The Miyazawa-Jernigan Contact Energies Revisited

Hui Zeng, Ke-Song Liu and Wei-Mou Zheng*

Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China

Abstract: The Miyazawa-Jernigan (MJ) contact potential for globular proteins is a widely used knowledge-based potential. It is well known that MJ's contact energies mainly come from one-body terms. Directly in the framework of the MJ energy for a protein, we derive the one-body term based on a probabilistic model, and compare the term with several hydrophobicity scales of amino acids. This derivation is based on a set of native structures, and the only information of structures manipulated in the analysis is the contact numbers of each residue. Contact numbers strongly correlate with layers of a protein when it is viewed as an ellipsoid. Using an entropic clustering approach, we obtain two coarse-grained states by maximizing the mutual information between coordination numbers and residue types, and find their differences in the two-body correction. A contact definition using sidechain centers roughly estimated from C_α atoms results in no significant changes.

Keywords: Protein contact energies, Protein potential function.

1. INTRODUCTION

Comprehending the nature of the physical interactions between different types of amino acids is crucial for understanding protein folding and structure prediction. The physics-based potential functions are based on full atomic models and therefore require high computational cost. Furthermore, they need not fully capture all of the important physical interactions. The known three-dimensional structures of proteins contain a large amount of information on the forces stabilizing proteins. Potential functions and the rules governing protein stability can be revealed from statistical analysis on known structures. By assuming that frequently observed structural features correspond to low-energy states, the observed statistical frequencies of various features, after comparing with that in a reference state or a null model, are converted into effective free energies [1-5]. A recent review on the theory and methods used to derive such potentials is in ref. [6]. Although such potentials, implicitly incorporating many physical interactions, do not necessarily reflect real energies, their application in protein folding, protein-protein docking, and protein design has achieved impressive successes.

Minimal models of proteins, which have a significantly reduced number of degrees of freedom and give a coarse-grained yet accurate description of polypeptide chains, are widely used to obtain insights into folding mechanisms of proteins, as well as in structure prediction and sequence design. Minimal models also enhance the statistical significance of knowledge-based potentials. In the simplest model, only the α-carbon atoms are considered. In many applications, a sidechain is represented by a center attached to a C_α atom. For a coarse-grained representation of polypeptide chains, interaction potentials, either contact- or distance-dependent, can be extracted from a database of known structures.

The Miyazawa-Jernigan (MJ) contact potential for globular proteins is a widely used knowledge-based potential [2, 7]. In the MJ model, a residue is represented by its side-chain center, which for Gly is taken as the position of its C_α atom. A pair of residues are defined to be in contact if they are not nearest neighbors in sequence and the distance between their centers is less than 6.5Å. The number of different types of residue-residue contacts can be counted directly from the structure of proteins. Miyazawa and Jernigan also introduced an effective solvent molecule, which has the volume of an average residue, to consider the residue-solvent contacts for explicitly including the solvent effect. In this model, residues make the same number of contacts (coordination number) on average, with either effective solvent molecules or other residues. By means of the approximation that the solvent and solute molecules are in quasi-chemical equilibrium and an approximate treatment of the effects of chain connectivity, Miyazawa and Jernigan estimated their interresidue contact energies from known crystal structures of globular proteins.

The high correlation between MJ effective contact energies e_{ab} and the energies required to transfer amino acids from water to less polar environments is well known. Energy e_{ab} between residue types a and b may be decomposed into two components: the desolvation terms e'_{a0}, e'_{b0} and the mixing term e'_{ab}, and then written as $e_{ab} = e'_{ab} - e'_{a0} - e'_{b0}$, where subscript '0' is for the solvent or water. Among different types of contacts, the average difference of the desolvation terms is about 9 times larger than that of the mixing terms. This means that contact energies e_{ab} are dominant by the 'one-body' desolvation

*Address correspondence to this author at the Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China; Tel: 86-10-62541820; Fax: 86-10-62562587; E-mail: zheng@itp.ac.cn

1875-0362/12

term. Similar conclusion can be drawn by an eigenvalue decomposition analysis of the MJ matrix (e_{ab}) [8].

If energy e_{ab} came completely from the 'one-body' term, *i.e.* $e_{ab} = \varepsilon_a + \varepsilon_b$, the one-body ε_a would be well estimated as

$$\varepsilon_a = \frac{1}{210}\sum_{b,c}\frac{1}{2}\left(e_{ab} + e_{ac} - e_{bc}\right). \tag{1}$$

(The difference between ε_a and the least square estimation of Ref. [9] is insignificant.) The error of this one-body approximation is measured by the standard deviation of these 210 estimated values. The values ε_a and their standard deviations are listed in Table **1**. Indeed, the one-body approximation is quite good. (The estimation includes the special case: $\varepsilon_a = e_{aa}/2 = -e'_{a0}$.) It is our main purpose here to explore the origin of the one-body term and the two-body correction based on a probabilistic consideration, rather than to improve the potential function. The MJ energies e_{ij} were first derived in 1985. Later in 1996 the energies were updated based on a much larger database, but no significant effects of the database size were seen. We shall focus on the 1996 version of e_{ij}.

