sum_{\{n_{1\mathrm{s}}\}} \cdots \sum_{\{n_{\alpha\mathrm{s}}\}} \cdots \sum_{\{n_{\lambda\mathrm{s}}\}} \mathcal{N}(E | \sum_{\alpha}\sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}}\ n_{\alpha\mathrm{s}}, \sum_{\alpha}\sum_{\mathrm{s}} \sigma_{\alpha\mathrm{s}}^2\ n_{\alpha\mathrm{s}})\ \prod_{\alpha} \mathcal{M}_{\alpha}(\{n_{\alpha\mathrm{s}}\}|\{q_{\mathrm{s}}\}),$$

where $\mathcal{M}_{\alpha}(\{n_{\alpha\mathrm{s}}\}|\{q_{\mathrm{s}}\})$ is defined as the following polynomial distribution:

$$\mathcal{M}_{\alpha}(\{n_{\alpha\mathrm{s}}\}|\{q_{\mathrm{s}}\}) \equiv \frac{\nu_{\alpha}!}{\prod_{\mathrm{s}} n_{\alpha\mathrm{s}}!} \prod_{\mathrm{s}} q_{\mathrm{s}}^{n_{\alpha\mathrm{s}}},$$

and $\sum_{\{n_{\alpha s}\}}$ represents the summation over all possible states of $\{n_{\alpha 1}, n_{\alpha 2}, \cdots\}$ for α, that is,

$$\sum_{\{n_{\alpha\mathrm{s}}\}} \mathcal{M}_{\alpha}(\{n_{\alpha\mathrm{s}}\}|\{q_{\mathrm{s}}\}) = \sum_{\{n_{\alpha 1}, n_{\alpha 2}, \cdots\}} \mathcal{M}_{\alpha}(\{n_{\alpha\mathrm{s}}\}|\{q_{\mathrm{s}}\}) = 1.$$

Then, the probability density of energy E of the structure ensemble $(i = 1, 2, 3, \cdots, N)$ is given by

$$\frac{1}{N}\sum_{i=1}^{N} \sum_{\{n_{1\mathrm{s}}\}} \cdots \sum_{\{n_{\alpha\mathrm{s}}\}} \cdots \sum_{\{n_{\lambda\mathrm{s}}\}} \mathcal{N}(E | \sum_{\alpha}\sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}}\ n_{\alpha\mathrm{s}}, \sum_{\alpha}\sum_{\mathrm{s}} \sigma_{\alpha\mathrm{s}}^2\ n_{\alpha\mathrm{s}})\ \prod_{\alpha} \mathcal{M}_{\alpha}(\{n_{\alpha\mathrm{s}}\}|\{q_{\mathrm{s}}^{(i)}\}).$$

The mean $\langle E \rangle$, $\langle E^2 \rangle$ and variance $V[E]$ of energy E over the structure ensemble are given, respectively, as follows:

$$\langle E \rangle = \frac{1}{N}\sum_{i=1}^{N} \sum_{\alpha} \nu_{\alpha} \sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}}\, q_{\mathrm{s}}^{(i)}, \tag{9}$$

$$\langle E^2 \rangle = \frac{1}{N}\sum_{i=1}^{N} \left(\left(\sum_{\alpha} \nu_{\alpha} \left(\sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}}\, q_{\mathrm{s}}^{(i)} \right) \right)^2 + \sum_{\alpha} \nu_{\alpha} \left(\sum_{\mathrm{s}} ({\sigma_{\alpha\mathrm{s}}}^2 + {\epsilon_{\alpha\mathrm{s}}}^2) q_{\mathrm{s}}^{(i)} - \left(\sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}} q_{\mathrm{s}}^{(i)} \right)^2 \right) \right), \tag{10}$$

$$V[E] = \frac{1}{N}\sum_{i=1}^{N} \left(\left(\sum_{\alpha} \nu_{\alpha} \left(\sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}} (q_{\mathrm{s}}^{(i)} - \langle q_{\mathrm{s}} \rangle) \right) \right)^2 + \sum_{\alpha} \nu_{\alpha} \left(\sum_{\mathrm{s}} ({\sigma_{\alpha\mathrm{s}}}^2 + {\epsilon_{\alpha\mathrm{s}}}^2) q_{\mathrm{s}}^{(i)} - \left(\sum_{\mathrm{s}} \epsilon_{\alpha\mathrm{s}} q_{\mathrm{s}}^{(i)} \right)^2 \right) \right), \tag{11}$$

where

$$\langle q_{\mathrm{s}} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} q_{\mathrm{s}}^{(i)}.$$

The outline of the derivation of the above equations is described in the Appendix.

As a result, once the parameters $\{q_{\rm s}^{(i)}\}$ for each structure i and $\{\epsilon_{\alpha{\rm s}}, \sigma_{\alpha{\rm s}}\}$ for the structure ensemble are determined in a preliminary computer simulation, we can straightforwardly calculate the mean $\langle E\rangle$ and variance $V[E]$ of the secondary structure energy E over the structure ensemble with a given amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, \cdots, \lambda\}\}$.

3 Extracting parameters from a set of real protein structures

Here is the procedure of extracting parameters $\{\epsilon_{\alpha\beta}^{(i)}, \sigma_{\alpha\beta}^{(i)}, \overline{n}^{(i)}, \Delta n^{(i)}\}$ for the side-chain vs side-chain (or vs main-chain) interaction energy $E_{\rm ss}$ (or $E_{\rm sm}$), $\{\epsilon_\alpha^{(i)}, \sigma_\alpha^{(i)}\}$ for the hydration energy $E_{\rm hy}$, and $\{\epsilon_{\alpha{\rm s}}, \sigma_{\alpha{\rm s}}\}$ and $\{q_{\rm s}^{(i)}\}$ for the secondary structure energy $E_{\rm se}$ (see Fig. **??**).

(1) Retrieve a set of various globular main-chain structures ($i = 1, 2, 3, \cdots, N$) with the same chain length of ν from the Protein Data Bank (PDB).

(2) Mount a randomly generated sequence with a random amino acid composition onto a main-chain structure i, and subsequently optimize the rotamer states of all side-chains at a constant temperature. There are no gaps along the main-chain in the mounting process.

(3) Calculate the following quantities.

For $E_{\rm ss}$, calculate a frequency distribution of the interaction energy $e_{\alpha\beta}$ between interacting residues α and β ($\alpha\beta = 11, 12, 13, \cdots, \lambda\lambda$), and calculate the number n of all pairs of interacting residues.

For $E_{\rm sm}$, calculate a frequency distribution of the interaction energy $e_{\alpha\beta}$ between the interacting side-chain residue α and main-chain atom β, and calculate the number n of all pairs of interacting side-chain residue and main-chain atoms.

For $E_{\rm hy}$, calculate a frequency distribution of the hydration energy e_α contributed by a residue α ($\alpha \in \{1, 2, 3, \cdots, \lambda\}$). Although the composition of glycine affects the number of C$_\beta$ atoms, the effect is not so sensitive to the whole hydration energy E, except for special cases where almost all residues are glycine. Therefore, we assume that the C$_\beta$ atoms are located at every site in the main-chain structure.

