Evaluation of Short-Range Interactions as Secondary Structure Energies for Protein Fold and Sequence Recognition

Sanzo Miyazawa[1,2]* **and Robert L. Jernigan**[2]
[1]*Faculty of Technology, Gunma University, Kiryu, Gunma, Japan*
[2]*Laboratory of Experimental and Computational Biology, DBS, National Cancer Institute, National Institutes of Health, Bethesda, Maryland*

ABSTRACT **Short-range interactions for secondary structures of proteins are evaluated as potentials of mean force from the observed frequencies of secondary structures in known protein structures which are assumed to have an equilibrium distribution with the Boltzmann factor of secondary structure energies. A secondary conformation at each residue position in a protein is described by a tripeptide, including one nearest neighbor on each side. The secondary structure potentials are approximated as additive contributions from neighboring residues along the sequence. These are part of an empirical potential to provide a crude estimate of protein conformational energy at a residue level. Unlike previous works, interactions are decoupled into intrinsic potentials of residues, potentials of backbone-backbone interactions, and of side chain-backbone interactions. Also interactions are decoupled into one-body, two-body, and higher order interactions between peptide backbone and side chain and between backbones. These decouplings are essential to correctly evaluate the total secondary structure energy of a protein structure without overcounting interactions. Each interaction potential is evaluated separately by taking account of the correlation in the amino acid order of protein sequences. Interactions among side chains are neglected, because of the relatively limited number of protein structures. Proteins 1999;36:347–356.**
Published 1999 Wiley-Liss, Inc.[†]

Key words: empirical potentials; protein inverse folding; protein sequence threading; statistical potentials

INTRODUCTION

The vast conformational space available to proteins has precluded the *a priori* prediction of protein structures from sequences based on completely rigorous calculations. Consequently, simplified models are required at an appropriate level of coarse graining for complete descriptions of the energy landscape in protein conformational space. Statistical potentials to represent propensities of residues for backbone and side-chain dihedral angles,[1,2] the distributions of residues between the interior and exterior of protein molecules,[3] ion-pair substructures in proteins,[4] cis- and trans-conformations of proline residues,[5] the sizes of empty cavities,[6] and interactions among residues[7–15] have been devised as simplified empirical potentials, and extracted from known protein structures. They have been used to predict secondary structures, to measure compatibilities between protein sequence and structure,[16–25] to predict the docking of protein structures[26] and protein binding,[27,28] to predict stability changes caused by amino acid mutations,[29–31] and to simulate protein folding.[32–36] In particular, many potentials have been devised for pairwise residue interactions.[7–14] It has been shown that they are useful to distinguish native protein structures from non-native folds, and to also recognize folds with compatible sequences,[16–25] although there are certainly limitations.[37–39]

Long-range interactions among residues are principal forces to recognize the native pairs of protein sequence and structure among other pairs because they are responsible for proteins folding cooperatively into their unique native structures. However, short-range interactions should not be neglected even for fold-sequence recognition because they contribute significantly to the formation of secondary structures in proteins, which are also essential parts of protein structures. The classification into short-range and long-range terms here is based on the distance of separation between residues along a protein sequence and not on the physical range of interactions; short-range interactions are those between residues close along the protein sequence, and long-range interactions are those between sequentially distant residues.

Statistical potentials for secondary structures have also been evaluated from the frequencies of secondary structures in known protein structures, including only short-range interactions.[20,24,32] Secondary structure potentials were extracted by regarding the observed frequencies of secondary structure segments as an equilibrium distribution with a Boltzmann factor of secondary structure energies within those segments. This is a simple procedure. However, it is not a trivial task to properly estimate the total secondary structure energy for a whole protein structure from these secondary structure energies. If the total secondary structure energy for a protein structure is

Published 1999 WILEY-LISS, INC.

*Correspondence to: Sanzo Miyazawa, Faculty of Technology, Gunma University, Kiryu, Gunma 376-8515, Japan. E-mail: miyazawa@smlab.sci.gunma-u.ac.jp

Received 24 September 1998; Accepted 8 April 1999

calculated as the sum of these secondary structure energies for sequence segments which overlap with neighboring segments in a sequence, this energy will include interactions among residues more than once because of the overlapping segments. To correctly include each interaction among residues along a sequence, the interactions among residues within a segment must be divided both into one-body, two-body, and higher order interactions and also into backbone-backbone, backbone-side chain, and side chain-side chain interactions.

Here, the energies of secondary structures are estimated on the basic assumption that the observed frequencies of secondary structures in proteins can be regarded as an equilibrium distribution with respect to the secondary structure energies. Then, only the effects of short-range interactions on secondary structures are estimated as potentials of mean force from the observed frequencies of secondary structures in known protein structures, ignoring the effects of long-range interactions. However, this does not mean that either structure is arrived at independently of the other. The potentials of mean force for secondary structures are decoupled into several components: intrinsic, backbone-backbone and backbone-side chain potentials. Also, they are decoupled into one-body, two-body, and higher orders of interaction, in order to avoid multiple counts of each term in the estimation of secondary structure energies. These short-range potentials for secondary structure can be used additively with the long-range contact energies and repulsive packing energies[8] for evaluating the total conformational energies of proteins. It will be shown in the following paper[21] that their inclusion can substantially improve our capability for the recognition of native structures.

MATERIALS AND METHODS

Secondary Structure Potential

The total secondary energy of a protein is represented here as a sum over contributions from residues along the chain as

$$E^{sec} \equiv \sum_{p} E_p^{sec} \tag{1}$$

where p indexes residue position. The contribution of the pth residue to secondary structures is approximated to originate only in the short-range interactions.

$$E_p^{sec} \simeq e^s(\ldots ; i_{p-1}, s_{p-1}; i_p, s_p; i_{p+1}, s_{p+1}; \ldots) \tag{2}$$

$e^s(\ldots ; i_{p-1}, s_{p-1}; i_p, s_p; i_{p+1}, s_{p+1}; \ldots)$ is the short-range interaction energy within the secondary structure, $(\ldots ; i_{p-1}, s_{p-1}; i_p, s_p; i_{p+1}, s_{p+1}; \ldots)$, where i_p is the residue type at p, and s_p means the secondary structure of that residue. Thus, s_p designates a backbone conformation and i_p the residue type at position p. The ellipses indicate others yet unspecified, but nonetheless to be of limited range.

Here, the effects of short-range interactions on secondary structures are estimated by a potential of mean force based on the observed frequencies of secondary structures in known protein structures. The correlations between long- and short-range interactions are neglected, and the effects of long-range interactions are taken into account only as a mean field. As Némethy and Scheraga[40] pointed out from comparisons between the conformational maps of the 20 naturally occurring amino acid residues and the distribution of (ϕ, ψ) values in observed structures, mean fields can cause shifts in the (ϕ, ψ) values in proteins, within the low energy regions, away from the minima for the isolated residues, and cause a spread of (ϕ, ψ) values.

