Long- and short-range interactions

in native protein structures

are consistent/minimally-frustrated in sequence space

Sanzo Miyazawa

Gunma University, Faculty of Technology

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

presented at

The 3rd Annual Meeting of the Protein Science Society of Japan

held in June 23rd to 25th of 2003, Sapporo, Japan.

(June 25, 2003)

ABSTRACT

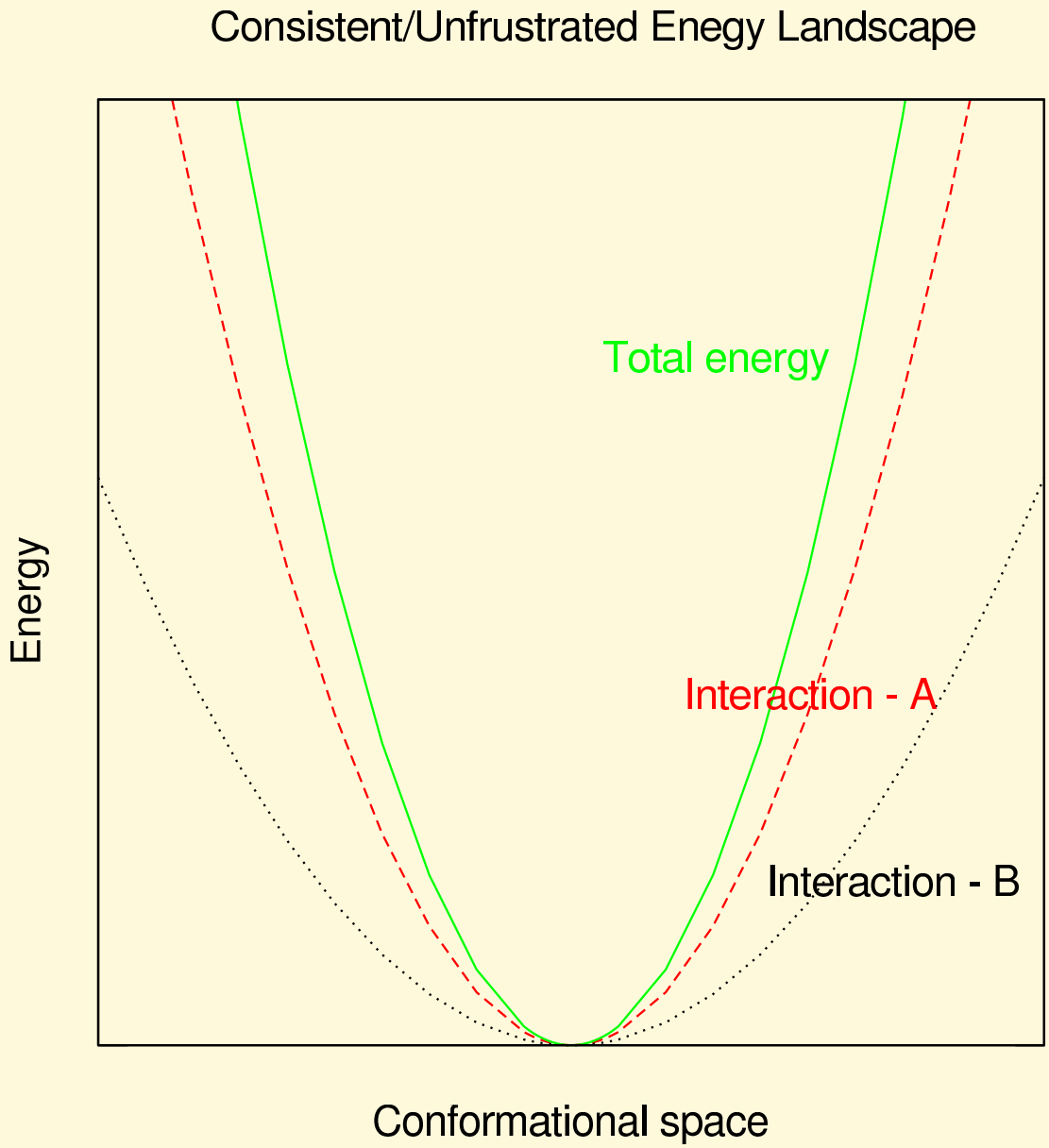
We show that long- and short-range interactions in almost all protein native structures are actually consistent with each other for coarse-grained energy scales; specifically we mean the long-range inter-residue contact energies and the short-range secondary structure energies based on peptide dihedral angles, which are potentials of mean force evaluated from residue distributions observed in protein native structures. This consistency is observed at equilibrium in sequence space rather than in conformational space. Statistical ensembles of sequences are generated by exchanging residues for each of 797 protein native structures with the Metropolis method. It is shown that adding the other category of interaction to either the short- or long-range interactions decreases the means and variances of those energies for essentially all protein native structures, indicating that both interactions consistently work by more-or-less restricting sequence spaces available to one of the interactions. In addition to this consistency, independence between these interaction classes is also indicated by the fact that there are almost no correlations between them when equilibrated using both interactions and significant but small, positive correlations at equilibrium using only one of the interactions. Evidence is provided that protein native sequences can be regarded approximately as samples from the statistical ensembles of sequences with these energy scales, and that all proteins have the same effective conformational temperature. Designing protein structures and sequences to be consistent and minimally-frustrated among the various interactions is the most effective way to increase protein stability and foldability.

Reference: Sanzo Miyazawa & R. L. Jernigan, *Proteins* **50**:35-43, 2003.

1. INTRODUCTION

A principle of consistency among various interactions in protein native structures