For E_{se}, calculate a frequency $q_{\mathrm{s}}^{(i)}$ ($\sum_{\mathrm{s}} q_{\mathrm{s}}^{(i)} = 1$) for a secondary structure s in the main-chain structure i, and a frequency distribution of the secondary structure energy $e_{\alpha\mathrm{s}}$ for an amino acid residue α which is located on the secondary structure s ($\mathrm{s} = 1, 2, 3, \cdots$).

(4) Repeat the procedures (2)-(3) except for the calculation of $q_{\mathrm{s}}^{(i)}$s (for E_{se}) several times. For E_{ss} and E_{sm}, calculate the mean $\epsilon_{\alpha\beta}^{(i)}$ and standard deviation $\sigma_{\alpha\beta}^{(i)}$ of $e_{\alpha\beta}$. Calculate the mean $\overline{n}^{(i)}$ and standard deviation $\Delta n^{(i)}$ of n. For E_{hy}, calculate the mean $\epsilon_{\alpha}^{(i)}$ and the standard deviation $\sigma_{\alpha}^{(i)}$ of e_{α}.

(5) The procedures (1)-(4) are conducted through all the structures.

4 Customizing the formulae for the practical use

For E_{ss}, let l ($l = 1, 2, \cdots, L$) be the index number for an arbitrary pair (α, β) and replace $\epsilon_{\alpha\beta}$, $\sigma_{\alpha\beta}$ and $q_{\alpha\beta}$ in eqns (1)-(4) by ϵ_l, σ_l and q_l, respectively. Calculate the following parameters

$$\begin{aligned} A_l &\equiv \langle \overline{n}\epsilon_l \rangle \\ B_l &\equiv \langle \overline{n}({\sigma_l}^2 + {\epsilon_l}^2) \rangle \\ C_l &\equiv \langle (\Delta n^2 - \overline{n} + \overline{n}^2){\epsilon_l}^2 \rangle \\ D_{ll'} &\equiv \langle 2(\Delta n^2 - \overline{n} + \overline{n}^2)\epsilon_l \epsilon_{l'} \rangle. \end{aligned}$$

Given an amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, 3, \cdots, \lambda\}\}$, we can calculate the mean $\langle E_{\mathrm{ss}} \rangle$ (eqn. (2)) and variance $V[E_{\mathrm{ss}}]$ (eqn. (4)) over the structure ensemble by using

$$\langle E_{\mathrm{ss}} \rangle = \sum_{l=1}^{L} A_l q_l \tag{12}$$

$$\langle {E_{\mathrm{ss}}}^2 \rangle = \sum_{l=1}^{L} (B_l q_l + C_l q_l^2) + \sum_{l=1}^{L-1} \sum_{l'=l+1}^{L} D_{ll'} q_l q_{l'}, \tag{13}$$

where $L = \binom{\lambda+1}{2}$ and q_l is equivalent to $q_{\alpha\beta}$ in eqn. (1).

For E_{sm}, the mean $\langle E_{\mathrm{sm}} \rangle$ (eqn. (2)) and variance $V[E_{\mathrm{sm}}]$ (eqn. (4)) over the structure ensemble are given by using the same equations as eqn. (12) and (13), with that $L = 5\lambda$ and q_l is equivalent to $q_{\alpha\beta}$ in eqn. (5).

For $E_{\rm hy}$, calculate the following parameters

$$\begin{aligned} A_\alpha &\equiv \langle \epsilon_\alpha \rangle \\ B_\alpha &\equiv \langle {\sigma_\alpha}^2 \rangle \\ C_\alpha &\equiv \langle {\epsilon_\alpha}^2 \rangle \\ D_{\alpha\alpha'} &\equiv \langle 2\epsilon_\alpha \epsilon_{\alpha'} \rangle. \end{aligned}$$

Given an amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, 3, \cdots, \lambda\}\}$, we can calculate the mean $\langle E_{\rm hy} \rangle$ (eqn. (6)) and variance $V[E_{\rm hy}]$ (eqn. (8)) over the structure ensemble by using

$$\langle E_{\rm hy} \rangle = \sum_{\alpha=1}^{\lambda} A_\alpha \nu_\alpha \tag{14}$$

$$\langle {E_{\rm hy}}^2 \rangle = \sum_{\alpha=1}^{\lambda} (B_\alpha \nu_\alpha + C_\alpha \nu_\alpha^2) + \sum_{\alpha=1}^{\lambda-1} \sum_{\alpha'=\alpha+1}^{\lambda} D_{\alpha\alpha'} \nu_\alpha \nu_{\alpha'}. \tag{15}$$

For $E_{\rm se}$, calculate the mean $\epsilon_{\alpha{\rm s}}$ and standard deviation $\sigma_{\alpha{\rm s}}$ of $e_{\alpha{\rm s}}$. The mean and deviation are taken over the structure ensemble, because these statistics seem to be uncorrelated with the topology of each main-chain structure. Subsequently, calculate the following parameters

$$\begin{aligned} A_\alpha &\equiv \langle \sum_{\rm s} \epsilon_{\alpha{\rm s}} q_{\rm s} \rangle \\ B_\alpha &\equiv \langle \sum_{\rm s} ({\sigma_{\alpha{\rm s}}}^2 + {\epsilon_{\alpha{\rm s}}}^2) q_{\rm s} \rangle \\ C_\alpha &\equiv \langle (\sum_{\rm s} \epsilon_{\alpha{\rm s}} q_{\rm s})^2 \rangle \\ D_{\alpha\alpha'} &\equiv \langle 2(\sum_{\rm s} \epsilon_{\alpha{\rm s}} q_{\rm s})(\sum_{\rm s} \epsilon_{\alpha'{\rm s}} q_{\rm s}) \rangle. \end{aligned}$$

Given an amino acid composition $\{\nu_\alpha | \alpha \in \{1, 2, 3, \cdots, \lambda\}\}$, we can calculate the mean $\langle E_{\rm se} \rangle$ (eqn. (9)) and variance $V[E_{\rm se}]$ (eqn. (11)) over the structure ensemble by using

$$\langle E_{\rm se} \rangle = \sum_{\alpha=1}^{\lambda} A_\alpha \nu_\alpha \tag{16}$$

$$\langle {E_{\rm se}}^2 \rangle = \sum_{\alpha=1}^{\lambda} (B_\alpha \nu_\alpha + C_\alpha (\nu_\alpha^2 - \nu_\alpha)) + \sum_{\alpha=1}^{\lambda-1} \sum_{\alpha'=\alpha+1}^{\lambda} D_{\alpha\alpha'} \nu_\alpha \nu_{\alpha'}. \tag{17}$$

The overall energy $E_{\rm total}$ for each structure is given by

$$E_{\rm total} = E_{\rm ss} + E_{\rm sm} + E_{\rm hy} + E_{\rm se}.$$

Assuming that there is no correlation between different energy terms, that is the covariance between different energy terms is equal to zero, we can obtain the mean $\langle E_{\mathrm{total}}\rangle$ and variance $V[E_{\mathrm{total}}]$ over the structure ensemble as follows:

$$\begin{aligned}\langle E_{\mathrm{total}}\rangle &= \langle E_{\mathrm{ss}}\rangle + \langle E_{\mathrm{sm}}\rangle + \langle E_{\mathrm{hy}}\rangle + \langle E_{\mathrm{se}}\rangle \\ V[E_{\mathrm{total}}] &= V[E_{\mathrm{ss}}] + V[E_{\mathrm{sm}}] + V[E_{\mathrm{hy}}] + V[E_{\mathrm{se}}].\end{aligned}$$

5 Result and Discussion

We used a set of 159 globular proteins with the chain length of $\nu = 108$ as a main-chain structure ensemble. Since a small number of natural proteins with $\nu = 108$ is registered in the Protein Data Bank (PDB), we prepared the natural or artificial 159 main-chain structures with $\nu = 108$ by the following procedure.