2. METHODS

To make a close correspondence with the 1996 version of e_{ij}, in the present study we use the PDB_select25 of 2001 Sep Release, instead of the current version. The database contains representative protein structures with sequence identity less than 25% [10]. We exclude those with chain lengths less than 50 and those annotated as membrane proteins. The final number of structures used in the analysis is 1274. The center representing the i-th residue a_i of a protein is a point along the joint line from the C_α atom of a_i to its C_β atom and at a distance ℓ from the C_α atom, where values of ℓ depend on the residue types, and, taken from Park & Levitt, have been listed in Table **1** [11]. For Gly, $\ell = 0$, *i.e.* the center is just its C_α atom. A pair of residues are held to be in contact if the distance between their centers are less than 6.5Å and one is not the nearest neighbors of the other along the chain.

2.1. A Probabilistic Model

According to Miyazawa and Jernigan, the total contact energy of a protein native structure is defined as the difference between the energy of its native structure and that of its extended conformation:

$$E = \sum_{a,b} e_{ab} n_{ab}, \tag{2}$$

where n_{ab} is the number of contacts or 'sides' between residue types a and b. If e_{ab} can be contributed completely to the one-body term, *i.e.* $e_{ab} = \varepsilon_a + \varepsilon_b$, we have

$$E = \sum_i \varepsilon_{a_i} \nu_i, \tag{3}$$

where i is the site index, a_i indicates the residue type at site i, and ν_i is the contact number of a_i or its coordination number. In this form of the total contact energy, the only structural information is $\{\nu_i\}$, the coordination number at each site. Correspondingly, it is reasonable to assume that the joint probability $P(A,S)$ for the sequence $A = a_1 a_2 \ldots a_n$ to take its native structure S is

$$P(A,S) = \prod_i P(a_i, \nu_i), \tag{4}$$

where $P(a_i, \nu_i)$ is the shorthand notation for $P(a = a_i, \nu = \nu_i)$, and $P(a, \nu)$ is the probability for a residue of type a to have ν contacts. We may associate an energy U with the logarithmic probability:

$$U = -\log P(A,S) = -\sum_i \log\left[P(a_i, \nu_i)\right]. \tag{5}$$

Usually, instead of the probability, in statistical modeling some probability ratio with respect to certain null model or reference state is considered. Correspondingly, the energy U is associated with a log-odds.

Table 1. Estimated Single-Body ε from the MJ Matrix, their Standard Deviations σ, and the Distances ℓ between the C_α Atoms of Residues and their Representative Centers

	C	M	F	I	L	V	W	Y	A	G
ε	–2.19	–2.60	–3.43	–3.05	–3.44	–2.51	–2.55	–2.16	–1.26	–0.91
σ	0.27	0.25	0.25	0.27	0.27	0.26	0.20	0.17	0.18	0.18
ℓ(Å)	2.0	3.0	3.4	2.3	2.6	2.0	3.9	3.8	1.5	0.0
	T	**S**	**N**	**Q**	**D**	**E**	**H**	**R**	**K**	**P**
ε	–1.03	–0.72	–0.66	–0.76	–0.55	–0.53	–1.34	–0.81	–0.20	–0.86
σ	0.18	0.20	0.25	0.18	0.36	0.33	0.20	0.30	0.33	0.16
ℓ(Å)	1.9	1.9	2.5	3.1	2.5	3.1	3.1	4.1	3.5	1.9

The proper reference state concerns the problem under study. For example, the gapless threading and the Rosetta's fragment assembling should use different reference states. One essential feature of the MJ model missing in $P(A,S)$ or U is the solvent or 'water' effect. In the MJ model residues are in contact not only with other residues, but also with 'unseen' solvent. To connect U with E of Eq. (3), we have to distribute $u_i = -\log P(a_i, \nu_i)$ to its ν_i sides in some way. We have inspected the sign of $\log[P(a \mid \nu)/P(a)]$, where $P(a)$ is the fraction of residue type a in the entire database, and found that the sign changes only once when ν increases from zero for all residue types except Thr. In the picture of MJ, $\nu = 0$ may be interpreted as the case when a residue is fully exposed to the solvent. A very large ν then corresponds to the case when a residue is deeply buried into the interior of a protein. Thus, for a given residue type a, at ν_a^* where $P(a \mid \nu_a^*) = P(a)$, residue a would freely contact with either other residue or water. Increasing ν from ν^* by 1 means that a contact with water is converted to that with residue. Regarding the case when the coordination number of residue type a is always ν_a^* as the reference state, we may estimate the contribution u_a of a contact 'emitted' from residue type a to the energy U as

$$u_a = -\sum_{\nu} \frac{N(a,\nu)}{N(a)} \frac{1}{\nu - \nu_a^*} \log\left[\frac{P(a \mid \nu)}{P(a)}\right], \quad (6)$$

where $N(a,\nu)$ is the total number of residues of type a and with coordination number ν in the entire database, and $N(a) = \sum_{\nu} N(a,\nu)$.