Because of the limited number of available protein structures, the secondary structure potential, e^s, is approximated as a sum of additive contributions from neighboring residues along a sequence, with neglect of side chain-side chain interactions. Non-additive contributions are simply neglected. In addition, the effects here from neighboring residues are limited to a dependence on their amino acid type but not on their secondary structures. These can include for example capping interactions exterior to secondary structure elements[41] or the sort of charge-helix dipole interactions treated long ago by Ptitsyn and Finkelshtein.[42] The conformational specification here is limited to a tripeptide. Thus, the secondary structure potential, e^s, is approximated as a sum of the following contributions.

$$e^s(\ldots ; i_{-1}, s_{-1}; i_0, s_0; i_1, s_1; \ldots) \simeq e^s(s_{-1}, s_0, s_1) + \sum_{-3 \le p \le 3} \delta e^s(s_{p-1}, s_p, s_{p+1}, i_0) \tag{3}$$

or

$$\simeq e^s(s_{-1}, s_0, s_1) + \sum_{-3 \le p \le 3} \delta e^s(s_{-1}, s_0, s_1, i_p). \tag{4}$$

The residue under consideration is indexed as zero, and negative and positive numbers represent relative residue positions towards the N-terminal and the C-terminal sides. It is assumed that the short-range interaction potentials for secondary structures do not depend on the absolute positions of residues in a sequence, but the relative positions between them. The first terms in Eq. 3 and 4 represent the backbone-backbone interactions and the second terms correspond to side chain-backbone interactions either within a residue or among residues. Altogether side chain-backbone interactions within five consecutive backbone units on each side of a side chain are included in the short-range interactions. The present evaluation of secondary structure energies differs substantially from those reported elsewhere, e.g., Bahar and Jernigan.[43]

Each term in Eqs. 3–4 is represented in the following form consisting of one-body and higher order interactions within backbones and between backbones and side chains, in order to avoid multiple counts of each interaction in the estimation of the total secondary structure energy of Eq. 1.

$$e^s(s_{-1}, s_0, s_1) \equiv e^s(s_0) + ½[\Delta e^s(s_{-1}, s_0) + \Delta e^s(s_0, s_1)] + \Delta e^s(s_{-1}, s_0, s_1) \tag{5}$$

$$\delta e^s(s_{-1}, i_0, s_0, s_1) \equiv \Delta e^s(i_0, s_0) + \{\Delta e^s(s_{-1}, i_0) + \Delta e^s(s_{-1}, i_0, s_0)\} + \{\Delta e^s(i_0, s_1) + \Delta e^s(i_0, s_0, s_1)\} + \Delta e^s(s_{-1}, i_0, s_0, s_1) \quad (6)$$

and for $p > 0$,

$$\delta e^s(s_{-1}, s_0, s_1, i_p) \equiv \delta e^s(s_{-p-1}, s_{-p}, s_{-p+1}, i_0) \equiv \Delta e^s(s_{-1}, i_p) + \Delta e^s(s_{-1}, s_0, i_p) + \Delta e^s(s_{-1}, s_0, s_1, i_p) \quad (7)$$

$$\delta e^s(i_{-p}, s_{-1}, s_0, s_1) \equiv \delta e^s(i_0, s_{p-1}, s_p, s_{p+1}) \equiv \Delta e^s(i_{-p}, s_1) + \Delta e^s(i_{-p}, s_0, s_1) + \Delta e^s(i_{-p}, s_{-1}, s_0, s_1). \quad (8)$$

The first term $e^s(s_0)$ in Eq. 5 represents the intrinsic propensity for secondary conformations in proteins, and $\Delta e^s(s_{-1}, s_0)$ in the second term corresponds to the nearest neighbor interactions between two consecutive backbone conformations, s_{-1} and s_0. The third term $\Delta e^s(s_{-1}, s_0, s_1)$ includes three-body interactions among three consecutive backbone conformations, s_{-1}, s_0 and s_1, in addition to the two-body backbone-backbone interactions between s_{-1} and s_1. These three terms include backbone-backbone interaction terms and do not depend on the type of side chain.

The potentials represented by Eqs. 7 and 8 are incremental energies and are not the total interaction energies between the side chain at position p or $-p$ and the tripeptide backbone at the center.

Each term in Eqs. 5–8 is estimated as follows in RT units as the potential of mean force from the observed frequencies of secondary structures in known protein structures. In the following formulation, the nearest neighbor correlation in the amino acid order of protein sequences is taken into account in the estimation of each term of the potential function.

$$e^s(s_p) = -\log\left(\frac{N(s_p)}{N}\right) + \text{Constant} \quad (9)$$

$$\Delta e^s(s_p, s_{p+1}) = -\log\left(\frac{N(s_p, s_{p+1})}{N}\right) + \log\left(\sum_{i_p, i_{p+1}} \frac{N(i_p, s_p)N(i_p, i_{p+1})N(i_{p+1}, s_{p+1})}{NN(i_p)N(i_{p+1})}\right) + \text{Constant} \quad (10)$$

$$\Delta e^s(s_{p-1}, s_p, s_{p+1}) = -\log\left(\frac{N(s_{p-1}, s_p, s_{p+1})}{N}\right) - \Delta e^s(s_{p-1}, s_p) - \Delta e^s(s_p, s_{p+1}) + \log\left(\sum_{i_{p-1}, i_p, i_{p+1}} \frac{N(i_{p-1}, s_{p-1})N(i_{p-1}, i_p)N(i_p, s_p)N(i_p, i_{p+1})N(i_{p+1}, s_{p+1})}{NN(i_{p-1})N(i_p)N(i_p)N(i_{p+1})}\right) + \text{Constant} \quad (11)$$

$$\Delta e^s(i_p, s_p) = -\log\left(\frac{N(i_p, s_p)}{N(i_p)}\right) + \log\left(\frac{N(s_p)}{N}\right) + \text{Constant} \quad (12)$$

$$\Delta e^s(i_p, s_p, s_{p+1}) = -\log\left(\frac{N(i_p, s_p, s_{p+1})}{N(i_p)}\right) + \log\left(\sum_{i_{p+1}} \frac{N(i_p, i_{p+1})N(s_p, i_{p+1}, s_{p+1})}{N(i_p)N(i_{p+1})}\right) - \Delta e^s(i_p, s_p) - \Delta e^s(i_p, s_{p+1}) + \text{Constant} \quad (13)$$

$$\Delta e^s(s_{p-1}, i_p, s_p) = -\log\left(\frac{N(s_{p-1}, i_p, s_p)}{N(i_p)}\right) + \log\left(\sum_{i_{p-1}} \frac{N(i_{p-1}, i_p)N(i_{p-1}, s_{p-1}, s_p)}{N(i_{p-1})N(i_p)}\right) - \Delta e^s(s_{p-1}, i_p) - \Delta e^s(i_p, s_p) + \text{Constant} \quad (14)$$

$$\Delta e^s(s_{p-1}, i_p, s_p, s_{p+1}) = -\log\left(\frac{N(s_{p-1}, i_p, s_p, s_{p+1})}{N(i_p)}\right) + \log\left(\sum_{i_{p-1}} \frac{N(i_{p-1}, i_p)N(i_{p-1}, s_{p-1}, s_p, s_{p+1})}{N(i_p)N(i_{p-1})}\right) + \log\left(\sum_{i_{p+1}} \frac{N(i_p, i_{p+1})N(s_{p-1}, s_p, i_{p+1}, s_{p+1})}{N(i_p)N(i_{p+1})}\right) - \log\left(\frac{N(s_{p-1}, s_p, s_{p+1})}{N}\right) - \Delta e^s(s_{p-1}, i_p, s_p) - \Delta e^s(i_p, s_p, s_{p+1}) - \Delta e^s(s_{p-1}, i_p) - \Delta e^s(i_p, s_p) - \Delta e^s(i_p, s_{p+1}) + \text{Constant}. \quad (15)$$