Initially, we retrieved various globular proteins with the chain length of $108 \leq \nu \leq 200$ from the PDB. Regarding each of the proteins with $\nu \geq 109$, all possible fragments with the chain length of 108 are prepared by truncation of the main-chain. From among all the fragments, a particular fragment, whose root means square deviation

$$RMSD = \sqrt{\frac{1}{108}\sum_{j=j_0}^{j_0+107}(x_j-\overline{x})^2+(y_j-\overline{y})^2+(z_j-\overline{z})^2},$$

is the smallest while less than 14.0 $\overset{\circ}{A}$ is added into the main-chain structure ensemble, where (x_j, y_j, z_j) is the coordinate of α carbon at the jth site.

We carried out the procedure of extracting parameters from the set of the 159 main-chain structures. The energy function we used in this study is a knowledge-based potential function, which is based on a rotamer library defined by Ota *et al.* (Ota *et al.*,2001). When a main-chain structure of a target protein is given in the inverse folding problem, the energy of the system is defined as a function of both an amino acid sequence and a set of rotamer states of all side-chains. Once a query sequence is given, the optimal set of rotamer states can be determined through energy optimization. Therefore, the energy function we used has a single argument (=variable), that is the query sequence of the amino acids with the optimal set of rotamer states. The energy of a protein sequence

P mounted on a given main-chain structure i is composed of the following four energy terms:

$$E_{\rm total}({\rm P},i) = E_{\rm ss}({\rm P},i) + E_{\rm sm}({\rm P},i) + E_{\rm hy}({\rm P},i) + E_{\rm se}({\rm P},i).$$

For $E_{\rm ss}$ and $E_{\rm sm}$, the cutoff distance $R_{\rm c}$ for the interaction is set to be $10\overset{\circ}{\rm A}$ and all pairs whose distance-through-bond is less than 3 are excluded. The details are presented in the original papers cited above.

As a preliminary examination, we tested **assumption 1** and **2** for $E_{\rm ss}$ and $E_{\rm sm}$. In Fig. 1, we plotted the conditional probability that the distance-through-space is less than $10\overset{\circ}{\rm A}$ when the distance-through-bond is given. It is remarkable that the probability for the pairs of different side-chain centers and that for pairs of side-chain centers and main-chain atoms are almost the same as each other. The probability calculated from the ensemble of 159 main-chain structures is about 0.1 in cases where the distance-through-bond is greater than 10. This result roughly exemplifies **assumption 1**, although pairs having a small distance-through-bond do not satisfy **assumption 1** and the probabilities for individual proteins fluctuate more or less around 0.1 (In Fig. 1, we show the case for streptomyces subtilisin inhibitor, whose PDB code is 3ssi). In Fig. 2, we plotted the distributions of the distance-through-space for pairs whose distance-through-space is less than $10\overset{\circ}{\rm A}$. The distributions calculated over the ensemble of 159 main-chain structures, take a similar feature throughout the whole range of distance-through-bonds, while the distributions for individual proteins (e.g. 3ssi shown in Fig. 2(a)) fluctuate more or less around the averaged feature. This result roughly exemplifies **assumption 2**.

In order to compare the mean $\langle E_{\rm xx}\rangle$ and variance $V[E_{\rm xx}]$ derived from a computer simulation with those predicted from the formulae (eqns (12)-(17)), for each of the individual energy terms (”xx”=”ss”, ”sm”, ”hy”, ”se” and ”total”) we carried out the gapless threading simulation as follows. We adopted 760 amino acid compositions $\{\nu_\alpha|\alpha \in \{1,2,3,\cdots,\lambda\}; \sum_\alpha \nu_\alpha = 108\}$, which were randomly generated by the Monte Carlo method. Mathematically, the most expected composition is given by

$$\{\frac{\nu_\alpha}{\nu}|\alpha = 1,2,\cdots,\lambda\} = \{\frac{1}{\lambda}\sum_{m=r}^{\lambda}\frac{1}{m}|r = 1,2,\cdots,\lambda\}, \tag{18}$$

where r represents the rank of percentage of amino acid composition in descending order (Webb, 1974). Fig. 3 shows the most expected mole-fraction of amino acids calculated from eqn. (18) with $\lambda = 20$. Ten different amino acid sequences were randomly generated for each of the 760 amino acid compositions. Each of the amino acid sequences was mounted onto a main-chain structure without gaps, and subsequently, the rotamer states of all the side-chains were optimized at a constant temperature. The individual energy terms for the sequences were calculated over the structure ensemble. Then, we obtained the ensemble mean $\langle E_{\mathrm{xx}} \rangle$ and ensemble variance $V[E_{\mathrm{xx}}]$ for each of the sequences. Fig. 4 shows the average and standard deviation of $\langle E_{\mathrm{xx}} \rangle$ over 10 different sequences for each of the 760 amino acid compositions. Similarly, Fig. 5 shows the average and standard deviation of $SD[E_{\mathrm{xx}}] = \sqrt{V[E_{\mathrm{xx}}]}$ over 10 different sequences for each amino acid composition.

As a result, the mean $\langle E_{\mathrm{xx}} \rangle$ and standard deviation $SD[E_{\mathrm{xx}}]$ resulting from the simulation are well or roughly consistent with the theoretical values calculated from the formulae (eqns (12)-(17)). In spite that the theoretical values for $\langle E_{\mathrm{ss}} \rangle$ and $\langle E_{\mathrm{sm}} \rangle$ were based on the same **assumptions** $1 \sim 3$, the theoretical values for $\langle E_{\mathrm{ss}} \rangle$ show the largest discrepancy among all the energy terms ($r = 0.83$), whereas those for $\langle E_{\mathrm{sm}} \rangle$ show an excellent agreement with the simulated values ($r = 1.0$). The difference between the former and the latter seems to stem from the following statistical reasons. The atoms (N, C, O, C_{α}, C_{β}) in a main chain are evenly distributed in a globular protein structure, and then the statistically sufficient number of individual residue-atom pairs produces interaction, whereas the number of individual residue-residue pairs producing interaction is likely to be insufficient for statistical treatment. Regarding the result for $\langle E_{\mathrm{ss}} \rangle$, we examined what features are in the amino acid compositions showing large discrepancy. It turned out that the small amino acids, such as alanine, glycine and serine, and proline have small fractions, whereas phenylalanine and tyrosine have large fractions. Following these features, we infer that the significant frustration between large side-chains makes the optimal rotamer states far from **assumption 1** and **2**, for amino acid sequences consisting of many large amino acid residues. Regarding the result for $SD[E_{\mathrm{xx}}]$, the discrepancy tends to be linearly larger as the $SD[E_{\mathrm{xx}}]$ becomes larger. The prediction of $SD[E_{\mathrm{xx}}]$ can be improved by using the linear relationship between the theoretical and simulated values.