We have only a few discrete integer values of ν. Usually, at any ν, $P(a \mid \nu)$ never exactly equals $P(a)$. We determine ν_a^* by interpolation as:

$$\nu_a^* = \nu_- + \frac{P(a) - P(a \mid \nu_-)}{P(a \mid \nu_+) - P(a \mid \nu_-)}, \quad (7)$$

where $\nu_+ = \nu_- + 1$ and $[P(a \mid \nu_+) - P(a)][P(a \mid \nu_-) - P(a)] < 0$.

2.2. Two-Body Correction

So far, we have ignored any explicit interaction between residues. To include the interaction between residue pairs, we may consider

$$P(A,S) = \prod_i P(a_i, \nu_i) \left[\prod_{j \in \partial a_i} Q(a_j \mid a_i, \nu_i)/Q(a_j)\right]^{1/2}, \quad (8)$$

where ∂a_i denotes the contacts of a_i, $Q(a_j \mid a_i, \nu_i)$ is the probability for a residue of the type a_j to be in contact with a_i conditional on the coordination number of a_i being ν_i, $Q(a_j)$ is the probability for a residue of type a_j to be in contact with any other residues, and power $\frac{1}{2}$ is used to avoid double counting of pairs. We can then derive the two-

Table 2. Counts $N(a,\nu)$, the Total Number of Residues of Type a with Coordination Number Being ν

ν	0	1	2	3	4	5	6	7	8	9	10	11	12	13
A	938	1613	2302	2493	2580	2678	2368	1608	650	227	53	10	2	1
V	358	738	1374	1895	2324	2822	2762	2103	977	355	77	20	1	0
C	34	103	251	443	658	839	765	525	236	82	18	1	0	0
D	1332	2206	2975	2523	1810	1271	713	397	156	42	8	1	0	0
E	1871	3473	3528	2590	1658	1037	570	272	91	29	4	0	0	0
F	279	556	729	1101	1628	1959	1597	932	341	95	23	2	0	0
G	1218	2421	2869	2672	2450	1999	1428	762	328	111	25	4	0	0
H	415	693	855	830	933	733	522	259	82	20	4	0	0	0
I	262	523	973	1408	1990	2422	2375	1725	925	324	68	11	1	0
W	118	200	340	438	556	679	514	279	104	17	3	1	0	0
K	1956	3309	3444	2519	1779	943	427	157	52	13	2	0	0	0
L	400	1055	1664	2285	3190	3737	3580	2462	1135	360	79	9	1	0
M	269	411	522	592	686	826	692	481	255	87	18	2	2	0
N	988	1747	2108	1906	1460	1069	681	340	146	42	10	2	0	0
Y	315	600	770	1071	1434	1590	1322	725	241	73	9	1	0	0
P	1492	1785	2116	1589	1302	1022	669	336	140	41	5	2	0	0
Q	886	1787	1955	1655	1245	822	497	208	88	18	7	1	0	0
R	1152	2114	2327	1971	1564	1128	603	306	110	18	6	0	0	0
S	1049	1995	2727	2497	1992	1611	1219	636	272	68	16	1	2	0
T	668	1497	2518	2412	1993	1628	1177	701	338	102	35	4	0	2

body energy u_{ab} for correction to u_a as

$$u_{ab,\nu} = -\frac{1}{2}\log[Q(b \mid a,\nu)/Q(b)], \quad (9)$$

where the dependence of u_{ab} on ν has been explicitly indicated.

3. RESULTS AND DISCUSSION

The fundamental numbers for analysis are counts $N(a,\nu)$, the total number of residues of type a with coordination number being ν, in the entire database. We do not count each protein separately. These counts are listed in Table **2**. As seen from the table, the counts at large ν are small. We have to combine them into a big bin. Specifically speaking, each ν from 0 to 8 is a bin while all $\nu \geq 9$ form a single bin of $\nu = 9$. Undetermined residue types B, Z and X are included in counting ν, but their statistics is not considered. By estimating $P(a \mid \nu)$ directly from $N(a,\nu)$ (without adding pseudocounts), we calculate ν_a^* using Eq. (7) and then the one-body u_a using Eq. (6). Their values (together with the average coordination numbers $\bar{\nu}_a$) are listed in Table **3**.

3.1. A Binary Partition of Residue States According to the Coordination Number

Our analysis on the two-body correction is based on counts like $N(a,\nu;b)$, which is the number of the 'sides' emitted from a residue of type a with coordination number ν and reaching at a residue of type b. There are $20 \times 20 = 400$ types of pair ab, when distributed to different ν the counts become too small to be statistically significant. Thus, a coarse-graining is considered to give a binary partition of these ν. When we reduce ν into two classes, say '0' and '1', the mutual information between residue type and coordination state

$$I(a;\nu) = \sum_{a,\nu} P(a,\nu)\log\left\{\frac{P(a,\nu)}{P(a)P(\nu)}\right\} \rightarrow$$

$$I(a;\sigma) = \sum_a P(a,\text{'0'})\log\left\{\frac{P(a,\text{'0'})}{P(a)P(\text{'0'})}\right\} + \sum_a P(a,\text{'1'})\log\left\{\frac{P(a,\text{'1'})}{P(a)P(\text{'1'})}\right\}, \quad (10)$$

where $P(\text{'0'}) = \sum_{\nu<\nu^c} P(\nu)$, $P(\text{'1'}) = \sum_{\nu\geq\nu^c} P(\nu)$, the definitions of $P(a,\text{'0'})$ and $P(a,\text{'1'})$ are analogous, and ν^c is the parameter for partition. It has been proven that the coarse-grained mutual information $I(a;\sigma)$ is never greater than the original one, and an optimal binary partition can be obtained by maximizing $I(a;\sigma)$ with respect to ν^c [12]. We find that the optimal partition is at $\nu^c = 4$, so we may call '0' the 'exposed', and '1' the 'buried'. Note that ν^c is very close to the values ν_a^* listed in Table **3**. (For Thr, its sign of $\log[P(a,\nu)/P(a)]$ changes twice; only the one closest to $\nu^c = 4$ is kept.)