For $p \neq q$,

$$\Delta e^s(i_p, s_q) = -\log\left(\frac{N(i_p, s_q)}{N(i_p)}\right) + \log\left(\sum_{i_q} \frac{N(i_p, i_q)N(i_q, s_q)}{N(i_p)N(i_q)}\right) + \text{Constant} \quad (16)$$

$$
\begin{aligned}
\Delta e^{s}(i_p, s_{q-1}, s_q) = -\log\left(\frac{N(i_p, s_{q-1}, s_q)}{N(i_p)}\right) \\
&+ \log\left(\sum_{i_q} \frac{N(i_p, i_q)N(s_{q-1}, i_q, s_q)}{N(i_p)N(i_q)}\right) \\
&+ \log\left(\sum_{i_{q-1}} \frac{N(i_p, i_{q-1})N(i_{q-1}, s_{q-1}, s_q)}{N(i_p)N(i_{q-1})}\right) \\
&- \log\left(\frac{N(s_{q-1}, s_q)}{N}\right) \\
&- \Delta e^{s}(i_p, s_{q-1}) - \Delta e^{s}(i_p, s_q) \qquad (17) \\
&+ \text{Constant}
\end{aligned}
$$

$$
\begin{aligned}
&\Delta e^{s}(i_p, s_{q-1}, s_q, s_{q+1}) \\
&\quad = -\log\left(\frac{N(i_p, s_{q-1}, s_q, s_{q+1})}{N(i_p)}\right) \\
&\quad + \log\left(\sum_{i_{q-1}} \frac{N(i_p, i_{q-1})N(i_{q-1}, s_{q-1}, s_q, s_{q+1})}{N(i_p)N(i_{q-1})}\right) \\
&\quad + \log\left(\sum_{i_q} \frac{N(i_p, i_q)N(s_{q-1}, i_q, s_q, s_{q+1})}{N(i_p)N(i_q)}\right) \\
&\quad + \log\left(\sum_{i_{q+1}} \frac{N(i_p, i_{q+1})N(s_{q-1}, s_q, i_{q+1}, s_{q+1})}{N(i_p)N(i_{q+1})}\right) \\
&\quad -2\log\left(\frac{N(s_{q-1}, s_q, s_{q+1})}{N}\right) - \Delta e^{s}(i_p, s_{q-1}, s_q) \\
&\quad - \Delta e^{s}(i_p, s_q, s_{q+1}) - \Delta e^{s}(i_p, s_{q-1}) \\
&\quad - \Delta e^{s}(i_p, s_q) - \Delta e^{s}(i_p, s_{q+1}) + \text{Constant}. \qquad (18)
\end{aligned}
$$

N is the total number of residues in the set of all protein structures. $N(s_p)$ is the number of residues taking the conformational state s_p, and $N(i_p, s_p)$ is the number of residues of type i_p in conformational state s_p. $N(i_p, i_q)$ is the number of residue pairs with type i_p at position p and type i_q at position q. The indices, p and q, are taken to be relative to the 0th residue, that is, the residue under consideration. In the formulation above, the nearest neighbor correlation in the amino acid order of protein sequences is taken into account in the estimation of each term of the potential function. If there were no correlation in the amino acid order, then the corresponding terms in these equations would be reduced to simple terms consisting of backbone-backbone interaction energies. To avoid divergence of the logarithmic functions in Eqs. 9–18, a small number, which corresponds to the sampling error of value 0.5/N per triplet state (s_{p-1}, s_p, s_{p+1}) of secondary structure, is added to the arguments in the logarithmic functions.

All "constants" in Eq. 9 to Eq. 18 are terms that do not depend on the conformation and can take any value for conventional use as energy functions. However, the statistical averages of these energies should be set to zero for use as scoring functions for compatibilities between sequences and structures; refer to Miyazawa and Jernigan.[21]

TABLE I. Definition of Secondary Conformation States[†]

Secondary conformation states	Definition[a]
α	$\phi \leq 0$ and $-120 < \psi \leq 60$
β	$\phi < -90$ and not α
β_p (proline β)	$-90 \leq \phi \leq 0$ and not α
α_L (left handed α)	$\phi > 0$ and $120 > \psi \geq -60$
β_L (left handed β)	$\phi > 0$ and not α_L

[†] This classification is based on database analysis of Wilmot and Thornton.[45]

[a] Both ϕ and ψ are defined between −180 and 180 degrees.

$$\sum_{s_p} \frac{N(s_p)}{N} e^{s}(s_p) = 0 \qquad (19)$$

$$\sum_{s_p}\sum_{s_{p+1}} \frac{N(s_p, s_{p+1})}{N} \Delta e^{s}(s_p, s_{p+1}) = 0 \qquad (20)$$

$$\sum_{s_{p-1}}\sum_{s_p}\sum_{s_{p+1}} \frac{N(s_{p-1}, s_p, s_{p+1})}{N} \Delta e^{s}(s_{p-1}, s_p, s_{p+1}) = 0 \qquad (21)$$

$$\sum_{s_q} \frac{N(i_p, s_q)}{N(i_p)} \Delta e^{s}(i_p, s_q) = 0 \qquad (22)$$

$$\sum_{s_q}\sum_{s_{q+1}} \frac{N(i_p, s_q, s_{q+1})}{N(i_p)} \Delta e^{s}(i_p, s_q, s_{q+1}) = 0 \qquad (23)$$

$$\sum_{s_{q-1}}\sum_{s_q}\sum_{s_{q+1}} \frac{N(i_p, s_{q-1}, s_q, s_{q+1})}{N(i_p)} \cdot \Delta e^{s}(i_p, s_{q-1}, s_q, s_{q+1}) = 0 \qquad (24)$$

RESULTS

Sample Weighting

According to the procedure described in Miyazawa and Jernigan,[8] 1,168 protein structures in the Protein Data Bank (PDB),[44] whose structures were analyzed by X-ray and whose resolutions are better than 2.5 Å, are chosen, and then each of the 1,661 sequences in those structures is sampled with a weight based on the sequence identity matrix. These weights correctly reduce the contributions of repeated or nearly identical structures. As listed in Table 1 of Miyazawa and Jernigan,[8] the effective number of proteins that is defined as the sum of all sampling weights is 251. The effective number of residues is 54,356. These sampling weights and the corresponding data set of proteins were used to estimate contact energies and repulsive packing energies for all types of amino acids by Miyazawa and Jernigan.[8] Here, these same structures have been used to estimate short-range potential energies for secondary structures, for consistency.

Secondary Structure Potential

Because there are correlations of (φ, ψ) among neighboring residues and also because coarse-graining is necessary for the recognition of sequence—structure compatibilities,

TABLE II. Empirical Potentials† for Secondary Conformations in RT Units Where the Specified Side Chain is at the Middle of the Segment