Fig. 6 shows the comparison of $Z = -(E_{\rm xx}^{\rm (native)} - \langle E_{\rm xx} \rangle)/SD[E_{\rm xx}]$ resulting from the computer simulation with that predicted from the formulae for each of the individual energy terms (”xx”=”ss”, ”sm”, ”hy”, ”se” and ”total”). We used $E_{\rm ss}^{\rm (native)} = -44.3$, $E_{\rm sm}^{\rm (native)} = 4.6$, $E_{\rm hy}^{\rm (native)} = -12.3$, and $E_{\rm se}^{\rm (native)} = -17.2$ and $E_{\rm total}^{\rm (native)} = -69.1$, which are the individual energies of the native sequence mounted on the corresponding structure of thioredoxin (PDB code is 2trxA). We can also see good agreement between the theoretical and simulated Z-scores.

These results suggest that the statistical model of proteins we adopted in this study is valid to calculate the energy distribution over the main-chain structure ensemble when an amino acid composition is given. The model is based on **assumptions** $1 \sim 5$, which make the effect of the amino acid sequence negligible. The effect of the sequences with the same amino acid composition is very small for $\langle E_{\rm xx} \rangle$ and Z-scores (Figs 4 and 6). This supports the idea that the energy distributions are significantly dependent on amino acid compositions and the effect of the sequence is almost negligible (Pande *et al.*,1997; Miyazawa & Jernigan,1999; Mirny *et al.*,2000). The predictability of the statistical formulae may increase as the diversity in the main-chain structure ensemble increases, the length of main-chain increases or the conformational entropy of rotamer states increases at a high temperature.

In evaluation of the relative stability of a protein structure against the reference structure ensemble, a difficult problem is determining what the reference structures are. In the conventional case, most engineers take the reference structures from the PDB. The resulting library of the structures is, however, insufficient to cover the whole shape space consisting of all conceivable protein structures. Therefore, it seems that the discrepancy between the theory and the simulation should not be the pending issue. Since the theoretical energy distributions are derived from the simple statistical model of proteins, we can apply the theory to the rough prediction of a protein main-chain structure into which a query amino acid sequence folds, on the assumption of the validity of the model.

In practical use, the formulae enable us to perform the high-through-put screening of sequences in the inverse folding problem. Additionally, the formulae seem to efficiently handle the problem analytically through the mean field theory (Zou & Saven,2000; Kono & Saven,2001).

Acknowledgments

This work was performed as a part of the R&D Project of the Industrial Science and Technology Frontier Program supported by NEDO (New Energy and Industrial Technology Development Organization). The knowledge-based potential and the set of main-chain structure ensemble were kindly provided from Dr. Motonori Ota (Center for Information Biology and DDBJ, National Institute of Genetics, Japan). We thank him for helpful comments on earlier drafts. We thank also Dr. Osamu Gotoh (Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Japan) for his encouragement.

References

[1] Babajide, A., Hofacker, I.L., Sippl, M.J. & Stadler, P.F. (1997). Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Folding* & *Design* **2**, 261-269.

[2] Bryngelson, J. & Wolynes, P. (1987). Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc.Natl.Acad.Sci. USA* **84**, 7524-7528.

[3] Casari, G. & Sippl, M.J. (1992). Structure-derived hydrophobic potential: Hydrophobic potentials derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725-732.

[4] Choy, W.Y. & Forman-Kay, J.D. (2001). Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **308**, 1011-1032.

[5] Isogai, Y., Ota, M., Fujisawa, T., Izuno, H., Mukai, M., Nakamura, H., Iizuka, T. & Nishikawa, K. (1999) Design and synthesis of a globin fold. *Biochemistry* **38** 7431-7443.

[6] Kono, H. & Saven, J.G. (2001). Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J. Mol. Biol.* **306**, 607-628.

[7] Miyazawa, S. & Jernigan, R.L. (1999) An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins* **36**, 357-369.

[8] Mirny, L.A., Finkelstein, A.V. & Shakhnovich, E.I. (2000) Statistical significance of protein structure prediction by threading. *Proc.Natl.Acad.Sci. USA* **97**, 9978-9983.

[9] Ota, M., Isogai, Y. & Nishikawa, K. (2001) Knowledge-based potential defined for a rotamer library to design protein sequences. *Protein Eng.* **14** 557-564.

[10] Paci, E., Smith, L.J., Dobson, C.M. & Karplus, M. (2001) Exploration of partially unfolded states of human alpha-lactalbumin by molecular dynamics simulation. *J. Mol. Biol.* **306** 329-347.

[11] Pande, V.S., Grosberg, A.Y. & Tanaka, T. (1997) Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73** 3192-3210.

[12] Zou, J. & Saven, J.G. (2000). Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J. Mol. Biol.* **296**, 281-294.

[13] Webb, D.J. (1974). The statistics of relative abundance and diversity. *J. Theor. Biol.* **43**, 277-291.

6 Appendix: Derivation of the mean and variance of the secondary structure energy over the structure ensemble

Using the following equations:

$$\begin{aligned}
\sum_{\{n_{\alpha\mathrm{s}}\}} \mathcal{M}_\alpha(\{n_{\alpha\mathrm{s}}\}|\{q_\mathrm{s}\}) = \sum_{\{n_{\alpha 1}, n_{\alpha 2}, \cdots\}} \frac{\nu_\alpha!}{\prod_\mathrm{s} n_{\alpha\mathrm{s}}!} \prod_\mathrm{s} q_\mathrm{s}^{n_{\alpha\mathrm{s}}} &= 1 \\
\sum_{\{n_{\alpha\mathrm{s}}\}} n_{\alpha\mathrm{s}}\, \mathcal{M}_\alpha(\{n_{\alpha\mathrm{s}}\}|\{q_\mathrm{s}\}) &= \nu_\alpha q_\mathrm{s} \\
\sum_{\{n_{\alpha\mathrm{s}}\}} {n_{\alpha\mathrm{s}}}^2\, \mathcal{M}_\alpha(\{n_{\alpha\mathrm{s}}\}|\{q_\mathrm{s}\}) &= \nu_\alpha q_\mathrm{s}(1-q_\mathrm{s}) + (\nu_\alpha q_\mathrm{s})^2 \\
\sum_{\{n_{\alpha\mathrm{s}}\}} n_{\alpha\mathrm{s}}\, n_{\alpha\mathrm{s}'}\, \mathcal{M}_\alpha(\{n_{\alpha\mathrm{s}}\}|\{q_\mathrm{s}\}) &= \nu_\alpha(\nu_\alpha - 1) q_\mathrm{s}\, q_{\mathrm{s}'},
\end{aligned}$$