3.2. Additivity of the Contribution to u_a from Each Contact

In the calculation of u_a according to Eq. (6), it is implied that the contribution to u_a from each contact is additive. To examine the additivity, we plot $u_{a,\nu} = -\log[P(a \mid \nu)/P(a)]$ versus ν in Figs. (**1a** and **1b**), where $\nu = 9$ is actually the combined bin of $\nu \geq 9$. The additivity corresponds to the linearity of the curves. Roughly speaking, this linearity is still recognizable, but two slopes separated at $\nu^c = 4$ are seen rather clearly for most residue types.

3.3. Comparison Among u_a, MJ's ε_a and some Hydrophobicity Scales

The one-body terms ε_a of MJ's contact energies have been attributed to the hydrophobic effect. We notice that the correlation between MJ's ε_a and MJ's 'average contact energies' e_a is very high (with the correlation coefficient

Table 3. The Values ν_a^* for the 'Neutral' Reference State, the Average Coordination Number $\bar{\nu}_a$, and the One-Body Energies u_a Estimated According to the Probabilistic Model

	C	M	F	I	L	V	W	Y	A	G
ν_a^*	3.69	4.11	3.55	3.83	3.71	3.90	3.47	3.44	3.70	4.34
$\bar{\nu}_a$	5.02	4.27	4.50	4.94	4.74	4.80	4.27	4.26	4.04	3.30
u_a	–0.34	–0.19	–0.31	–0.31	–0.28	–0.25	–0.23	–0.24	–0.10	0.08
	T	**S**	**N**	**Q**	**D**	**E**	**H**	**R**	**K**	**P**
ν_a^*	4.43	3.84	3.80	3.66	3.72	3.26	5.15	3.64	3.37	2.91
$\bar{\nu}_a$	3.51	3.25	2.95	2.74	2.84	2.41	3.39	2.78	2.31	2.76
u_a	0.08	0.12	0.17	0.21	0.21	0.31	0.07	0.18	0.29	0.20

Table 4. Correlations between u_a, ε_a and some Hydrophobicity Scales. WS: the Scale of Wertz-Scheraga; Guy: the Scale of Guy Averaged Over four Datasets; w_a : the Scale Based on the Binary Partition at $\nu^c = 4$

	ε_a	u_a	WS	Guy
u_a	0.935			
WS	0.914	0.906		
Guy	0.953	0.941	0.952	
w_a	0.915	0.992	0.893	0.943

$r = 0.996$), and the strong correlation between e_a and the experimental transfer free energies of amino acids have been examined [7].

Several hydrophobicity scales for amino acid side chains based on statistical analyses of residue distributions of known structures have been reported [13-16]. In these analyses it is assumed that residues are distributed between the surface and interior of the protein in the same way they would be distributed between water and a solvent with a polarity similar to that of the protein's interior. For example, Wertz and Scheraga (WS) classified all residues as being either buried in the protein or exposed to water based on the number of times that a grid of lines parallel with each of three orthogonal axes intersects the solvent exposed van der Waals surface of each residue. An apparent transfer energy for residue type a is then estimated as

$$\Delta F_a = RT \log(M_a^e / M_a^b), \tag{11}$$

where M_a^e and M_a^b are the mole fractions of residue type a that are exposed to water and buried in the protein, respectively. Most side chains are neither entirely buried nor entirely exposed. Taking this into account, Guy modeled the spacial distribution of residues as a function of their distance from the protein surface by dividing the protein ellipsoid 'body' into layers parallel to the protein surface [17]. This layer analysis leaded to a refined hydrophobicity scale. We have examined correlations between u_a, MJ's ε_a, WS's scale and Guy's scale. We have also added a scale w_a derived from the similar version of binary partition according to $\nu^c = 4$. The comparison among them is displayed in Table **4**. It is seen that u_a has a strong correlation with MJ's ε_a, but the correlation of Guy's scale is even stronger. (We have also examined the reference state with the above 'neutral' ν^* replaced by $\nu^* = 0$ or $\nu^* = q_a$, MJ's average coordination numbers. However, the correlations between the obtained u_a and MJ's ε_a are much weaker.)

The linear regression lines of MJ's ε_a versus the probabilistic energies u_a is shown in Fig. (**2**). In fact, the regression between ε_a and u_a (or Guy's hydrophobicity scale ΔF_a) shows rather strong correlation for polar as well as apolar residues, except for typical hydrophobic residue types Leu, Phe and Ala, besides Cys which also involves the disulfide bond. The regression line between ε_a and u_a seems to imply that the correlation of u_a with ε_a is not restricted just on apolar residues.