Side chain	#aa	α	β	β_p	α_L	β_L	$\alpha\alpha\alpha$	$\beta\beta\beta$	$\beta_p\beta\beta_p$	$\beta\beta\beta_p$	$\beta_p\beta\beta$	$\beta_p\beta_p\beta_p$
Backbone[a]	51129.1	−0.52	0.01	0.69	1.72	2.55	−0.82	−0.57	−0.18	−0.11	−0.04	0.44
LYS	3028.8	−0.15	0.17	0.16	0.31	1.34	−0.09	0.58	−0.39	0.04	0.08	0.06
ARG	2155.0	−0.11	0.03	0.21	0.75	1.47	−0.12	0.39	−0.59	−0.01	−0.13	0.28
GLN	1815.9	−0.12	0.07	0.21	0.48	1.52	−0.13	0.43	−0.30	−0.05	−0.21	0.01
GLU	2928.4	−0.22	0.41	0.22	0.93	1.44	−0.24	0.57	0.13	0.28	0.91	0.25
ALA	4407.4	−0.19	0.40	0.06	1.19	1.21	−0.26	0.31	0.50	0.46	0.52	−0.04
MET	955.0	−0.09	−0.03	0.35	1.15	3.67	−0.15	−0.13	−0.42	−0.16	−0.15	0.25
LEU	4172.5	−0.08	0.03	0.04	1.39	2.10	−0.15	−0.16	0.01	−0.08	−0.02	0.05
PHE	1989.6	0.09	−0.26	0.28	1.29	2.08	0.06	−0.37	0.25	−0.40	−0.10	0.93
ILE	2714.9	0.18	−0.33	0.50	2.58	3.11	0.10	−0.49	−0.31	−0.48	−0.53	0.16
VAL	3564.7	0.33	−0.40	0.45	2.76	2.47	0.28	−0.57	−0.56	−0.44	−0.63	0.55
TRP	776.4	0.02	−0.13	0.09	1.28	2.36	−0.01	−0.28	0.36	−0.08	−0.15	2.39
TYR	1919.6	0.21	−0.33	0.19	0.85	1.83	0.21	−0.45	0.52	−0.38	−0.17	0.78
CYS	943.0	0.27	−0.34	0.10	0.70	1.14	0.33	−0.31	−0.42	−0.11	−0.42	0.05
THR	3043.3	0.15	−0.27	0.15	2.07	1.27	0.35	−0.16	−0.01	−0.10	−0.29	0.77
SER	3426.2	0.03	−0.09	−0.07	0.69	0.61	0.27	0.12	0.33	0.40	0.34	0.77
ASP	3139.7	−0.08	0.27	−0.14	−0.01	0.97	0.25	1.19	1.44	1.35	1.47	0.60
ASN	2307.1	0.06	0.12	0.19	−0.78	0.94	0.33	0.81	0.80	0.71	0.75	0.86
HIS	1139.2	0.01	−0.09	0.18	−0.07	1.49	0.04	0.28	0.22	0.13	0.06	0.32
GLY	4408.5	1.49	1.45	1.36	−1.10	−1.45	1.51	1.44	2.07	1.75	1.81	1.92
PRO	2293.8	0.75	3.41	−0.71	4.58	4.22	2.32	4.44	3.40	3.32	3.16	−0.82

† $\Delta e^s(i_0, s_0)$ and $\delta e^s(s_{-1}, i_0, s_0, s_1)$.

[a] The values of $e^s(s_0)$ and $e^s(s_{-1}, s_0, s_1)$.

coarse-graining is employed directly for the backbone conformation of a tripeptide. The (ϕ, ψ) conformation of a residue is classified into five conformational states.[45] These are α, β, proline $\beta(\beta_p)$, left handed $\alpha(\alpha_L)$, and left handed $\beta(\beta_L)$, as defined in Table I. There is no differentiation between the trans and rare cis peptide conformations. Consequently, the total number of backbone conformational states for a tripeptide is 125. For these 125 backbone conformational states, the intrinsic potential energies of backbone conformations and the interaction energies between backbones and side chains are evaluated from the observed frequencies of those conformations of tripeptides according to Eqs. 9–18. The short interaction range is assumed to include only those within five residues, which corresponds to about one turn of an α helix; see Eq. 3. As noted in the Methods section, side chain-side chain interactions are neglected not because they do not have significant contributions to the stability of secondary structures but because the number of available protein structures is limited and thus they cannot be evaluated reliably.

Table II shows side chain-backbone interaction energies $\Delta e^s(i_0, s_0)$ between the backbone of secondary conformation s_0 and its side chain of type i_0, and $\delta e^s(s_{-1}, i_0, s_0, s_1)$ between the backbone (s_{-1}, s_0, s_1) of a tripeptide and a side chain of type i_0 located at the middle position of tripeptides in regular α and β secondary conformations. Intrinsic energies $e^s(s_0)$ of single peptides s_0 and backbone-backbone interaction energies $e^s(s_{-1}, s_0, s_1)$ including those intrinsic energies of single peptides s_0 at the middle of tripeptides and interaction energies between the single peptide s_0 and both neighbors (s_{-1}, s_1) are also shown in the top row of this table. These interaction energies are defined by Eqs. 3–8 and calculated from Eqs. 9–24. Many features in this table could have been anticipated. Glu and Ala are side chains that individually favor the α helical conformation but disfavor the β strand conformation. Side chains of Lys, Arg, Gln, Met, and Leu favor both α ($\alpha\alpha\alpha$) and β ($\beta_p\beta\beta_p$ or $\beta\beta\beta$) structures. Phe, Ile, Val, Trp, Tyr, Cys, and Thr are β ($\beta\beta\beta$, $\beta\beta\beta_p$, $\beta_p\beta\beta$, or $\beta_p\beta\beta_p$) strand formers but α ($\alpha\alpha\alpha$) helix breakers except for Trp. Ser, Asp, Asn, and His do not have any preference for either regular α or β structures. Pro is the strongest breaker of these regular structures. Also, Gly is not preferable in the middle of any of these regular structures, probably because of its flexible backbone conformation. Asp and Asn are β strand breakers nearly to the same extent as Gly. Position effects of side chains on the interaction energies between side chains and tripeptide backbones are also observed.

In Figure 1, the short-range energies per residue for a representative set of proteins are plotted against the stated X-ray resolutions of the protein structures. It should be noted that the ordinate is defined to include side chain-backbone interaction energies only, that is, the backbone-backbone interaction energies corresponding to the first term in Eqs. 3 and 4 are not included in the ordinate. Open circles show proteins whose structures were determined by NMR analyses. The protein structures used here are 189 proteins that differ from each other by at least 35% in sequence identity and were those selected by Orengo, et al[46]; see their Table I. Proteins with many unknown atomic coordinates are not included. There appears to be a clear difference in the average short-range energies between proteins with resolutions worse than 2.5

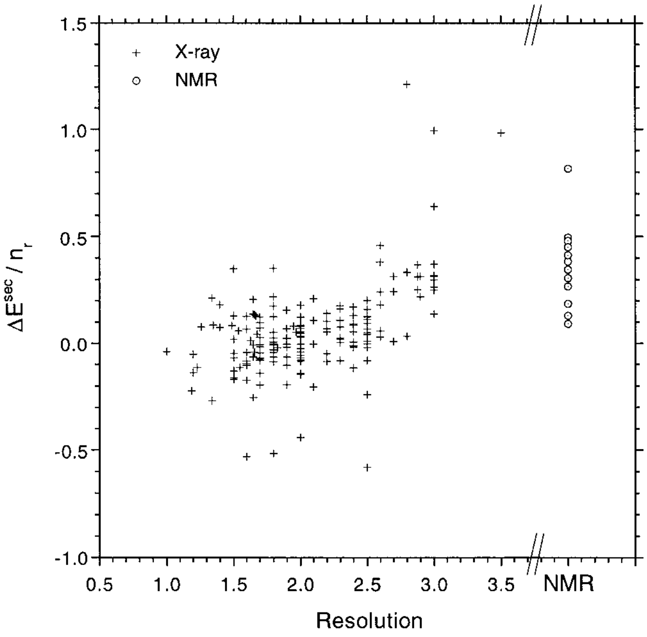

Fig. 1. The relationship between the average secondary structure energy per residue for representative proteins and their resolution from X-ray analyses; here the ordinate is defined not to include the backbone-backbone interaction energies corresponding to the first term in Eqs. 3 and 4, that is, $\Delta E^{sec} \equiv \Sigma_p \Sigma_{p-3 \le q \le p+3} \delta e^s(s_{q-1}, s_q, s_{q+1}, i_p)$. All energies here are given in RT units. The representative protein structures used here are 189 protein structures that differ from each other by at least 35% sequence identity and are those selected by Orengo et al.[46]; see their Table I. Proteins with many unknown atomic coordinates are not included. Open circles show proteins whose structures are determined by NMR analyses. Here it should be noted that the present estimates of the short-range energies have been compiled from a dataset comprised of protein structures having resolutions better than 2.5 Å. The entry names and sequence identifiers of the PDB files used for this figure are:

Membrane proteins:
1PRC-L 1PRC-M 1PRC-C 2POR 1SN3 1VSG-A 1HGE-A 1HGE-B 1PRC-H

Metal binding proteins:
1CY3 1PRC-C 5RXN 2HIP-A 2CDV

DNA binding proteins:
1HDD-C

Inhibitors without an enzyme:
1HOE 1PI2 3EBX 2OVO 5PTI

Multimeric proteins without subunit interactions:
2WRP-R 1UTG 1ROP-A 2TMV-P 2RHE 2STV 3PGM 6LDH 1PYP

Structures determined by NMR:
1C5A 1HCC 1ATX 1SH1 2SH1 1EPG 4TGF 3TRX 1EGO 1APS
1IL8-A 2GB1

Other monomeric proteins:
1MBC 1MBA 1ECD 2LH3 2LHB 1R69 4ICB 4CPV 1LE2 1YCC
1CC5 451C 1IFC 1RBP 1SGT 4PTP 2SGA 2ALP 2SNV 1CD8
1CD4 1ACX 1PAZ 1PCY 1GCR 2CNA 3PSG 1F3G 8I1B 1ALD
1PII 6XIA 2TAA-A 4ENL 5P21 4FXN 2FCR 2FX2 3CHY 5CPA
8DFR 3DFR 3ADK 1GKY 1RHD 4PFK 3PGK 2GBP 8ABP 2LIV
1TRB 1IPD 4ICD 1PGD 8ADH 2TS1 1PHH 3LZM 1LZ1 1RNH
7RSA 1CRN 1CTF 1FXD 2FXB 4FD1 1FDX 4CLA 9RNT 1RNB-A
1FKF 1SNC 1UBQ 3B5C 9PAP 3BLM 2CPP 1CSC 1ACE 1COX
1GLY 1LAP 1LFI 2CYP 8ACN 2CA2

Other multimeric proteins:
1HBB-A 2SDH-A 1ITH-A 1COL-A 1LMB-A 3SDP-A 2SCP-A 2HMZ-A 256B-A 2CCY-A
1GMF-A 1BBP-A 2FB4-H 3HLA-B 1COB-A 2AZA-A 2PAB-A 1BMV-1 1BMV-2 2PLV-1
1TNF-A 2MEV-1 2MEV-2 2MEV-3 2PLV-2 2PLV-3 2LTN-A 2RSP-A 2ER7-E 5HVP-A
1NSB-A 5TIM-A 2TRX-A 1CSE-E 1GP1-A 4DFR-A 8CAT-A 4MDH-A 1GD1-O 7AAT-A
1HRH-A 1RVE-A 2SIC-I 8ATC-B 2TSC-A 2SAR-A 1MSB-A 1BOV-A 1FXI-A 1TGS-I
1TPK-A 9WGA-A 3HLA-A 8ATC-A 2CPK-E 1GST-A 1OVA-A 7API-A 1WSY-B 2GLS-A
2PMG-A 6TMN-E 3GAP-A

Å and those with better resolution. Probably the relatively large values of the average secondary structure energies of side chain-backbone interactions reflect the poor resolutions of these protein structures.

Position Effects of Side Chains on the Secondary Structure Potentials

Position effects of side chains on the interaction energies between side chains and tripeptide backbones are shown in Figure 2 for the $\alpha\alpha\alpha$ conformation and in Figure 3 for the $\beta\beta\beta$ conformation. Note that backbone conformations are fixed only for a tripeptide at the center, and otherwise the conformation is unspecified.

As expected, a strongly asymmetric positional effect of Pro on a regular α helix is detected; Pro breaks the helical structure at its C-terminal side. The positively charged side chains, His, Lys, and Arg, prefer the C-terminal over the N-terminal side in α helical conformations, probably because of the electrostatic interactions between peptide dipoles and side chain charges. Lys especially shows such a position effect. Also, a similar but inverse position effect is observed for the negatively charged side chains, Glu and Asp. Asp shows an especially strong position effect in that it disfavors the α helical conformation more strongly at other positions than near the N-terminal end. An unexpected feature is the similar position effect detected for Thr and Ser.

For the $\beta\beta\beta$ conformation shown in Figure 3, large differences in preference between the N- and C-terminal sides of a tripeptide are not detected for any side chains except for the positively charged side chains, Lys, Arg and His, for Gln, and for Trp and Cys. Lys, Arg, His and Gln dislike the C-terminal side of β strands as well as the middle of β strands. The convex curves for Glu, Gln and Ala, as well as Pro, Gly, Asp and Asn reflect their disfavor for β structure; Table II shows that Gln favors the $\beta_p\beta\beta_p$ conformation. Oppositely, the concave curves for Phe, Ile, and Val reflect their preference for β structure. The potential curves for Met and Leu are concave but positive at the ±3th positions, indicating that those residues tend to be located in short β strands rather than in long β strands.

Effects of Mis-Alignments on Secondary Structure Energies

Let us consider the average increases in incremental backbone-side chain interaction energies due to the mis-alignment of a residue of type i_0 to structure type j_0 in a sequence—structure alignment.

$$<\delta e^s(s_{p-1}, s_p, s_{p+1}, i_0)>_{j_0} \equiv \sum_{s_{p-1}} \sum_{s_p} \sum_{s_{p+1}} \frac{N(s_{p-1}, s_p, s_{p+1}, j_0)}{N(j_0)} \delta e^s(s_{p-1}, s_p, s_{p+1}, i_0) \qquad (25)$$

Table III shows the average total increments of the secondary structure energies accompanied by the mis-alignment of an amino acid of type i_0 to structure type j_0, that is,

$$\sum_{-3 \le p \le 3} <\delta e^s(s_{p-1}, s_p, s_{p+1}, i_0)>_{j_0} . \qquad (26)$$