we derived the mean and variance of the secondary structure energy over the structure ensemble as follows:

$$\begin{aligned}
\langle E \rangle &= \frac{1}{N}\sum_{i=1}^{N} \sum_{\{n_{1\mathrm{s}}\}} \cdots \sum_{\{n_{\alpha\mathrm{s}}\}} \cdots \sum_{\{n_{\lambda\mathrm{s}}\}} \sum_\alpha (\sum_\mathrm{s} \epsilon_{\alpha\mathrm{s}}\, n_{\alpha\mathrm{s}})\, \mathcal{M}_\alpha(\{n_{\alpha\mathrm{s}}\}|\{q_\mathrm{s}^{(i)}\}) \prod_{\alpha' \neq \alpha} \mathcal{M}_{\alpha'}(\{n_{\alpha'\mathrm{s}}\}|\{q_\mathrm{s}^{(i)}\}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \sum_\alpha \sum_\mathrm{s} \epsilon_{\alpha\mathrm{s}} \nu_\alpha q_\mathrm{s}^{(i)} \\
&= \sum_\alpha \nu_\alpha \sum_\mathrm{s} \epsilon_{\alpha\mathrm{s}} \langle q_\mathrm{s} \rangle
\end{aligned}$$

$$\begin{aligned}
\langle E^2 \rangle &= \frac{1}{N}\sum_{i=1}^{N} \sum_{\{n_{1\mathrm{s}}\}} \cdots \sum_{\{n_{\alpha\mathrm{s}}\}} \cdots \sum_{\{n_{\lambda\mathrm{s}}\}} \left(\sum_{\alpha\mathrm{s}} {\sigma_{\alpha\mathrm{s}}}^2 n_{\alpha\mathrm{s}} + \left(\sum_{\alpha\mathrm{s}} \epsilon_{\alpha\mathrm{s}} n_{\alpha\mathrm{s}} \right)^2 \right) \prod_\alpha \mathcal{M}_\alpha(\{n_{\alpha\mathrm{s}}\}|\{q_\mathrm{s}^{(i)}\}) \\
&= \sum_\alpha \nu_\alpha (\sum_\mathrm{s} {\sigma_{\alpha\mathrm{s}}}^2 \langle q_\mathrm{s} \rangle) + \sum_\alpha {\nu_\alpha}^2 (\sum_\mathrm{s} {\epsilon_{\alpha\mathrm{s}}}^2 \langle {q_\mathrm{s}}^2 \rangle) + \sum_\alpha \nu_\alpha (\sum_\mathrm{s} {\epsilon_{\alpha\mathrm{s}}}^2 (\langle q_\mathrm{s} \rangle - \langle {q_\mathrm{s}}^2 \rangle)) \\
&\quad + 2 \sum_\alpha {\nu_\alpha}^2 \sum_{\mathrm{s}<\mathrm{s}'} \epsilon_{\alpha\mathrm{s}} \epsilon_{\alpha\mathrm{s}'} \langle q_\mathrm{s} q_{\mathrm{s}'} \rangle - 2 \sum_\alpha \nu_\alpha \sum_{\mathrm{s}<\mathrm{s}'} \epsilon_{\alpha\mathrm{s}} \epsilon_{\alpha\mathrm{s}'} \langle q_\mathrm{s} q_{\mathrm{s}'} \rangle + 2 \sum_{\alpha<\alpha'} \nu_\alpha \nu_{\alpha'} \sum_{\mathrm{s}<\mathrm{s}'} \epsilon_{\alpha\mathrm{s}} \epsilon_{\alpha'\mathrm{s}'} \langle q_\mathrm{s} q_{\mathrm{s}'} \rangle
\end{aligned}$$

$$\langle E \rangle^2 = \sum_\alpha {\nu_\alpha}^2 (\sum_\mathrm{s} {\epsilon_{\alpha\mathrm{s}}}^2 \langle q_\mathrm{s} \rangle^2) + 2 \sum_\alpha {\nu_\alpha}^2 \sum_{\mathrm{s}<\mathrm{s}'} \epsilon_{\alpha\mathrm{s}} \epsilon_{\alpha\mathrm{s}'} \langle q_\mathrm{s} \rangle \langle q_{\mathrm{s}'} \rangle + 2 \sum_{\alpha<\alpha'} \nu_\alpha \nu_{\alpha'} \sum_{\mathrm{s}<\mathrm{s}'} \epsilon_{\alpha\mathrm{s}} \epsilon_{\alpha'\mathrm{s}'} \langle q_\mathrm{s} \rangle \langle q_{\mathrm{s}'} \rangle$$

$$\begin{aligned}
V[E] &= \langle E^2 \rangle - \langle E \rangle^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} \Bigg(\sum_\alpha {\nu_\alpha}^2 (\sum_\mathrm{s} \epsilon_{\alpha\mathrm{s}} (q_\mathrm{s}^{(i)} - \langle q_\mathrm{s} \rangle))^2 + 2 \sum_{\alpha<\alpha'} \nu_\alpha \nu_{\alpha'} (\sum_\mathrm{s} \epsilon_{\alpha\mathrm{s}} (q_\mathrm{s}^{(i)} - \langle q_\mathrm{s} \rangle))(\sum_\mathrm{s} \epsilon_{\alpha'\mathrm{s}} (q_\mathrm{s}^{(i)} - \langle q_\mathrm{s} \rangle)) \\
&+ \sum_\alpha \nu_\alpha \left(\sum_\mathrm{s} ({\sigma_{\alpha\mathrm{s}}}^2 + {\epsilon_{\alpha\mathrm{s}}}^2) q_\mathrm{s}^{(i)} - \left(\sum_\mathrm{s} \epsilon_{\alpha\mathrm{s}} q_\mathrm{s}^{(i)} \right)^2 \right) \Bigg)
\end{aligned}$$