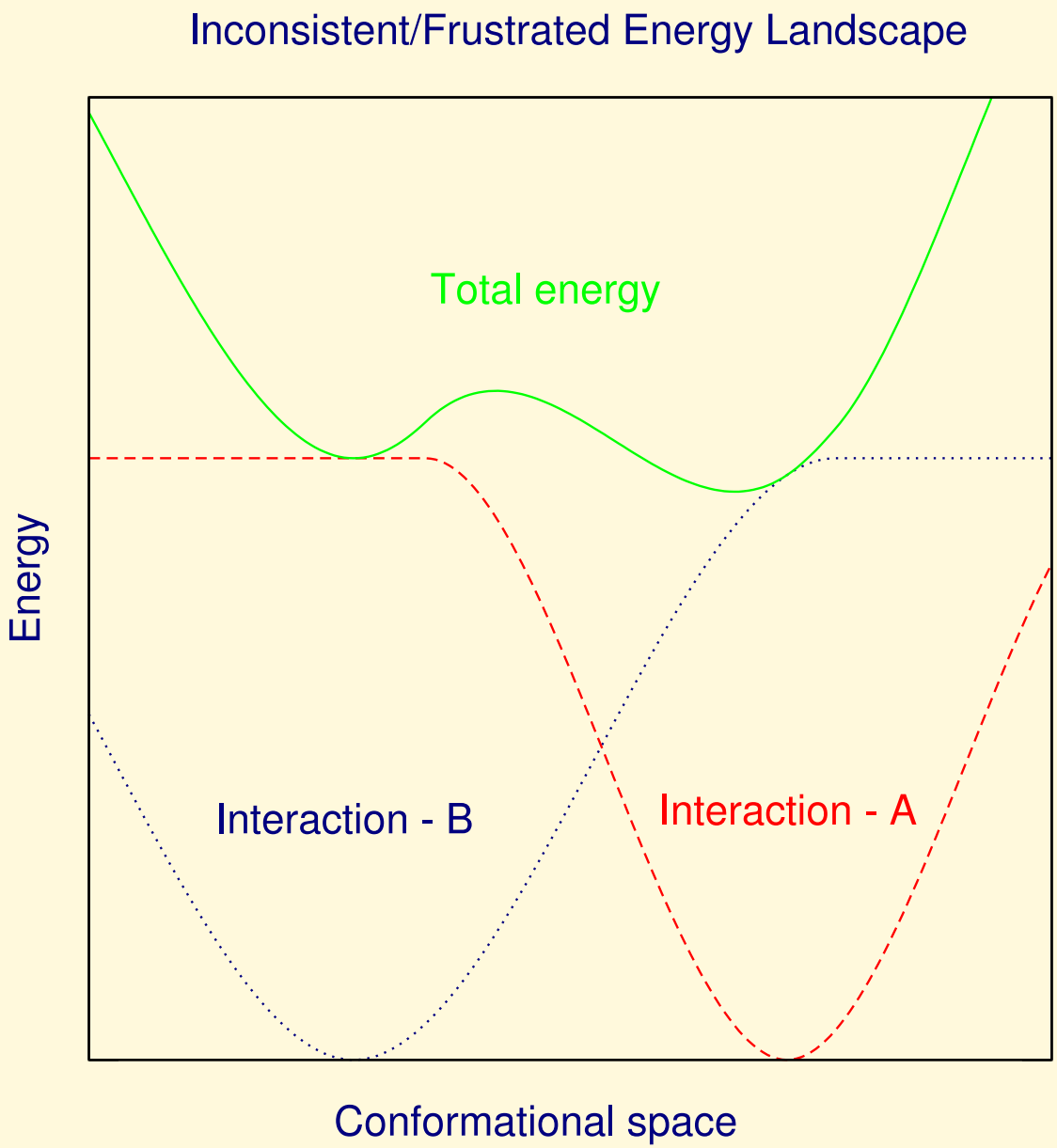
- It was proposed by Go (1983) from the fact that various prediction methods of protein secondary structure based solely on short-range interactions are fairly successful, even though long-range interactions are essential to fold protein structures,
- The consistency among interactions is an effective way for proteins to increase structural stabilities.


Consistent/Unfrustrated Enegy Landscape
Total energy
Interaction - A
Interaction - B
Energy
Conformational space

A principle of minimal frustration in the energy landscape of natural proteins

Bryngelson & Wolynes (1987) pointed out that:

- A rough energy landscape, a frustrated situation which is caused by many competing interactions, is a characteristic of random copolymers and often causes glass transitions.
- The energy landscape for natural proteins must be minimally frustrated between smooth and rough energy landscapes and must resemble funnels for proteins to fold into single stable structures within a limited time.

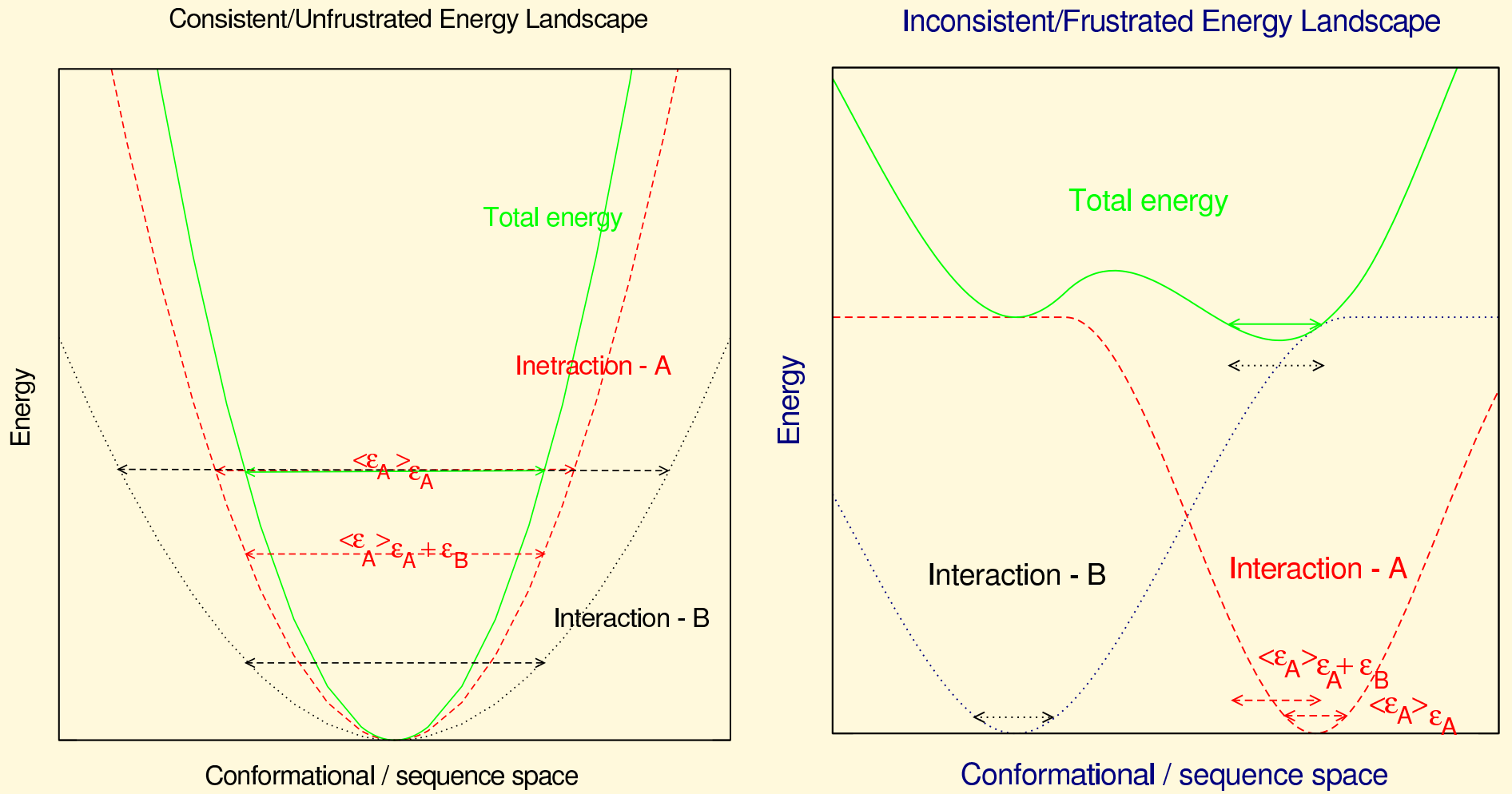

Inconsistent/Frustrated Energy Landscape
Total energy
Energy
Interaction - B
Interaction - A
Conformational space

