Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 609–621
polymer
www.elsevier.com/locate/polymer

Analysis of the amino acid effect on protein folding by atom pair contacts

Zhouting Jiang[a], Linxi Zhang[a,*], Jin Chen[a], Agen Xia[a], Delu Zhao[b]

[a]*Department of Physics, Zhejiang University, Hangzhou 310027, People's Republic of China*
[b]*Polymer Physics Laboratory, Center of Molecular Science, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100080, People's Republic of China*

Received 20 July 2003; received in revised form 22 October 2003; accepted 22 October 2003

Abstract

New atomic pair contacts with considering the coordinates of each atom in a residue are introduced here. We analyze the ability of all the 20 amino acid residues to form long-range and short-range contacts by calculating the average numbers of short- and long-range contacts between different amino acid pairs. It is concluded that Phe-Phe, Leu-Phe and Leu-Leu have a high tendency to form contacts. The relative ability to form atom pair contact does not depend on the limiting value of R_C. The average number of contacts per residue, which is the scale of the relative ability to form contacts for the 20 amino acid residue types, is also calculated. The result shows that hydrophobic residues with large numbers of long-range contacts more easily form long-range contacts, while the hydrophilic ones form long-range contacts less often. Linear regression analysis by a new method of counting contacts concludes that either contact order (CO) or total contact distance (TCD) parameter has a significant correlation with the logarithms of folding rates. The relative deviations between the experimentally observed ln k_f and the two parameters CO and TCD are smaller than that with previous methods. Moreover, the values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ between λ-type and μ-type amino acids are investigated. Comparisons between the Fauchere–Pliska hydrophobicity scale and the average number of contacts per residue formed are also made. The new knowledge of atomic pair contacts can help us understand the importance of amino acid residue type and its sequence in globular structure of the protein in detail.

Keywords: Atom pair contact; Contact order and total contact distance; Folding rate of protein

1. Introduction

Proteins are compact polymers [1]. There are many types of interactions in proteins, which play an important role in protein folding and the stability of a protein molecule. It is now well established that for small proteins the information contained in the amino acid sequence is sufficient to determine the folded structure [2,3]. The protein folding problem, i.e. prediction of a protein's (unknown) native structure from its (known) amino acid sequence is one of the most challenging open problems in computational physics, chemistry and biology. A related important task is to understand the relationship between sequences and folding rates of proteins. The folding rate of proteins that fold with two- or weakly three-state kinetics has a significant correlation with the average sequence separation of all contacting residues in the native state, defined by the parameter contact order (CO) [4], and the summation over all contacts, defined by the parameter total contact distance (TCD) [5]. Debe and Coddard presented a first principles approach based on a nucleation–condensation folding mechanism for predicting the experimentally determined folding rates [6]. Later, Neutral network methods were used for predicting the folding rate of proteins [7,8]. Miller et al. provide topological parameters to determine protein-folding rate [9]. However, the exact relationships between folding rates and protein structures have not been determined clearly yet. If we can learn the relationship between folding rate and protein structure, we can use these methods to predict the folding rate of many proteins prior to experiment. Among the interactions responsible for the stabilization of the native structures in globular proteins, those occurring between sequentially distant but spatially close amino acid residues are recognized to play a dominant role. The native state is stabilized by various residue-specific, non-bonded interactions that hold a protein together in a compact form. Although the determination of protein sequences may be fast, determining the detailed

* Corresponding author. Tel.: +86-571-87951328; fax: +86-571-87951328.
E-mail address: zhanglx@mail.hz.zj.cn (L. Zhang).

0032-3861/$ - see front matter © 2003 Published by Elsevier Ltd.
doi:10.1016/j.polymer.2003.10.067

three-dimensional structure of a protein is far more time consuming. The tertiary interaction problem is more difficult and understanding the nature of such interactions is crucial for protein structure prediction. The prediction of protein structures from their amino acid sequences remains an unsolved problem in computational biochemistry.

A common strategy for approaching the protein folding problem is to use simplified representations and energy functions. We begin with considering a zeroth order model in which all native interactions in a protein are equally favorable, i.e. a homogeneous contact model. In such a model, the free energy cost of forming different contacts in a protein depends solely on the entropic cost of restricting the chain to allow the contact. When many of the contacts in a protein are between residues distant in the primary sequence, a large portion of the chain must be ordered before even a few favorable contacts can form, leading to a large folding free energy barrier. There are many types of contacts (e.g. hydrogen bonds, hydrophobic–hydrophobic, aromatic–aromatic, aromatic-polar, etc.) in calculating potential energy of the protein [10]. Such an examination has already been described for one contact type [11]. Up to now, two elements (amino acids, ligands, etc.) were postulated as forming a contact if the closest distance between their atoms falls below a certain threshold [12,13]. For example, although a coarse-grained protein model, which treats amino acid residues as united interaction sites, offers a more practical approach for tackling the protein folding problem [14], this model omits many detailed features of atomic interactions. Recently atom pair interactions have been discussed [15]. Although we have discussed the importance of long-range contacts on the stability of protein structure by calculating long-range residue–residue contacts, [16] we ignore the fact that there are many types of contacts (e.g. hydrogen bonds, hydrophobic–hydrophobic, aromatic–aromatic, aromatic-polar, etc.) between two amino acid residues. In this paper we present a new method for considering the contacts between the atoms of two amino acid residues. This method may consider several contacts simultaneously between two amino acid residues. From our calculation, it is found that atom pair contacts can provide more detailed information about protein folding and the stability of protein structure.