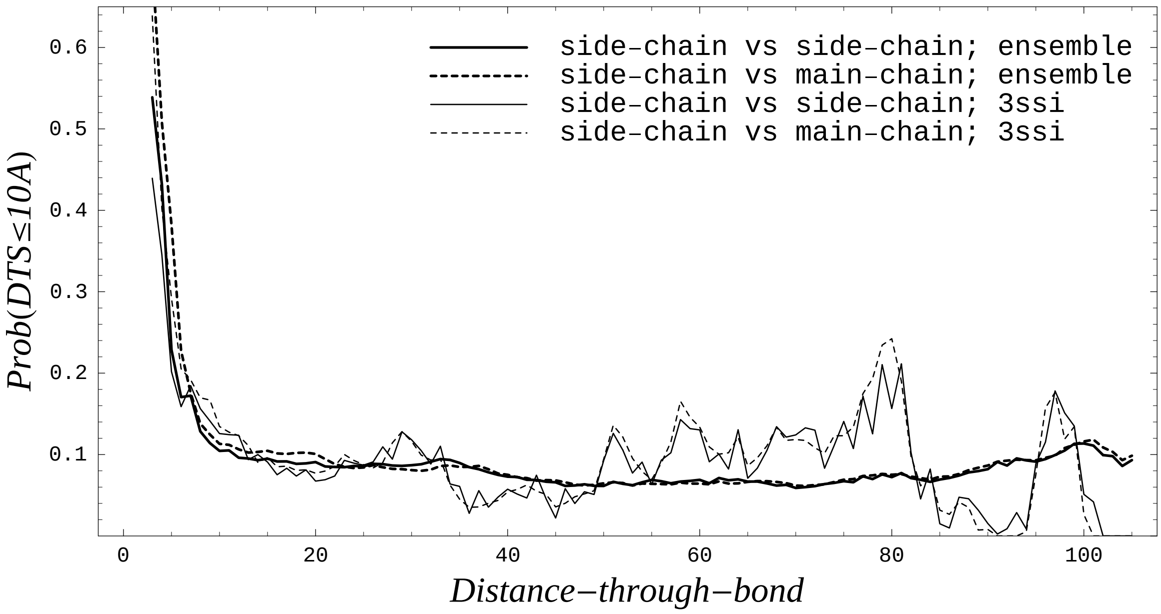

Figure 1: **The conditional probability that the distance-through-space (DTS) is less than 10Å when the distance-through-bond is given.** The solid lines are for the pairs of different side-chain centers. The dashed lines are for the pairs of side-chain centers and main-chain atoms. The thin line is for the main-chain structure of streptomyces subtilisin inhibitor (PDB code is 3ssi). The thick line is the result from the ensemble of 159 main-chain structures defined in the text.

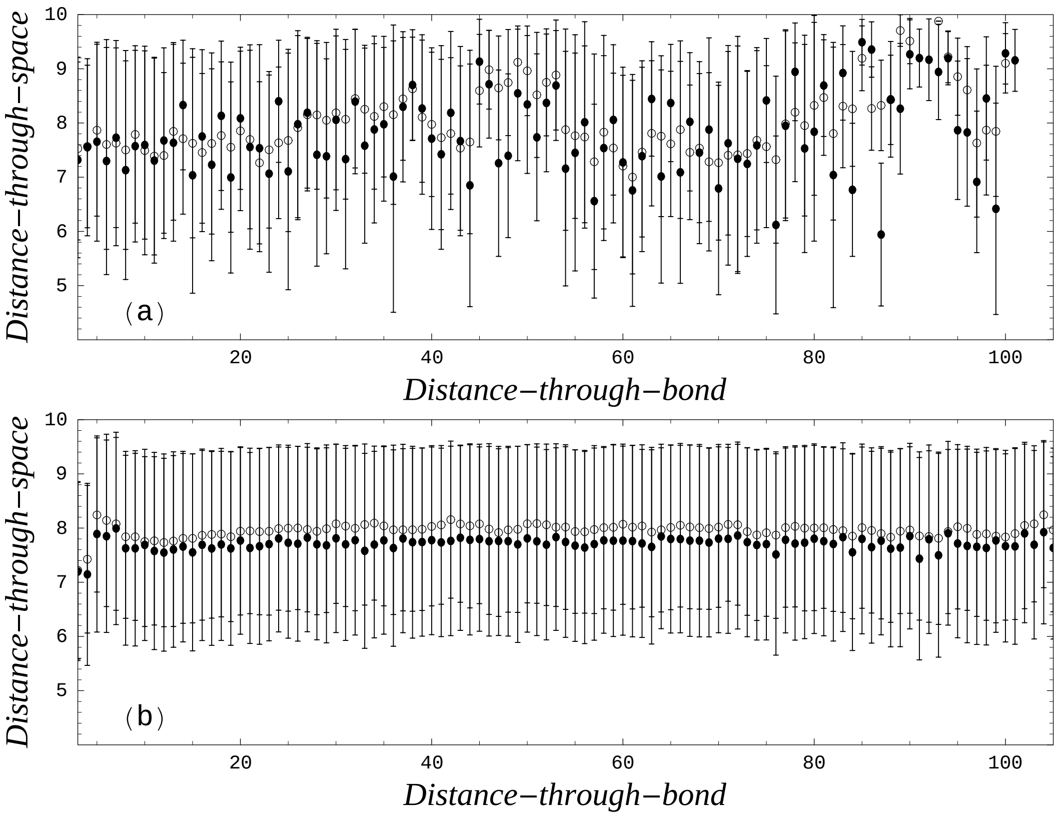

Figure 2: **The distributions of the distance-through-space (Å) for the pairs whose distance-through-space is less than 10Å.** The filled circles represent the mean for the pairs of different side-chain centers. The empty circles represent the mean for the pairs of side-chain centers and main-chain atoms. The error bars represent the standard deviations. (a) for the main-chain structure of streptomyces subtilisin inhibitor (PDB code is 3ssi). (b) over the structure ensemble defined in the text.