The consistency and minimal frustration among various interactions in protein native structures are essential for the stability and foldability of protein structures.

Here, we show short- and long-range interactions between residues in coarse-grained energy scales are consistent with each other for sequence selection of the more stable sequences for each protein.

- Interaction potentials used here are potentials of mean force evaluated from residue distributions observed in protein native structures:
  - ⋆ the long-range contact potentials between the 20 kinds of amino acids (Miyazawa & Jernigan, 1985, 1996, 1999) and
  - ⋆ the short-range secondary structure potentials based on peptide dihedral angles (Miyazawa & Jernigan, 1999).
- Sequence space is searched instead of conformational space by exchanging amino acids within each protein.
  - ⋆ Evidence is provided that protein native sequences can be regarded approximately as samples from the statistical ensembles of sequences with these energy scales.

Consistencies between short- and long-range interactions are examined for their effects on the mean and the variance of interaction energies at statistical equilibrium in sequence space.



2. METHODS

Stability of protein sequence and structure

An effective free energy, $\mathcal{F}$, which represents the stability of a sequence - structure pair, i.e., probability $P(s,i)$ of a specific conformation s for sequence i:

$$\beta\mathcal{F}(s|i) \equiv -\log(P(s|i)) \quad (1)$$

$$= \beta E^{conf}(s,i) + \log(\sum_{s} \exp(-\beta E^{conf}(s,i))) \quad (2)$$

where

β	$\equiv 1/(kT)$,
$E^{conf}(s,i)$	is the conformational energy of the conformational state s of sequence i.
$\sum_s$	is taken over all possible conformations.

The contribution from the partition function is approximated by assuming the condition under which native-like conformations are dominant:

$$
\begin{aligned}
&\log(\sum_{s} \exp(-\beta E^{conf}(s,i))) \\
&\simeq \log(\sum_{s\in\{\text{native-like}\}} \exp(-\beta E^{conf}(s,i))) \\
&\simeq \log(\sum_{s\in\{\text{native-like}\}} 1) - \beta \sum_{s\in\{\text{native-like}\}} E^{conf}(s,i) / (\sum_{s\in\{\text{native-like}\}} 1) \\
&\simeq n_r\sigma - \beta \langle E^{conf}(s,i)\rangle_{\beta=0}, \text{ native-like conf.}
\end{aligned}
\tag{3}
$$

where

n_r is the sequence length.

σ is a constant to represent the conformational entropy per residue in k units for native-like structures.

Coarse-grained conformational energy

- **Secondary structure energy:**

$$E^s(s,i) = \sum_p \delta e^s(s_{p-1}, i_p, s_p, s_{p+1}) \tag{4}$$

$$\langle E^s(s,i)\rangle_{\beta=0,\text{native-like conf.}} \simeq \sum_p \langle \delta e^s(s_{p-1}, i_p, s_p, s_{p+1})\rangle_{\text{all natives}} \tag{5}$$

where $\delta e^s(s_{p-1}, i_p, s_p, s_{p+1})$ is the interaction energy between the side chian of i_p type and the tripeptide of conformational state (s_{p-1}, s_p, s_{p+1}); p indicates a residue position. s_p is one of α, β, pro-β, L-α, L-β.

- **Pairwise contact energy:**

$$E^c(s,i) = \frac{1}{2}\sum_p \sum_j n^c_{i_p j}(e_{i_p j} - e_{rr}) \tag{6}$$

$$\langle E^c(s,i)\rangle_{\beta=0,\text{native-like conf.}} \simeq \frac{1}{2}\sum_p \langle n^c_{i_p j}\rangle_{\text{all natives}}(e_{i_p j} - e_{rr}) \tag{7}$$

where $e_{i_p j}$ is a contact energy between residues of i_p and j types, e_{rr} is a collapse energy independent of residue type, and $n^c_{i_p j}$ is the number of contacts between residues of i and j types at pth residue.

Here, we consider only sequences having the same amino acid composition as the native sequence.

Statistical ensemble of sequences

The conditional probabilities $P(i|s)$ of sequences i for a given structure s :

$$P(i|s) = P(s|i)P(i)/\sum_i P(s|i)P(i) \quad (8)$$

$$P(i) = \text{constant} \quad (9)$$

where

$P(s\|i)$	is the probability of a specific conformation s for sequence i.
$P(i)$	is the $a\ priori$ probability for sequence i.
$\sum_i$	means the sum over all sequences with fixed length for a given structure; here, we consider only sequences having the same amino acid composition as the native sequence.

Thus, $P(i|s)$ is represented as:

$$P(i|s) = \frac{1}{\mathcal{Z}} \exp(-\beta \mathcal{E}(s,i)) \tag{10}$$

$$\mathcal{Z} \equiv \sum_i \exp(-\beta \mathcal{E}(s,i)) \tag{11}$$

$$\beta \mathcal{E}(s,i) \equiv \beta E^{conf}(s,i) - \beta \langle E^{conf}(s,i) \rangle_{\beta=0}, \text{ native-like conf.} \tag{12}$$

where

$\mathcal{Z}$	is a partition function for the ensemble of sequences
$\mathcal{E}(s,i)$	is the conformational energy relative to the average over native-like conformations.

Monte Carlo simulations to generate the statistical ensemble of sequences

- 100,000 residue exchanges per residue are tried in each protein with the Metropolis method.
- The conformational temperature $1/\beta$ is always taken to be one; so that the sum of the equilibrium distributions over all proteins are close to those observed in their native structures.

Datasets of protein structures used

- Proteins which belong to class 1 to 5 in Release 1.53 of the SCOP have been used.
- Only structures better than 2.5 Ådetermined by X-ray are used.
- Species representatives of 2129 proteins were used to estimate the statistical potentials.
- Family representatives of 797 proteins are used to analyze the statistical ensembles of sequences.