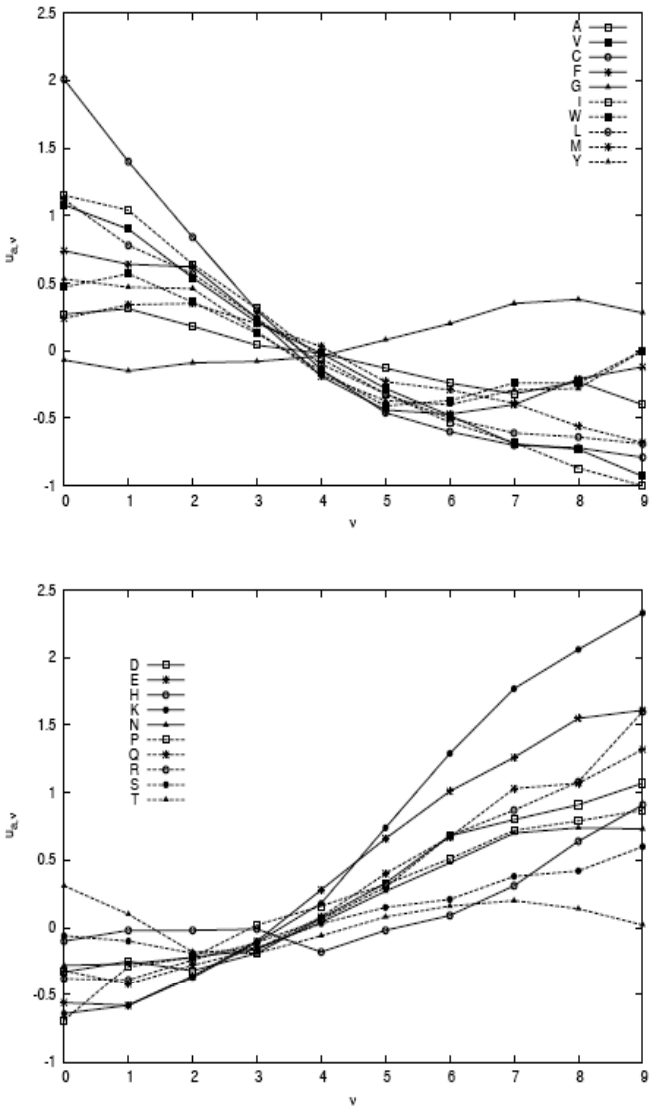

Fig. (1). Propensity of a to ν, $u_{a,\nu} = -\log[P(a|\nu)/P(a)]$ versus ν. Fig. (**1a**) is for $a \in \{G,A,V,C,F,I,W,L,M,Y\}$, and Fig. (**1b**) for $a \in \{D,E,N,Q,K,H,R,S,T,P\}$.

3.4. Two-Body Corrections

Instead of using (9), the two-body corrections are calculated for the binary partition of ν at $\nu^c = 4$ into '0' (exposed) and '1' (buried). The values of $u_{ab,\sigma} = -\frac{1}{2}\log[Q(b \mid a,\sigma)/Q(b)]$ with $\sigma \in \{\text{'0', '1'}\}$ are listed in Tables **5a** and **5b**. (Note that the values there have been multiplied by 200.) The two tables both show some asymmetry. The most striking difference between the two tables is: while the interactions between apolar residues are stronger in the exposed environment, the interactions between polar residues are stronger in the buried environment. This is consistent with the physics of residue interaction: apolar residues near the protein surface strongly attract each other to reduce their areas accessible to water, while ionizable sidechain groups in the protein interior are virtually always uncharged to reduce the permittivity effect. (There are some charges in the interior, but they are almost always functional.) It is not surprising that Cys, involving in disulfide bonds, behaves peculiarly in Cys-Cys pairing. In the tables, the environment propensities $u_{a,\sigma} = -\log[P(a \mid \sigma)/P(a)]$ of residues are also listed. (Our scale w_a of binary partition corresponds to $w_a = u_{a,\text{'1'}} - u_{a,\text{'0'}}$.)

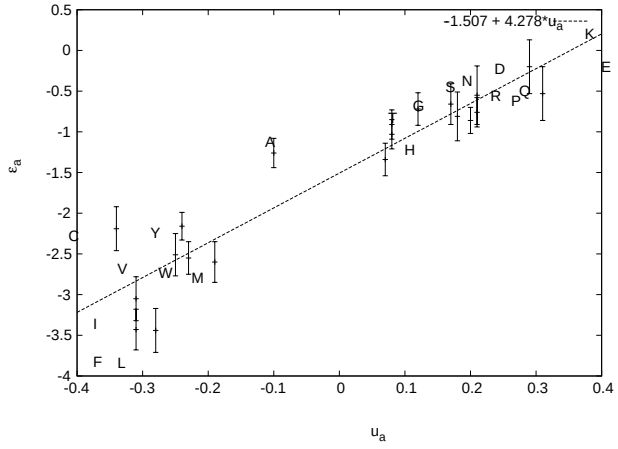

Fig. (2). MJ's ε_a versus the probabilistic energies u_a. The regression line is $\varepsilon_a = 4.278u_a - 1.507$.