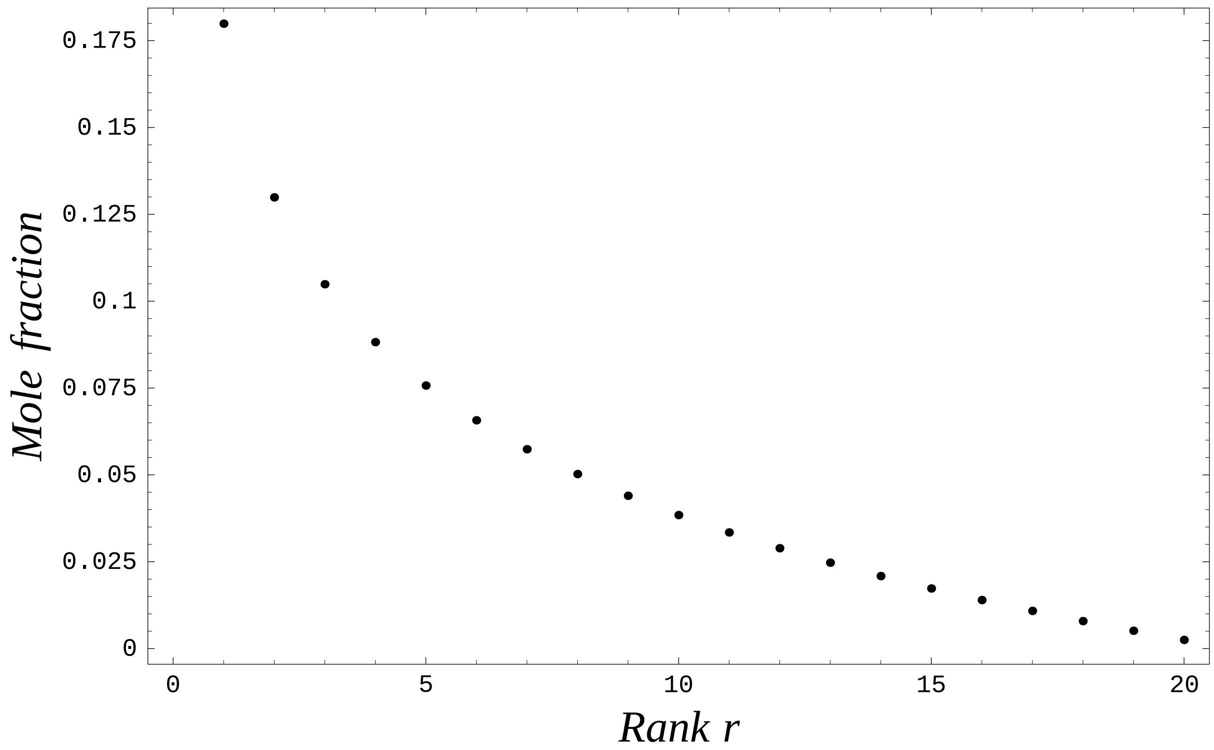

Figure 3: **The most expected mole-fraction of amino acids.** The abscissa is the rank of percentage of amino acid composition in descending order. The ordinate is the mole fraction, which is calculated from eqn. (18) with $\lambda = 20$.

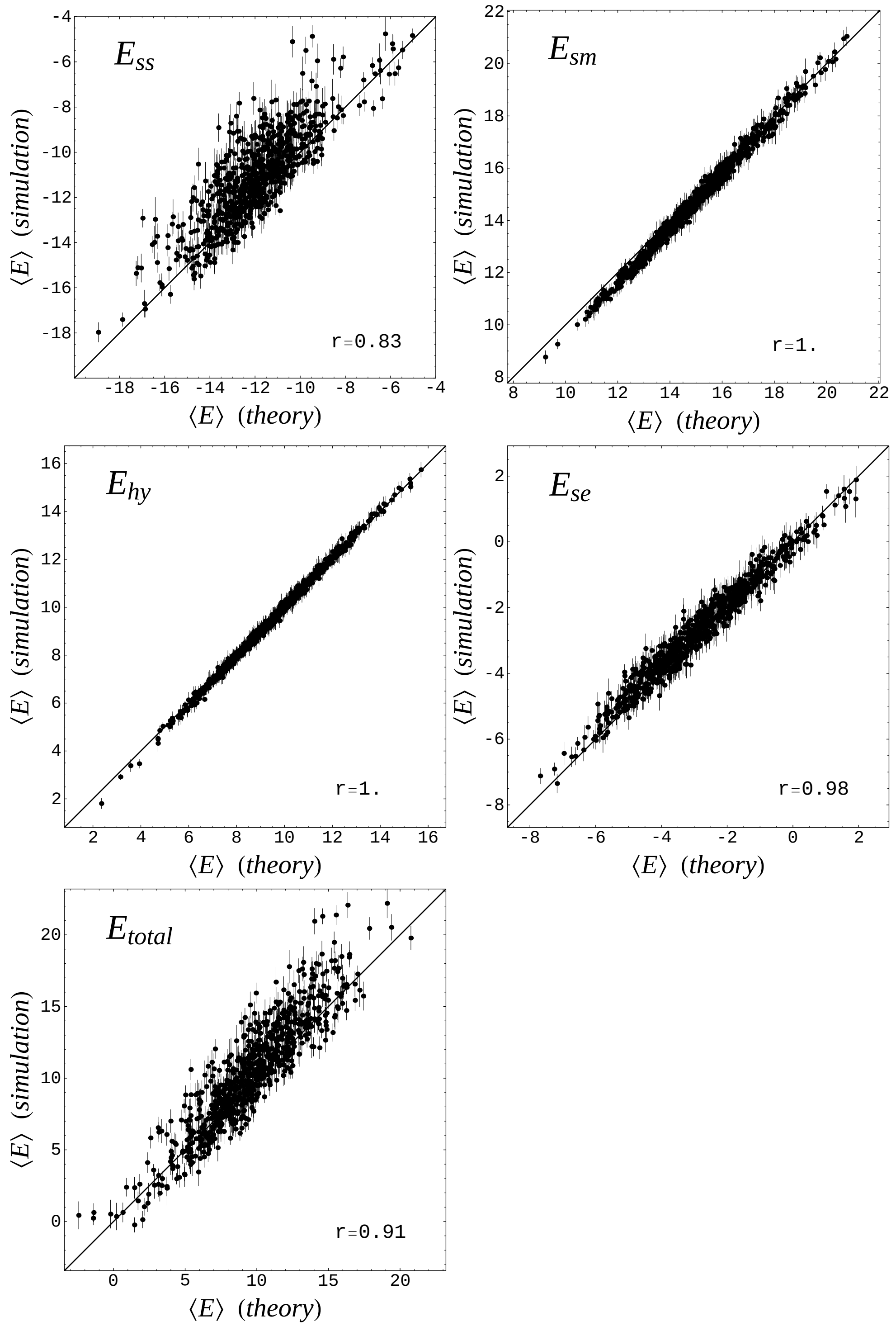

Figure 4: **Correlation of the ensemble mean of the reference structural energy, $\langle E \rangle$, derived from the theory and that from the gapless threading simulation.** The dots with error bars represent the average and standard deviation of $\langle E \rangle$ over 10 different sequences for each of the 760 amino acid compositions, which were randomly generated and roughly obey eqn. (18). The theoretical values were calculated from eqn. (12) for E_{ss} (and for E_{sm}), eqn. (14) for E_{hy} and eqn. (16) for E_{se}.