Notations for statistical averages which are calculated in the present analyses:

$$\langle X \rangle_Y \equiv \frac{1}{\mathcal{Z}(Y)} \sum_i X(s,i) \exp(-\beta Y(s,i)) \tag{13}$$
$$\mathcal{Z}(Y) \equiv \sum_i \exp(-\beta Y(s,i))$$

For example,

$$\langle \mathcal{E}^c \rangle_{\mathcal{E}^s+\mathcal{E}^c} \equiv \frac{1}{\mathcal{Z}} \sum_i \mathcal{E}^c(s,i) \exp(-\beta(\mathcal{E}^s(s,i) + \mathcal{E}^c(s,i))) \tag{14}$$
$$\langle (\Delta\mathcal{E}^c)^2 \rangle_{\mathcal{E}^s+\mathcal{E}^c} \equiv \frac{1}{\mathcal{Z}} \sum_i (\Delta\mathcal{E}^c(s,i))^2 \exp(-\beta(\mathcal{E}^s(s,i) + \mathcal{E}^c(s,i))) \tag{15}$$

where

$$\Delta\mathcal{E} \equiv \mathcal{E} - \langle \mathcal{E} \rangle \tag{16}$$

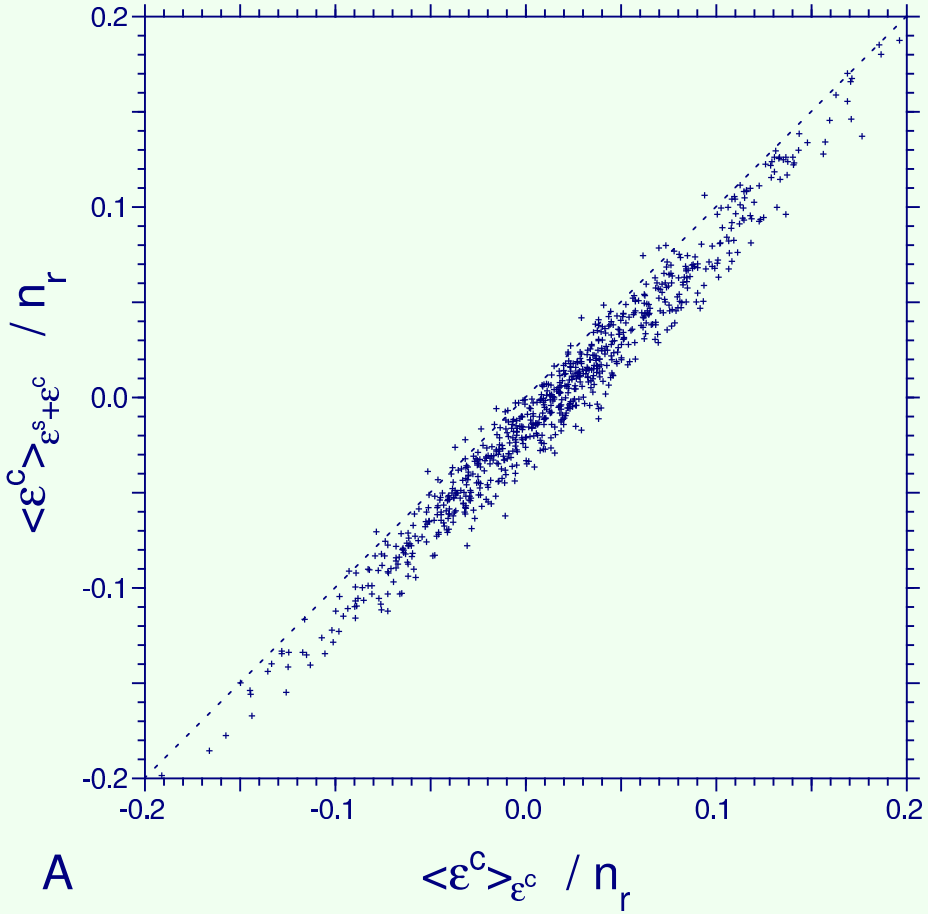
3. RESULTS

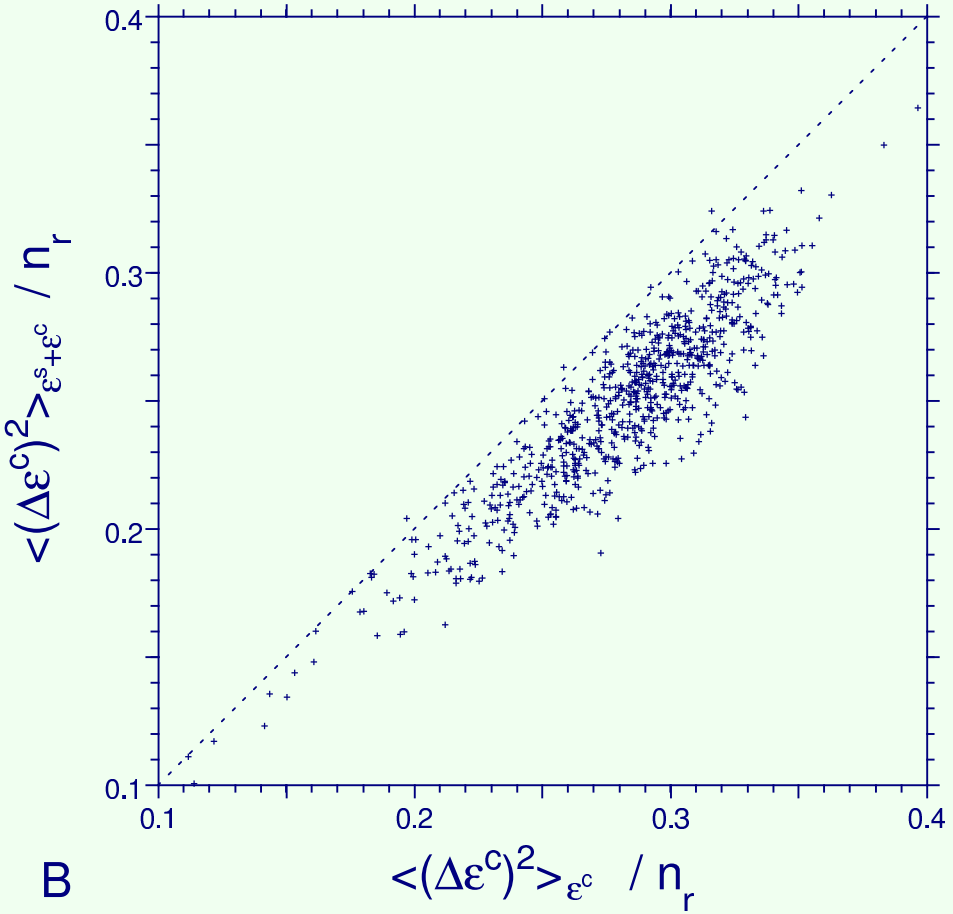
- $\langle \mathcal{E}^c \rangle_{\mathcal{E}^c} > \langle \mathcal{E}^c \rangle_{\mathcal{E}^s + \mathcal{E}^c}$

  for almost all proteins indicates that both classes of interactions are consistent with each other.