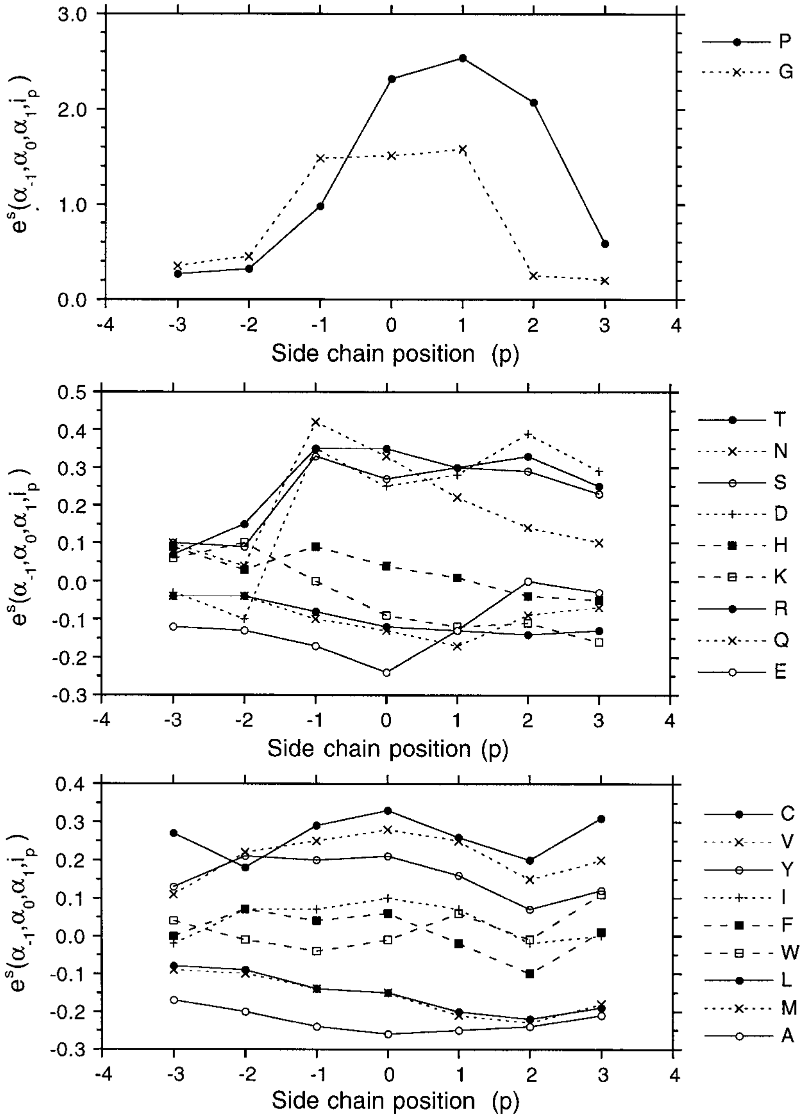

Fig. 2. Structure-derived position effects of a side chain on the interaction energy between a side chain and a helical tripeptide backbone in the ααα conformation. The interaction energies between a tripeptide backbone of the ααα conformation and side chains located on the N-terminal side (negative p) or C-terminal side (positive p) of the tripeptide are shown; for $p > 0$

$$e^s(s_{-1}, i_0, s_0, s_1) \equiv \delta e^s(s_{-1}, i_0, s_0, s_1)$$

$$e^s(s_{-1}, s_0, s_1, i_p) \equiv e^s(s_{-p-1}, s_{-p}, s_{p+1}, i_0)$$
$$\equiv \Delta e^s(s_1, i_p) + \Delta e^s(s_0, s_1, i_p) + \delta e^s(s_{-1}, s_0, s_1, i_p)$$

$$e^s(i_{-p}, s_{-1}, s_0, s_1) \equiv e^s(i_0, s_{p-1}, s_p, s_{p+1})$$
$$\equiv \Delta e^s(i_{-p}, s_{-1}) + \Delta e^s(i_{-p}, s_{-1}, s_0) + \delta e^s(i_{-p}, s_{-1}, s_0, s_1)$$

These interaction energies are calculated from Eqs. 12–24.

These side chain-backbone interaction energies are just the average energy increments accompanying an amino acid replacement from native type j_0 to type i_0. The amino acid types are sorted, so that the average total energies tend to be less positive for diagonal than for off-diagonal terms. Notably mis-aligning Pro to any other type of side chain costs a large positive energy, but the inverse mis-alignments to Pro are not accompanied by such large penalties. This is reasonable, because Pro is quite incompatible with α and β conformations that are commonly taken up by other types of residues, but conversely other types of residues can take the β_p conformation that is common for Pro. An asymmetry is observed for Gly alignments. Generally the mis-alignments of Gly to α helix formers tend to cost a larger penalty on average than inverse mis-alignments, but those to β strand formers have less penalty. The large positive increments in secondary structure energy occur with both the mis-alignments of Gly to other types of amino acids and the inverse mis-alignments, because of the flexibility of Gly. On average, mis-alignments of Met, Trp, Cys, Gly and Pro cost significantly larger energies than others; however, the energies estimated for Met, Trp and Cys are likely to be less reliable than others because of their smaller sample sizes. The average increment of secondary structure energy due to mis-alignments over all types of amino acids is 0.83 (RT units).

DISCUSSION

Many years ago, Némethy and Scheraga[40] pointed out that the observed conformations of each residue in proteins mostly fall in the low-energy regions of the (φ, ψ) map of individual residues. This fact supports the present basic assumption and indicates some consistency between the long and short-range interactions. Consistency among interactions in proteins was originally proposed as the "principle of structural consistency" by Go,[47] and also called the "principle of minimal frustration" in the energy landscape view of proteins advanced by Bryngelson and Wolynes.[48]

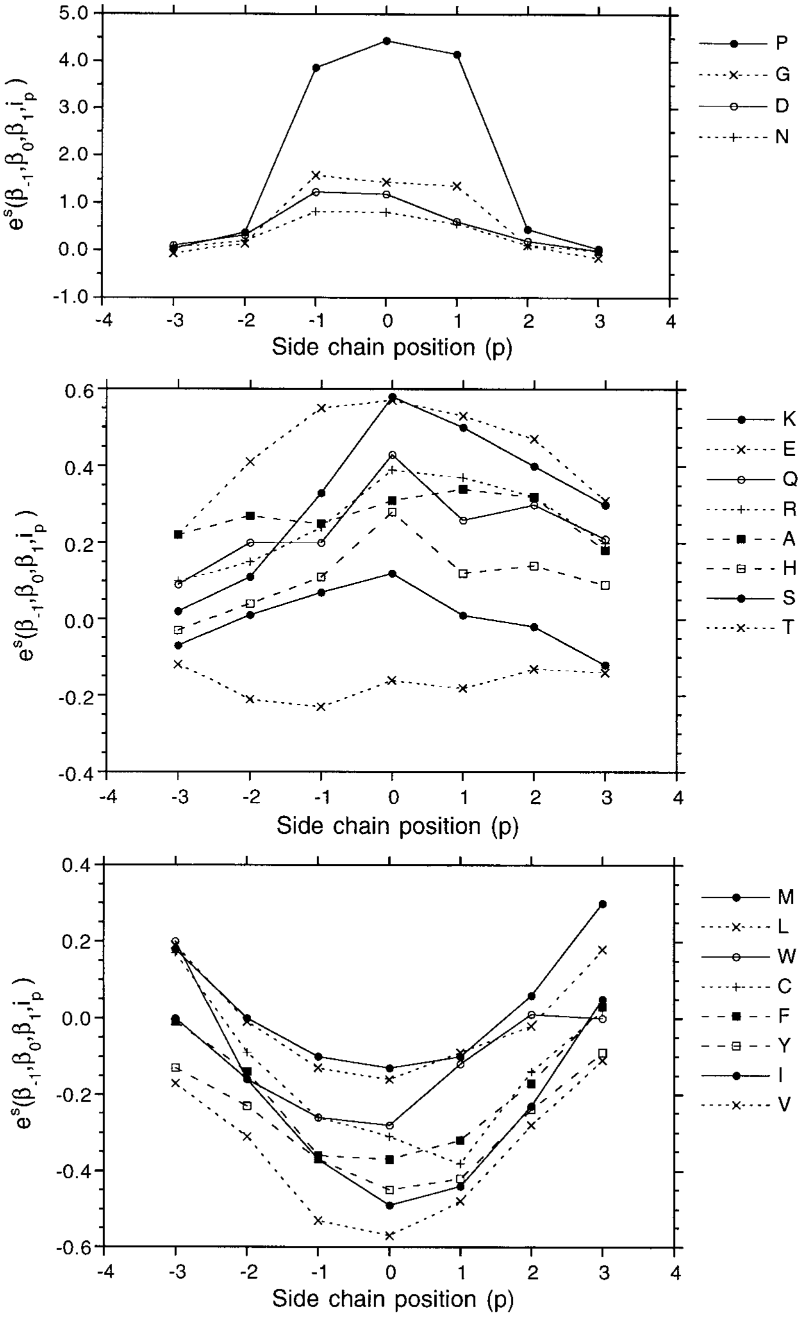

Fig. 3. Position effects of a side chain on the interactions between a side chain and a tripeptide backbone in the βββ conformation. The interaction energies between a tripeptide backbone of the βββ conformation and side chains located at the N-terminal side (negative p) or C-terminal side (positive p) of the tripeptide are shown; see the caption of Figure 2 for the definitions of these energies. These interaction energies are calculated from Eqs. 12–24.