In consistence with the demixing effect, the magnitude of diagonal entries is generally larger than that of the nondiagonal ones. The correlation between e_{ab} and $u_{ab} = u_a + u_b$ without two-body correction is $r = 0.921$ while the correlation using the one-body values ($w_a + w_b$) of the binary partition is $r = 0.902$. To include the two-body correction, we consider

$$u_{ab} = w_a + w_b + (u_{ab,\sigma} + u_{ba,\sigma}), \tag{12}$$

with $\sigma \in \{\text{'0','1'}\}$. The correlation between e_{ab} and this corrected u_{ab} is $r = 0.911$ for $\sigma =$'0', and $r = 0.909$ for $\sigma =$'1'. (The two-body correction deduced from the binary partition is able to improve the correlation between $u_{ab} = u_a + u_b$ and MJ's e_{ab} to $r = 0.931$, but only when one fifth of the correction is kept.)

3.5. Contacts Defined by C_α Atoms

In the simplest representation, the polypeptide chains are modeled using only the C_α atoms to provide a global picture of low-resolution structures. It is worth examining contact energies v_a with the contact definition involving only C_α atoms. Generally, when contacts are defined by a cutoff distance between C_α atoms, the correlation of v_a with ε_a is much weaker than that of u_a. At cutoff $R_c = 7.5$Å the correlation is only 0.86. Thus, proper representative centers for sidechains are advantageous to the quality and utility of v_a.

We have tried the simplest way to assume that the sidechain center of the i-th residue is along the direction from the midpoint of the line joining the $(i \pm 1)$-th C_α atoms to the i-th C_α, and at the distance ℓ associated with the sidechain residue type. The correlation between ε_a and u_a is $r = 0.929$, and the regression line is $\varepsilon_a = 4.20u_a - 1.49$. A more careful approximation takes the mean orientation of the sidechain center into account (mainly an off-plane angle of 40.1°), but the change in correlation is almost negligible. Although the simplest models are useful, a certain degree of complexity is needed for more realistic applications.

4. CONCLUDING REMARKS

Knowledge-based potentials for simplified models of protein are essential to understanding the protein structure and folding dynamics. Among numerous forms of potential functions for coarse-grained protein structures, the most widely used one is the weighted linear sum over pairwise contacts, which provides the great computational expediency. Such potentials include the MJ contact energies as a special case. It is not so obvious that many of them may be approximated with a simple function of individual residue properties such as hydrophobicity, demixing, and electrostatics [9]. (Note that the 'one-body' approximations of Ref. [9] may contain also 'two-body' terms in the sense of interactions, and that the one-body term of the MJ contact energies is the contribution of *each contact* associated with a residue of a given type to the energy of a protein.)

According to the equilibrium statistical thermodynamics, the relative probability $P(s)$ of a microstate s with energy $u(s)$ is given by the Boltzmann distribution $P(s) = Z^{-1}e^{-\beta u(s)}$, where $\beta = 1/kT$, k is the Boltzmann constant and Z the normalization factor or the partition function. This Boltzmann distribution is assumed in most statistical potential functions. However, from the viewpoint of probabilistic modelling, the derivation of knowledge-based force fields need not rely on statistical mechanics. A model constructed based on certain conditional independency can be assessed by the fit of that model to the

Table 5a. Two-Body Corrections $u_{ab,\,\text{`0'}}$ (Multiplied by 200) for the 'Exposed' Environment (with rows for $a$, Except for the Last Row which is $u_{a,\,\text{`0'}}$ Multiplied by 100)

	A	V	I	L	M	F	W	Y	H	K	R	D	E	N	Q	S	T	G	P	C
A	–43	–4	0	–1	2	14	33	8	20	12	5	3	17	0	6	–4	3	–9	6	23
V	2	–31	–30	–22	–15	–14	–2	–3	3	22	11	41	11	11	6	16	–2	4	–5	–14
I	–2	–28	–42	–30	–30	–22	–9	–20	–8	18	11	36	27	23	12	22	12	11	3	6
L	–7	–19	–34	–43	–29	–25	–13	–18	–1	18	4	41	19	25	–6	26	24	23	6	10
M	–4	–6	–26	–26	–56	–28	–24	–19	–5	14	11	21	11	23	–1	29	14	3	3	40
F	28	–13	–13	–30	–35	–57	–51	–49	–13	6	–18	39	30	13	13	26	46	19	–5	–10
W	22	20	–16	–14	–12	–29	–83	–34	–30	–5	–39	41	18	5	1	21	49	20	–30	37
Y	20	3	–26	–16	–27	–34	–41	–44	–15	6	–14	31	21	22	–8	22	23	8	–10	0
H	18	1	2	–4	–16	–23	–22	–16	–46	22	1	–9	0	2	8	2	18	7	–10	–1
K	8	7	8	7	16	12	–2	–6	16	40	57	–44	–54	1	–0	5	4	4	18	31
R	4	3	3	–1	9	–11	–17	–7	5	67	26	–33	–45	14	2	7	11	11	–5	14
D	5	33	32	31	12	24	38	33	–7	–51	–33	7	25	–25	0	–25	–13	8	9	35
E	12	6	7	7	9	25	20	19	0	–60	–43	29	30	3	9	–8	–10	26	–9	55
N	3	17	27	23	29	10	17	13	–4	1	8	–31	–0	–26	–12	–11	–13	–0	7	34
Q	5	3	3	–16	–13	10	–15	3	10	–1	4	5	12	–1	–13	–1	–4	8	–11	15
S	–2	12	22	18	24	22	26	27	–2	3	5	–26	–16	–6	–3	–16	–7	–0	1	6
T	1	4	14	25	24	31	32	30	14	5	15	–11	–16	–15	–12	–13	–24	–12	4	23
G	–12	–8	–2	13	4	–1	12	–1	13	10	11	20	35	–5	15	5	–8	–50	–2	–39
P	8	–9	2	7	–2	–23	–43	–22	–18	17	–7	21	–1	13	–10	5	13	–4	–10	–10
C	16	2	6	–0	–9	–3	52	–10	–31	48	23	47	62	60	17	34	35	11	–5	–254
$u_{a,\,\text{`0'}}$	19	61	73	63	31	56	41	40	–3	–42	–28	–28	–40	–24	–30	–15	–7	–11	–27	88