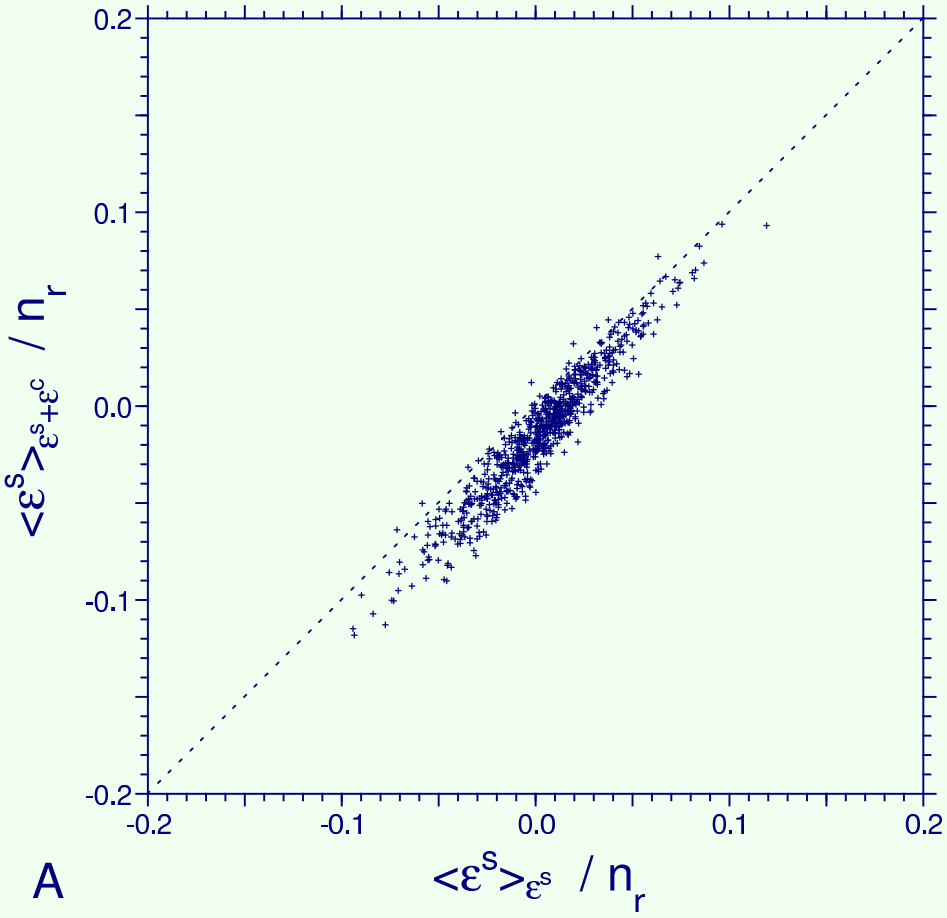
- $\langle (\Delta \mathcal{E}^c)^2 \rangle_{\mathcal{E}^c} > \langle (\Delta \mathcal{E}^c)^2 \rangle_{\mathcal{E}^s + \mathcal{E}^c}$

  for almost all proteins indicates that one class of interactions tend to reduce the available range of conformational space for the other class of interactions.