In the case of secondary structures, as pointed out by Némethy and Scheraga[40] long-range interactions can shift the (ϕ, ψ) values in proteins, within a low energy region, away from the minima for the isolated residues, and long-range interactions have the effect of spreading out the (ϕ, ψ) values observed in proteins more than for isolated residues. Such an effect of long-range interactions may depend on the overall conformation of proteins but here it is treated as a mean field in proteins, ignoring the details of such a dependence. As a result, this long-range effect of spreading out of the (ϕ, ψ) values relative to the actual potentials may effectively increase temperatures in Boltzmann distributions for secondary structure potentials.

Potentials of local conformations were also evaluated from statistical preferences observed in protein structures.[20,24,49] In Nishikawa and Matsuo,[24] the total local conformational energy of a protein fold was evaluated as a sum of the energies of pentapeptide conformations at residue positions along a sequence. The potential of a pentapeptide conformation was approximated as a sum of tripeptide potentials; see Eq. 6 of their paper. In this formalism, the nearest neighbor interactions between side chains and peptide backbones, $\Delta e^s(i_0, s_1)$ and $\Delta e^s(s_{-1}, i_0)$ in the present terminology, appear to be improperly counted twice in the total local conformational energy. These types of nearest neighbor interactions are not negligible; espe-

TABLE III. Average Increments Defined by Eq. 27 of the Secondary Structure Energies for the Mis-Alignment of an Amino Acid of Type i (Row) to Structure Type j (Column)[†]

	LYS	ARG	GLN	GLU	ALA	MET	LEU	PHE	ILE	VAL	TRP	TYR	CYS	THR	SER	ASP	ASN	HIS	GLY	PRO	Average
LYS	0.00	0.25	0.28	0.24	0.31	0.35	0.36	0.45	0.39	0.42	0.47	0.47	0.56	0.40	0.43	0.44	0.46	0.43	1.04	0.37	0.42
ARG	0.32	0.00	0.33	0.35	0.36	0.30	0.36	0.41	0.37	0.39	0.43	0.49	0.53	0.45	0.51	0.55	0.57	0.48	1.47	0.65	0.50
GLN	0.34	0.33	0.00	0.31	0.38	0.33	0.37	0.47	0.38	0.40	0.53	0.54	0.58	0.44	0.51	0.51	0.54	0.46	1.36	0.72	0.50
GLU	0.33	0.37	0.34	0.00	0.34	0.41	0.45	0.51	0.49	0.50	0.48	0.52	0.62	0.44	0.45	0.42	0.56	0.46	1.44	0.46	0.50
ALA	0.27	0.26	0.28	0.21	0.00	0.19	0.19	0.27	0.24	0.27	0.29	0.38	0.42	0.38	0.41	0.47	0.58	0.40	1.31	0.32	0.38
MET	1.06	0.87	0.95	1.00	0.83	0.00	0.63	0.68	0.61	0.63	0.80	0.91	1.07	1.11	1.32	1.52	1.58	1.16	3.25	1.44	1.17
LEU	0.53	0.43	0.46	0.52	0.35	0.18	0.00	0.25	0.16	0.19	0.35	0.38	0.42	0.55	0.76	0.91	0.94	0.59	1.98	0.72	0.59
PHE	0.79	0.63	0.69	0.71	0.57	0.36	0.38	0.00	0.30	0.34	0.40	0.40	0.65	0.71	0.87	1.05	1.05	0.71	2.10	1.19	0.76
ILE	0.84	0.71	0.75	0.84	0.64	0.41	0.43	0.43	0.00	0.22	0.53	0.62	0.72	0.75	1.12	1.46	1.65	1.00	3.49	1.10	0.98
VAL	0.73	0.63	0.63	0.72	0.57	0.39	0.37	0.37	0.16	0.00	0.48	0.48	0.56	0.64	0.93	1.28	1.35	0.84	2.79	1.03	0.82
TRP	1.46	1.23	1.30	1.16	1.13	0.93	0.96	0.89	0.95	0.99	0.00	1.09	1.38	1.27	1.52	1.56	1.76	1.39	3.38	1.66	1.41
TYR	0.67	0.57	0.58	0.59	0.55	0.38	0.40	0.27	0.34	0.32	0.32	0.00	0.48	0.52	0.69	0.80	0.80	0.56	1.91	0.98	0.65
CYS	1.41	1.21	1.24	1.26	1.18	0.97	0.94	0.91	0.87	0.86	1.04	1.01	0.00	1.21	1.35	1.39	1.47	1.16	2.59	1.59	1.28
THR	0.45	0.41	0.41	0.38	0.43	0.42	0.41	0.39	0.33	0.32	0.35	0.34	0.42	0.00	0.31	0.37	0.59	0.47	1.77	0.47	0.49
SER	0.38	0.39	0.39	0.33	0.41	0.47	0.50	0.46	0.50	0.47	0.41	0.41	0.44	0.26	0.00	0.12	0.25	0.33	0.86	0.31	0.39
ASP	0.73	0.78	0.74	0.62	0.77	0.94	0.95	0.94	1.08	1.10	0.86	0.89	0.90	0.69	0.50	0.00	0.32	0.64	1.02	0.54	0.75
ASN	0.64	0.64	0.63	0.60	0.72	0.78	0.82	0.77	0.90	0.90	0.73	0.74	0.79	0.61	0.49	0.23	0.00	0.53	0.83	0.81	0.66
HIS	0.65	0.59	0.58	0.57	0.62	0.57	0.61	0.57	0.66	0.67	0.64	0.60	0.77	0.66	0.66	0.56	0.54	0.00	1.64	0.93	0.71
GLY	1.72	1.72	1.70	1.66	1.70	1.78	1.79	1.74	1.85	1.83	1.71	1.73	1.69	1.73	1.61	1.49	1.40	1.58	0.00	1.92	1.56
PRO	3.18	3.24	3.33	2.96	2.93	3.25	3.13	3.46	3.47	3.57	3.19	3.55	3.48	3.29	3.15	3.06	3.73	3.48	5.36	0.00	3.30
Ave	0.76	0.72	0.74	0.70	0.69	0.67	0.67	0.70	0.68	0.69	0.71	0.75	0.81	0.75	0.81	0.84	0.94	0.80	1.76	0.79	0.83

[†] The average energy increment per residue for a random alignment is 0.83. All energies are in RT units.

cially those between a Pro side chain and a peptide backbone are strong; see our Figures 2 and 3. As a result, the estimates[31] of unfolding free energy changes due to amino acid replacements using this kind of potential may be qualitatively correct, but could be quantitatively incorrect; there are also uncertainties associated with the relative contributions of classes of interactions, hydration energies, side chain packing energies, hydrogen bonding energies, and local conformational energies, which they have summed with arbitrary weights.