Table 5b. Two-Body Corrections $u_{ab,\,\text{`1'}}$ (Multiplied by 200) for the 'Buried' Environment (with Rows for $a$, Except for the Last Row which is $u_{a,\,\text{`1'}}$ Multiplied by 100)

	A	V	I	L	M	F	W	Y	H	K	R	D	E	N	Q	S	T	G	P	C
A	–22	–11	–10	–5	–1	12	17	14	20	14	11	21	16	13	12	8	1	–11	11	15
V	–11	–29	–20	–16	–4	–3	9	4	22	16	25	39	28	35	22	25	12	17	10	22
I	–9	–20	–28	–19	–8	–7	5	–6	25	15	18	38	20	40	21	29	12	28	12	23
L	–4	–16	–18	–28	–8	–11	–8	–8	12	15	13	39	21	37	7	29	19	34	22	21
M	–0	–5	–8	–8	–31	–19	–12	–21	4	14	15	22	15	16	1	19	17	15	2	10
F	10	–4	–9	–11	–18	–41	–25	–23	1	21	15	27	19	19	13	18	23	22	–4	10
W	20	8	7	–7	–12	–26	–40	–24	–12	6	–10	6	5	12	–4	13	22	18	–34	17
Y	13	5	–4	–7	–19	–24	–23	–20	–9	–10	1	12	10	5	4	16	24	17	–18	13
H	21	25	26	15	9	5	–13	–7	–48	3	–11	–40	–29	–21	–6	–16	–4	–2	–9	–20
K	16	23	20	22	11	17	11	–3	10	4	38	–73	–82	–17	–23	–13	–10	1	6	22
R	13	33	25	19	18	16	–15	3	–12	24	9	–67	–72	–13	–17	–15	–8	–4	–21	23
D	22	45	44	48	29	34	7	13	–42	–71	–69	–20	–5	–53	–23	–39	–27	–23	–0	41
E	18	36	32	32	18	17	5	12	–31	–74	–73	–7	–4	–32	–15	–32	–20	5	–15	37
N	13	38	44	43	15	22	10	11	–18	–22	–9	–47	–30	–46	–21	–35	–29	–33	–12	22
Q	13	27	30	17	9	14	9	2	–7	–24	–18	–25	–19	–30	–34	–21	–22	–11	–10	13
S	7	27	31	32	20	19	11	15	–15	–10	–13	–37	–22	–37	–18	–31	–23	–25	–12	3
T	2	12	12	19	15	25	27	24	–2	–12	–10	–27	–16	–28	–16	–20	–20	–19	–6	13
G	–12	18	30	33	13	23	17	19	–4	10	2	–24	11	–27	–7	–26	–19	–52	–6	5
P	8	10	11	21	2	–1	–29	–16	–1	16	–13	–3	–13	–11	–3	–13	–10	–9	–16	–4
C	16	21	23	22	15	10	15	12	–16	21	18	33	38	17	11	–1	11	–2	–2	–178
$u_{a,\,\text{`1'}}$	–16	–38	–43	–39	–24	–37	–29	–29	3	76	40	41	71	32	45	18	8	12	39	–47

data. The Boltzmann distribution as an exponential form is related to the exponential decay of the probability measures for certain tail events in the so-called large deviation theory [18]. An example of deriving a 'temperature' factor for

protein stability is given in Ref. [19] (see Eq. (16.8) there). We have derived the one-body term based on a probabilistic model directly in the framework of the MJ energy of a protein. Although no direct relation is available between the scaling factor of our u_a and the 'room temperature', the correlation between u_a and MJ's ε_a is evident. The amino acid hydrophobicity estimated from a layer analysis by Guy showed strong correlation with MJ's ε_a, but their frameworks for the energy of a protein differ. The coordination number carries some interpretation as a layer index, so Guy's hydrophobicities strongly correlate with our u_a or w_a.