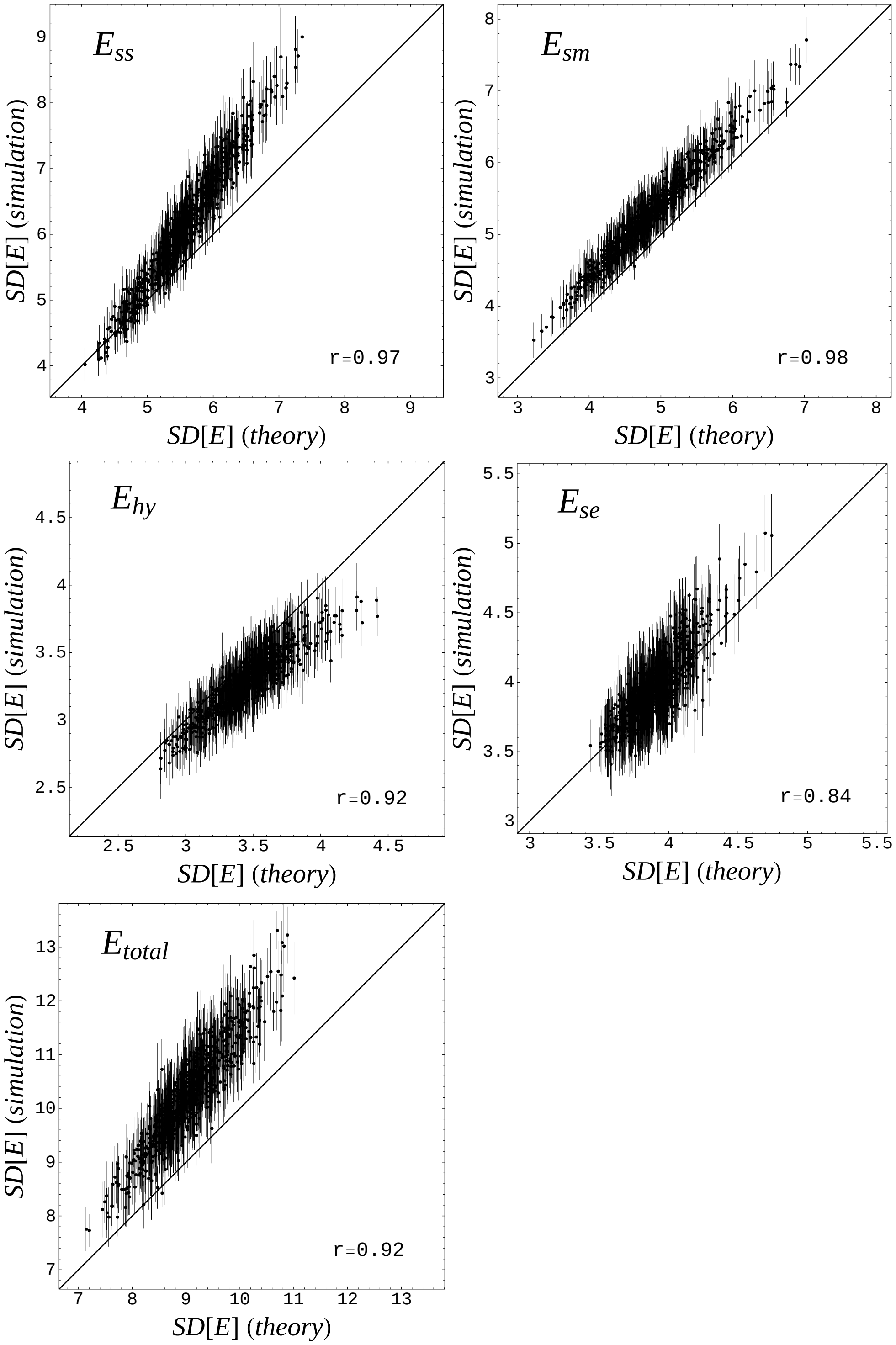

Figure 5: **Correlation of the ensemble standard deviation of the reference structural energy, $SD[E]$, derived from the theory and that from the gapless threading simulation.** The dots with error bars represent the average and standard deviation of $SD[E]$ over 10 different sequences for each of the 760 amino acid compositions, which were adopted in Fig. 4. The theoretical values were calculated from eqns. (12)-(17).

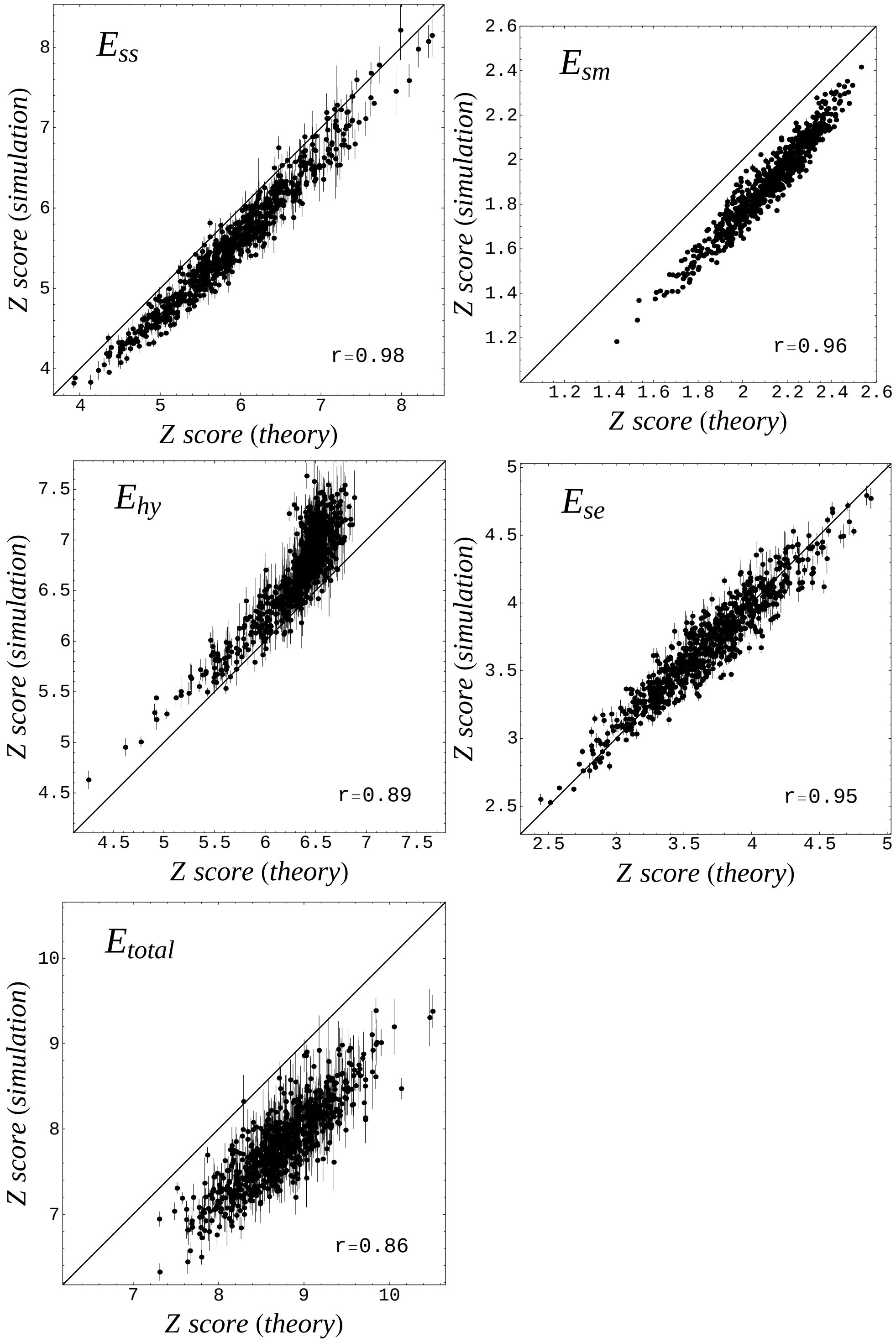

Figure 6: **Correlation of the Z-score derived from the theory and that from the gapless threading simulation.** The dots with error bars represent the average and standard deviation of the Z-score over 10 different sequences for each of the 760 amino acid compositions, which were adopted in Fig. 4. The theoretical values were calculated from eqns. (12)-(17). The details are shown in the text.