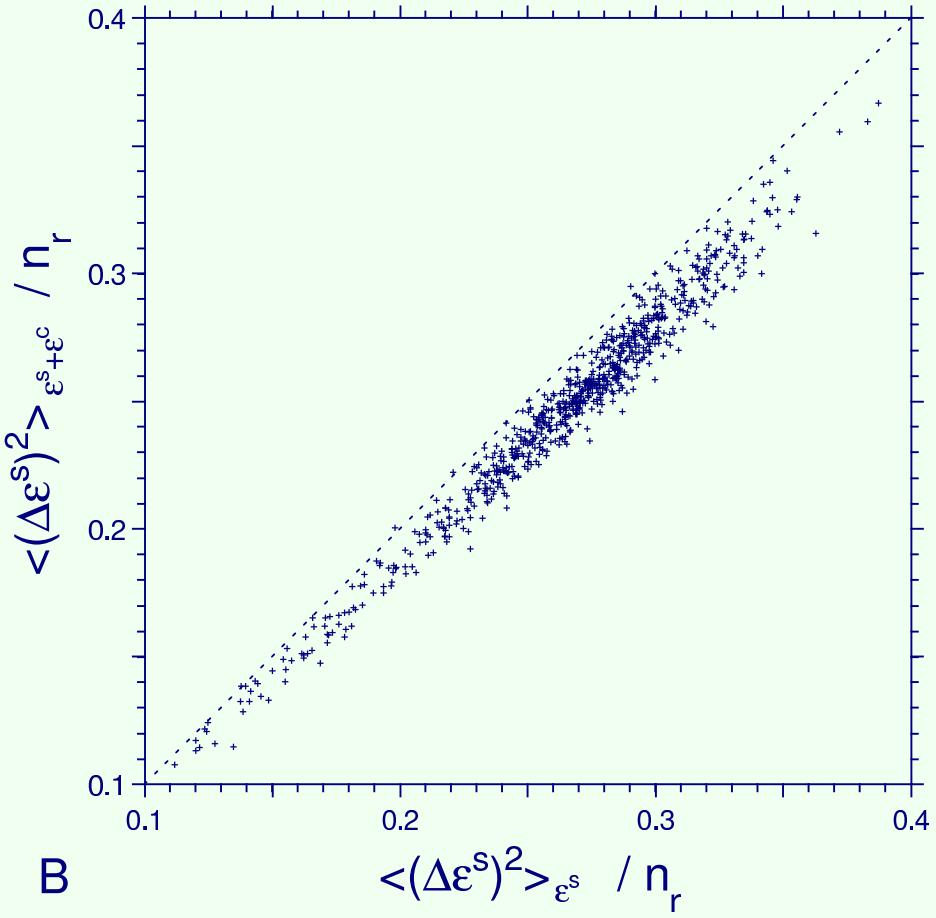




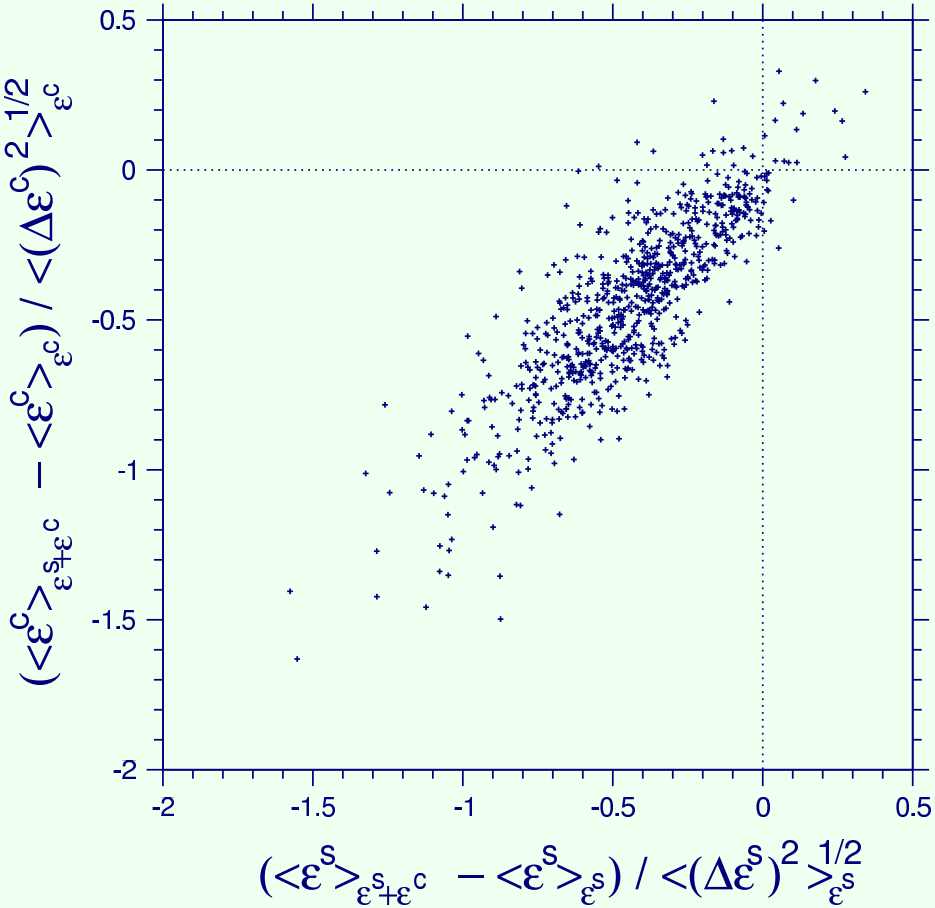
- $\langle \mathcal{E}^s \rangle_{\mathcal{E}^s} \quad > \quad \langle \mathcal{E}^s \rangle_{\mathcal{E}^s + \mathcal{E}^c}$



- $\langle (\Delta \mathcal{E}^s)^2 \rangle_{\mathcal{E}^s} \quad > \quad \langle (\Delta \mathcal{E}^s)^2 \rangle_{\mathcal{E}^s + \mathcal{E}^c}$



The decreases in the mean energies of one class by adding the other class of interactions range from 0 to -1 s.d. for both classes of interactions



Covariances between contact energies and secondary structure energies

Relation between the covariances and the increments of mean energies due to the change of interactions:

$$\int_0^1 \frac{\partial \langle \mathcal{E}^c \rangle_{x\mathcal{E}^s+\mathcal{E}^c}}{\partial x} dx = -\beta \int_0^1 \langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{x\mathcal{E}^s+\mathcal{E}^c} dx \tag{17}$$

$$\int_0^1 \frac{\partial \langle \mathcal{E}^s \rangle_{\mathcal{E}^s+y\mathcal{E}^c}}{\partial y} dy = -\beta \int_0^1 \langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^s+y\mathcal{E}^c} dy \tag{18}$$

- $\langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^s+\mathcal{E}^c} \sim 0$

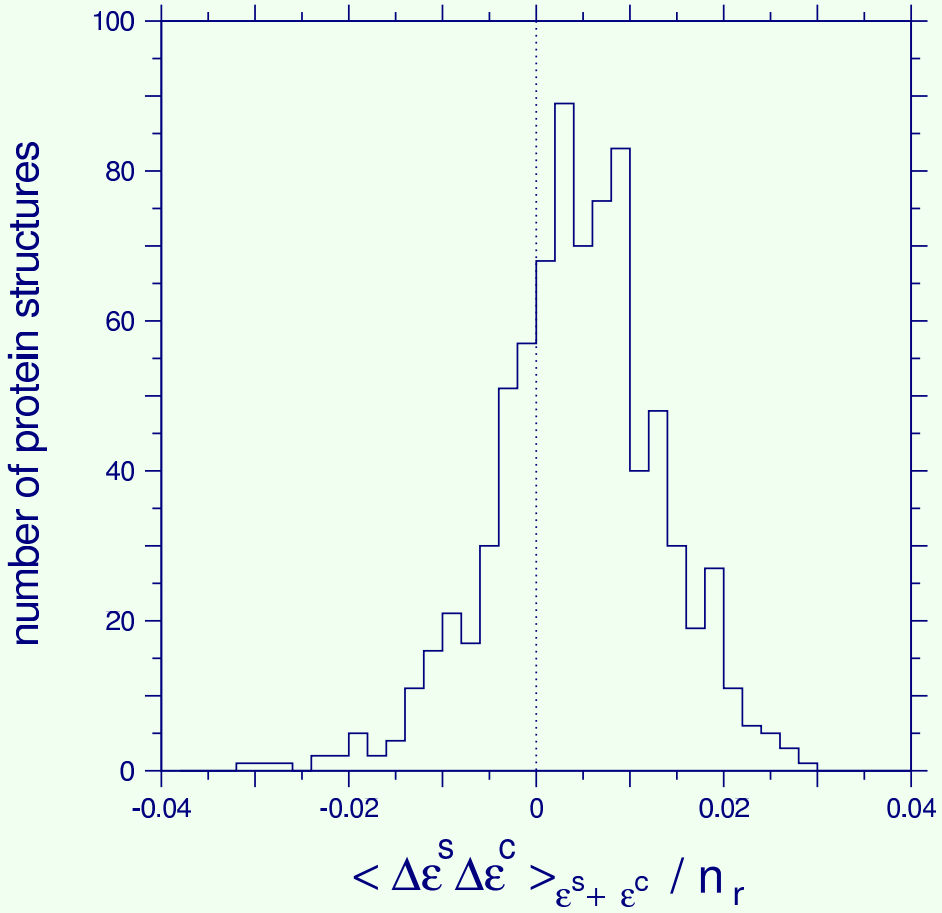
  for almost all proteins.