The structural information most concerning one-body energies is coordination numbers, which play an essential role in our analysis. According to the one-dimesional mean-field-like approximation of Ref. [20], the correlation between the mean coordination number $\bar{\nu}_a$ and one-body energy ε_a should be strong. Indeed, this is verified by our calculation of the correlation, $r = -0.909$.

More sophisticated contact energy potentials have been proposed, which incorporate several features of residues, such as their solvent exposure, their secondary structures, or orientation of side chains. For example, effective contact energies for an expanded 60-residue alphabet (including three secondary structural classes) have been estimated [21]. Using an entropic clustering approach, we have obtained two coarse-grained states by maximizing the mutual information between coordination numbers and residue types. Obvious differences between the two states are seen in their two-body corrections. The dependence of residue pair correlations on structural environment is worth a detailed investigation [22]. Nonlocal sidechain contacts between regular secondary structure elements, those cross-strand or those between different helices or β-sheets, which show stronger correlation than that of local contacts [22], would play a role different from local ones. Especially, the contacts between different helices and/or β-sheets should be more responsible to forming the stable structure core than other contacts, and would be expected being more conservative in sequence. Inclusion of such features in designing new contact potentials is important in practice.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China.

CONFLICTS OF INTEREST

None Declared.

REFERENCES

[1] S. Tanaka and H. A. Scheraga, "Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins", *Macromolecules*, vol. 9, pp. 945-950, 1976.

[2] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation", *Macromolecules*, vol. 18, pp. 534-552, 1985.

[3] R. L. Jernigan and I. Bahar, "Structure-derived potentials and protein simulations", *Curr. Opin. Struct. Biol.*, vol. 6, pp. 195-209, 1996.

[4] M. J. Sippl, "Knowledge-based potentials for proteins", *Curr. Opin. Struct. Biol.*, vol. 5, pp. 229-235, 1995.

[5] X. Li and J. Liang, "Knowledge-based energy functions for computational studies of proteins", In: *Computational Methods for Protein Structure Prediction and Modeling*, Springer: Berlin 2007, pp. 71-123.

[6] S. P. Leelananda, X. P. Feng, P. Gniewek, A. Kloczkowski, and R. L. Jernigan, "Statistical contact potentials in protein coarse-grained modeling: from pair to multi-body potentials", in *Multiscale Approaches to Protein Modeling*, Springer: Berlin 2011, pp. 127-157.

[7] S. Miyazawa and R. L. Jernigan, "Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term", *J. Mol. Biol.*, vol. 256, pp. 623-644, 1996.

[8] H. Li, C. Tang, and N.S. Wingreen, "Nature of driving force for protein folding: A result from analyzing the statistical potential", *Phys. Rev. Lett.*, vol. 79, pp. 765-768, 1997.

[9] P. Pokarowski, A. Kloczkowski, R. L. Jernigan, N. S. Kothari, M. Pokarowska, and A. Kolinski, "Inferring ideal amino acid interaction forms from statistical protein contact potentials", *Proteins*: *Struct. Func. Bioinf.*, vol. 59, pp. 49-57, 2005.

[10] U. Hobohm and C. Sander, "Enlarged representative set of protein structures", *Protein Sci*., vol. 3, pp. 522-524, 1994.

[11] B. Park and M. Levitt, "Energy functions that discriminate X-ray and near-native folds from well-constructed decoys", *J. Mol. Biol.*, vol. 258, pp. 367-392, 1996.

[12] W. M. Zheng, "Entropic approach for reduction of amino acid alphabets", [Available from: arxiv.org/abs/physics/0106074].

[13] C. Chothia, "The nature of the accessible and buried surfaces in proteins", *J. Mol. Biol*., vol. 105, pp. 1-14, 1976.

[14] D. H. Wertz and H. A. Scheraga, "Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule", *Macromolecules*, vol. 11(1), pp. 9-15, 1978.

[15] J. Janin, "Surface and inside volumes in globular proteins", *Nature*, vol. 277, pp. 491-492, 1979.

[16] B. Robson and D. J. Osguthorpe, "Refined models for computer simulation of protein folding", *J. Mol. Biol.*, vol. 132, pp. 19-51, 1979.

[17] H. R. Guy, "Amino acid side-chain partition energies and distribution of residues in soluble proteins", *Biophys. J.*, vol. 47(1), pp. 61-70, 1985.

[18] H. Touchette, "The large deviation approach to statistical mechanics", *Physics Reports*, vol. 478, pp. 1-69, 2009.

[19] A. V. Finkelstein and O. B. Ptitsyn, *Protein Physics*. Academic Press: Boston 2002.

[20] A. R. Kinjo and S. Miyazawa, "On the optimal contact potential of proteins", *Chem. Phys. Lett.*, vol. 451, pp. 132-135, 2008.

[21] C. Zhang and S. H. Kim, "Environment-dependent residue contact energies for proteins", *Proc. Natl. Acad. Sci. USA.*, vol. 97, pp. 2550-2555, 2000.

[22] A. P. Cootes, P. M. G. Curmi, R. Cunningham, C. Donnelly, and A. E. Torda, "The dependence of amino acid pair correlations on structural environment", *Proteins*, vol. 32, pp. 175-189, 1998.

Received: November 21, 2011 Revised: December 20, 2011 Accepted: December 23, 2011