On the other hand, in Kocher et al.,[20] three-body interactions between side chains and peptide backbones seem to be simply averaged in the residue-to-torsion potentials and also in the torsion-to-residue potentials regardless of the length of separation between side chains and peptide backbones. The intrinsic interaction energies of backbone conformations are included in the torsion-to-residue potentials. In order to properly evaluate the interactions between side chains and backbones, decoupling interactions into one-body, two-body and higher order interactions as done here is preferable.

Instead of (ϕ, ψ), DeWitte & Shakhnovich[49] used the dihedral angle formed with consecutive four C^α atoms to specify local conformational states in order to evaluate local conformational energies. The contributions of side chain-side chain interactions between nearest neighbors to secondary structure formation were evaluated by classifying residues into three categories, helix, sheet, and turn formers. The coarse graining over residue types would probably be a good way to overcome the limited number of available protein structures and to take account of side chain-side chain interactions.

Here, the (ϕ, ψ) space of a peptide backbone has been divided into a small number of discrete states, and then the secondary structure potentials have been evaluated for these discrete states. This kind of coarse-graining is appropriate for describing protein folds for fold recognition. However, finer-graining of secondary structure potentials may be required for simulations of protein conformations. However, the secondary structure potentials for a tripeptide cannot be evaluated with such a fine mesh, because of the relatively small number of protein structures. Hybrid potentials, in which $e^s(s_0)$ and $\Delta e^s(i_p, s_0)$ are evaluated with a fine mesh and other terms such as $\Delta e^s(i_0, s_{p-1}, s_p, s_{p+1})$ are evaluated with a coarser mesh, might be useful. For such a hybrid potential, the decoupling of interactions into one-body, two-body, and higher order interactions is essential. In the present analysis, the trans and cis conformations of a peptide backbone are not distinguished, but they may also need to be distinguished.

Most qualitative results of the effects of each type of side chain on secondary structures found here, such as designation of breakers and formers for α helix and β strand and the strong position effects of Pro, are already known, but the present type of detailed evaluation of the potential of mean force has never been carried out before. The position effects of charged side chains on helix formation are also well known.[50] However, new observations here include the position effect of side chains on a β strand. These secondary structure potentials could also be used for secondary structure predictions. But, our intention here has been to apply them in the next article[21] to fold and sequence recognition.

CONCLUSIONS

The short-range potentials for secondary structures here have been evaluated from the observed frequencies of secondary structures for tripeptides in proteins. The basic assumption is that observed frequencies of secondary structures over proteins can be regarded as an equilibrium distribution with respect to the secondary structure energies. Long-range interactions are assumed to be consistent with short-range interactions for the stabilization of secondary conformations, and the effects of the long-range interactions are treated as a mean field. Because they have been developed carefully in a self-consistent way, these short-range potentials for secondary structure can be used additively with the long-range contact energies and repulsive packing energies for evaluating the total conformational energies of proteins. It will be shown in the following article[21] that their inclusion can substantially improve the capability for the recognition of native structures.

REFERENCES

1. Pohl FM. Empirical protein energy maps. Nature New Biol 1971;234:277–279.
2. Pohl FM. Statistical analysis of protein structures. In Jaenicke R, editor. Protein folding, Amsterdam: Elsevier/North-Holland Biomedical Press; 1980. p 183–196.
3. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
4. Bryant S, Lawrence CE. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions. Proteins 1991;9:108–119.
5. MacArthur MW, Thornton JM. Influence of proline residues on protein formation. J Mol Biol 1991;218:397–412.
6. Rashin AA, Ionif M, Honig B. Internal cavities and buried waters in globular proteins. Biochemistry 1986;25:3619–3625.
7. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
8. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 1996;256: 623–644.
9. Miyazawa S, Jernigan RL. Self-consistent estimation of interresidue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1998;34:49–68.
10. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
11. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
12. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
13. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
14. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
15. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Prot Sci 1997;6:1467–1481.
16. Hendlich M, Lackner P, Weitckus S, Floechner H, Froschauer R, Gottsbachner K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models; the calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
17. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 1992;13:258–271.
18. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
19. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: application to super-secondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
20. Kocher JPA, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure Matches. J Mol Biol 1994;235:1598–1613.
21. Miyazawa S, Jernigan RL. An Empirical energy potential with a reference state for Protein Fold and Sequence Recognition. Proteins 1999;36:357–369.
22. Miyazawa S, Jernigan RL. A scoring function with structure-dependent gap penalties for identifying protein sequence-structure compatibilities. 1999; submitted.
23. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
24. Nishikawa K, Matsuo Y. Development of pseudoenergy potentials for assessing protein 3D-1D compatibility and detecting weak homologies. Prot Eng 1993;6:811–820.
25. Matsuo Y, Nakamura H, Nishikawa K. Detection of protein 3D-1D compatibility characterized by the evaluation of side-chain packing and electrostatic interactions. J Biochem 1995;118:137–148.
26. Pellegrini M, Doniach S. Computer simulation of antibody binding specificity. Proteins 1993;15:436–444.
27. Altuvia Y, Schueler O, Margalit H. Ranking potential binding peptides to MHC molecules by a computational threading approach. J Mol Biol 1995;249:244–250.
28. Wallqvist A, Jernigan RL, Covell DG. A preference-based free-energy parametrization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. Prot Sci 1995;4:1881–1903.
29. Miyazawa S, Jernigan RL. A New substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Engineering 1993;6:267–278.
30. Miyazawa S, Jernigan RL. Protein stability for single substitution mutants and the extent of local compactness in the denatured state. Prot Eng 1994;7:1209–1220.
31. Ota M, Kanaya S, Nishikawa K. Desk-top analysis of the structural stability of various point mutations introduced into ribonuclease H. J Mol Biol 1995;248:733–738.
32. Miyazawa S, Jernigan RL. Equilibrium folding and unfolding pathways for a model protein. Biopolymers 1982;21:1333–1363.
33. Wilson C, Doniach S. A computer model to dynamically simulate protein folding: studies with crambin. Proteins 1989;6:193–209.
34. Skolnick J, Kolinski A. Simulations of folding of a globular protein. Science 1990;250:1121–1125.
35. Sun S. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. Protein Sci 1993;2:762–785.
36. Koliński A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994;18:338–352.
37. Thomas PD, Dill K. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
38. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
39. Park BH, Huang ES, Levitt M. Factors Affecting the Ability of Energy Functions to Discriminate Correct from Incorrect Folds. J Mol Biol 1997;266:831–846.
40. Némethy G, Scheraga HA. Protein folding. Quart Rev Biophys 1977;10:239–352.
41. Aurora R, Rose GD. Helix capping. Protein Science 1998;7:21–38.
42. Ptitsyn OB, Finkelshtein AV. Prediction of helical portions of globular proteins according to their primary structure. Dokl Akad Nauk SSSR (Engl. transl.) 1970;195:221–224.
43. Bahar I, Jernigan RL. Short-rage conformational energies, secondary structure propensities, and recognition of correct sequence—structure matches. Proteins 1997;29:292–308.
44. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1997;112:535–542.
45. Wilmot CM, Thornton JM. β-turns and their distributions: a proposed new nomenclature. Prot Eng 1990;3:479–493.
46. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Prot Eng 1993;6:485–500.
47. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
48. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84: 7524–7528.
49. DeWitte RS, Shakhnovich EI. Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. Protein Sci 1994;3:1570–1581.
50. Jernigan RL, Miyazawa S, Szu SC. Stabilization of regular conformational regions in proteins by intra-region electrostatic interactions. Macromolecules 1980;13:518–525.