- $\langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^s} > 0 \ , \ \langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^c} > 0$

  for almost all proteins.

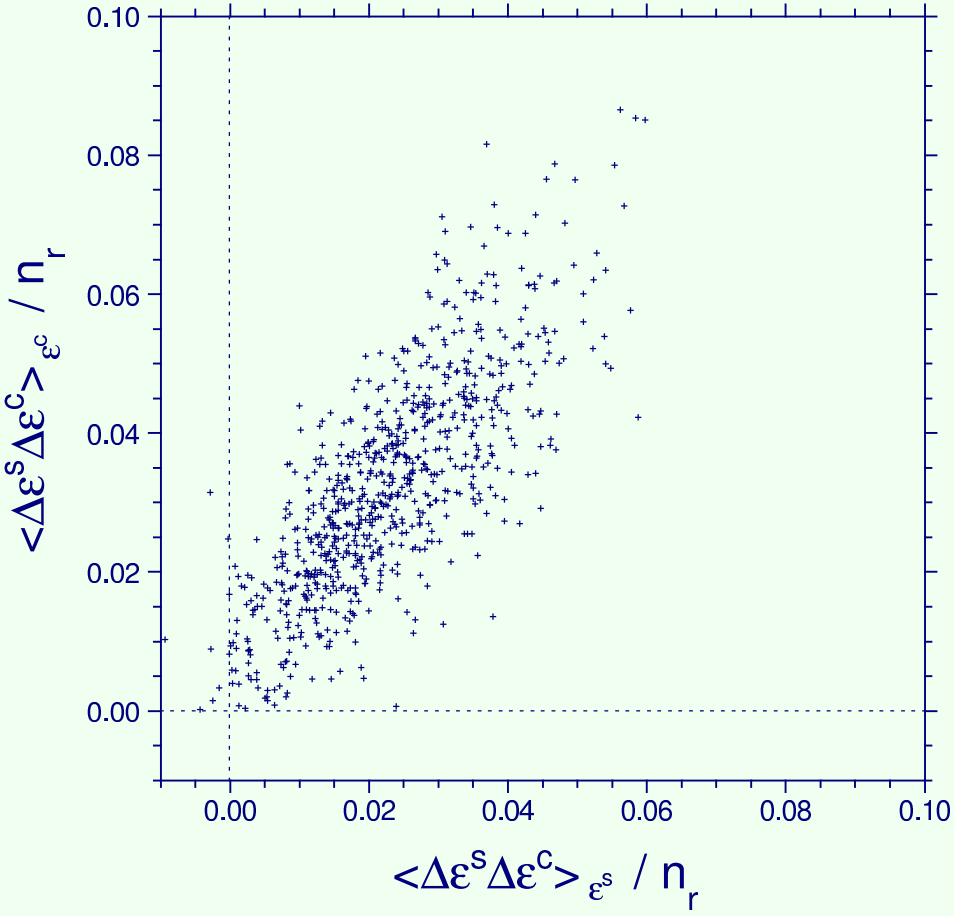
- $\langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^s+\mathcal{E}^c} \sim 0$

  for almost all proteins.



- $\langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^s} > 0$, $\langle \Delta\mathcal{E}^s \Delta\mathcal{E}^c \rangle_{\mathcal{E}^c} > 0$

  for almost all proteins.

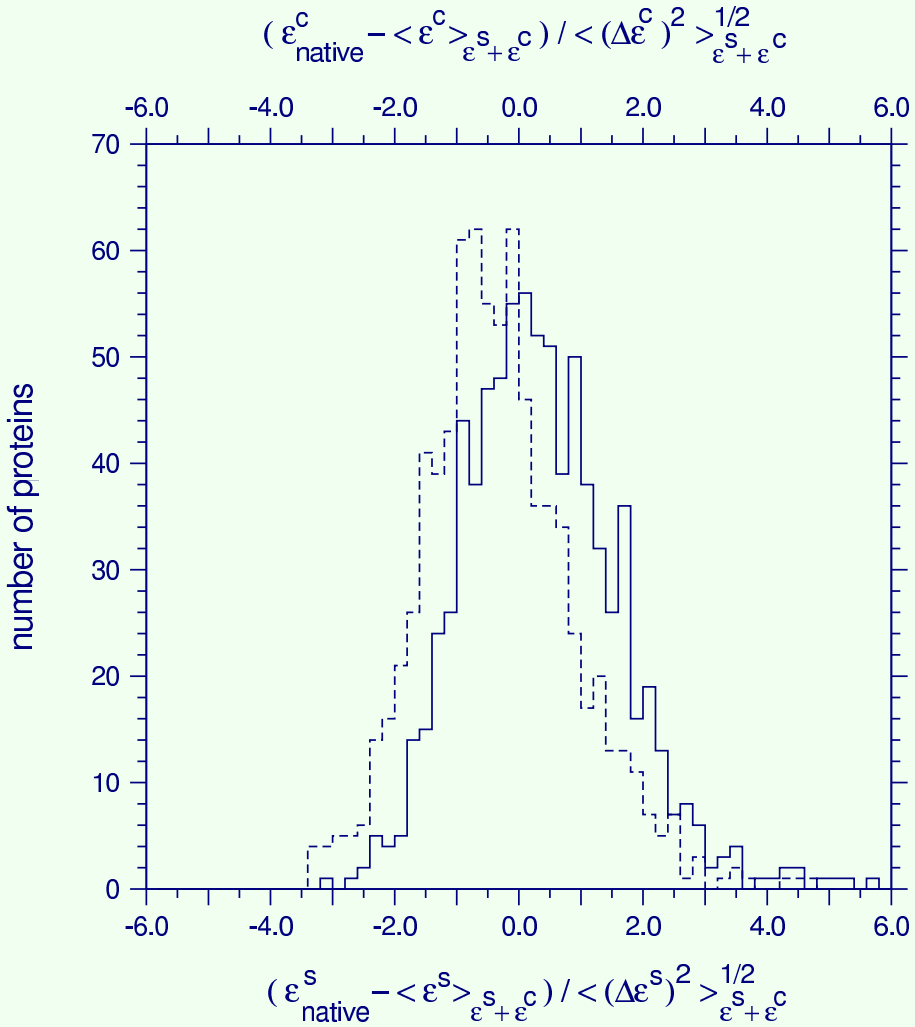


Native protein sequences are regarded as samples at equilibrium in sequence space.

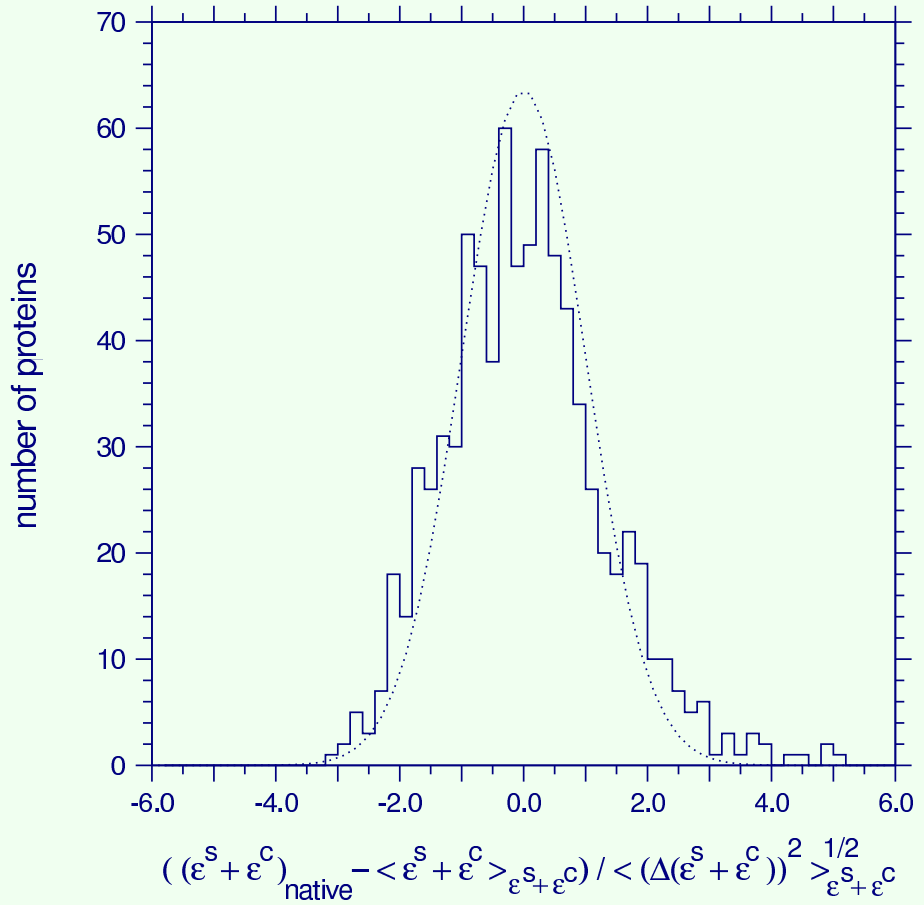
- The total contact frequencies between the 20 kinds of amino acids observed in many protein native structures can be regarded with small relative errors ($\langle 10$ %) as contact frequencies at statistical equilibrium in sequence space (Miyazawa&Jernigan, 1999).
- Here it is shown that contact energies and secondary structure energies of most native proteins lie mostly within the statistical fluctuations around equilibrium in sequence space, and that there is no correlation between the deviations of both native energies from their statistical averages.

The frequency distribution for the total energies of native proteins is similar to a Gaussian distribution.

- For the contact and secondary structure energies

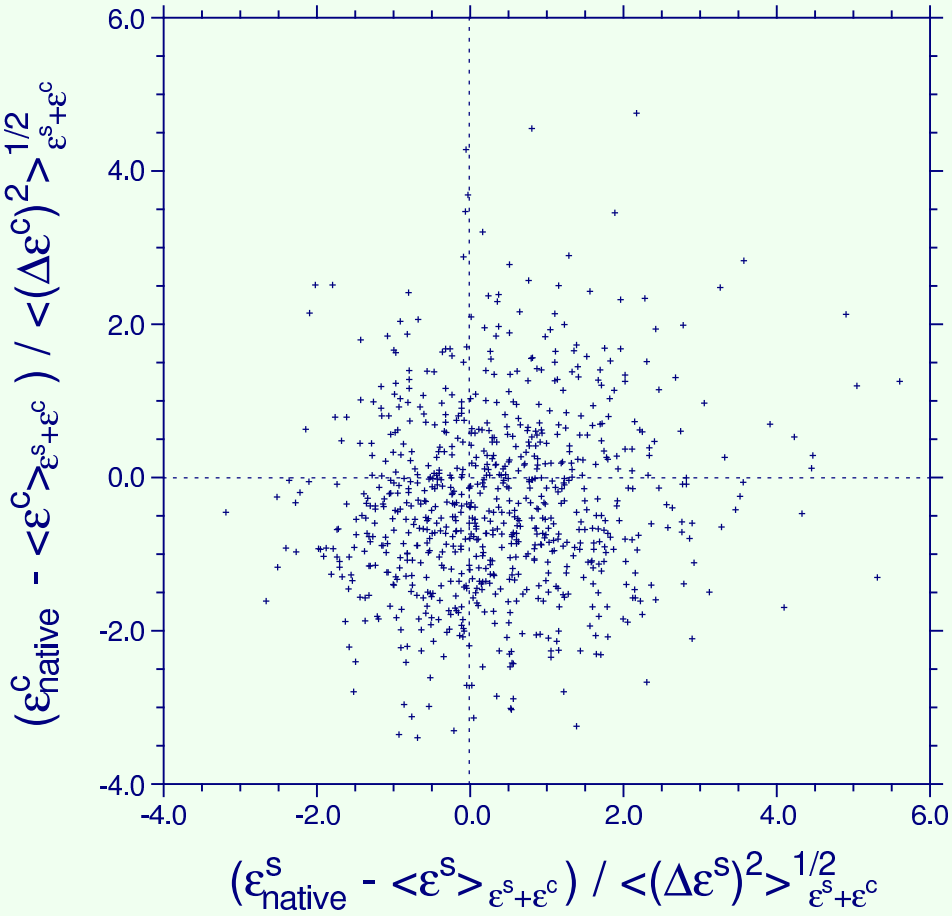


- For the total energies



There is clearly no correlation between the deviations of secondary structure and contact energies for each native protein from their statistical averages.

$\Longrightarrow$ All proteins have the same conformational temperature.



4. CONCLUSIONS

- **Short-range secondary structure interactions and long-range contact interactions in coarse-grained energy potentials are consistent/minimally-frustrated with each other for a statistical equilibrium with residue exchanges in protein sequences.**

  Proteins must have achieved these unique characteristics of smoothing the energy landscape on a coarse-grained conformational scale over the course of molecular evolution.

- **Protein native sequences can be regarded approximately as samples from equilibrium ensembles of sequences with these energy scales, and in addition all proteins have the same conformational temperature.**