2. Materials and methods

2.1. Database

In this paper, we have selected 200 proteins as the data source for our present work. The coordinates for each protein analyzed are obtained from the Protein Data Bank (PDB, www.rcsb.org). The selected proteins are non-homologous and the structures are determined to high resolution (resolution <2.5 Å). The selected proteins are from four different structural classes, namely, all-α, all-β, α + β and α/β, each with a set of 50 proteins. We obtain the information about the structural class from SCOP database [17]. The PDB codes for all the protein samples used in the present study are given in Table 1.

2.2. Computation of surrounding residues and contacts

The computation of surrounding residues in a protein molecule has been described in our earlier article [16]. In that paper, the residues in protein molecules are represented only by their C_α atoms. This kind of simple method to simulate the position of the residue is widely used [18–23]. However, we think that attractive interactions between atoms may lead to the formation of contacts. In our present work, we investigate atomic pair contacts by considering each coordinate of the heavy atoms (i.e. non-hydrogen atom) in the residues. An atom pair whose distance is closer than R_C is defined as an atom–atom contact which contributes to the interactions between amino acids. If two residues have several atomic pair contacts, in general there may be a stronger attractive interaction between two residues on average. Of course, the attractive interactions depend on the types of atoms and contacts, and here we discuss attractive interactions in the view of their statistics. It is reasonable to compare the intensity of atomic pair contact interactions, in some further detail. In previous calculation of contacts, this attractive interaction cannot be considered. There can be more than one atom–atom pair contact between two amino acid residues [15]. In the following discussion, it is found that this model is more reasonable for investigating statistical properties of proteins. Here the limiting values $R_C = 6.5$, 8.0 and 10.0 Å for contacts are chosen. These limits are widely used to describe the protein's structure and character of the folding behavior [22–26]. For a given residue, the composition of surrounding residues is analyzed in terms of the location at the sequence level, and the contributions from non-neighboring residues are treated as a contact. It is widely considered that the contributions from $\leq \pm 4$ residues are treated as short-range contacts, and $> \pm 4$ residues as long-range. Using the heavy atom coordinates, a sphere of radius R_C is centered on each atom, and the composition of all surrounding atoms is calculated. Although the contact is calculated between atoms, it affects only the count method. The mechanism of the protein folding is finally attributed to the interactions between the amino acids that the atoms belong to.

In order to solve the problem of the different occurrences for different amino acid residues in the 200 proteins, we define the average number of contacts $N_{\lambda\text{-}\mu}$ as

$$N_{\lambda\text{-}\mu} = \frac{n_{\lambda\text{-}\mu}}{\sqrt{n_\lambda n_\mu}}, \quad (\lambda \text{ or } \mu = \text{Ala, Asp, Cys, Glu, …, Tyr}) \quad (1)$$

Here n_λ and n_μ are the total number of λ- and μ-type residues in the 200 protein chains, respectively. $n_{\lambda\text{-}\mu}$ represents the total number of atom pair contacts between λ-type and μ-type residues in the 200 protein chains. $N_{\lambda\text{-}\mu}$

Table 1
The PDB codes of proteins used in this paper

All-α proteins	1ALA	1BBL	1BP2	1CCR	1EA8	1ECA	1ECD	1ECO	1F63	1FCS
	1FHA	1FIP	1H96	1HBG	1IFA	1LE4	1LH1	1LH2	1LPE	1MBA
	1MBC	1MBD	1MBS	1PPA	1RCB	1RRO	1UTG	1YCC	1YEA	2C2C
	2CDV	2CTS	2CY3	2CYP	2END	2FAL	2HBG	2LHB	2MHR	2PDE
	2WRP	3C2C	451C	4BP2	4CPV	4ICB	4MBN	5CPV	5CYT	5TNC
All-β proteins	1A45	1ACX	1CA2	1CD8	1CID	1EST	1F3G	1F53	1GPR	1H6X
	1HJC	1HOE	1IFC	1KL9	1MPP	1NN2	1PAZ	1PYP	1QNY	1RBP
	1SGT	1STP	1TEN	1TIE	1TLK	1TON	2ALP	2APR	2AYH	2BB2
	2CA2	2CAB	2CNA	2ER7	2GCH	2ILA	2MCM	2PCY	2REN	2RHE
	2SGA	2SNS	2SNV	2STV	3APP	3CNA	3EST	4FGF	4PEP	5PTP
α + β proteins	102L	125L	190L	1AQP	1BKF	1CTF	1CYO	1D9W	1DUR	1E3V
	1EZM	1FCL	1FD6-1	1FD6-6	1FD6-9	1FDD	1FKB	1FKD	1FKF	1FRH
	1FXD	1GWD	1I1Z	1I20	1IET	1IEU-1	1IEU-6	1L3F	1LHH	1LHI
	1LZ1	1POP	1PPN	1ROB	2AAK	2ACT	2LYZ	2LZM	3ILB	3LYZ
	3RN3	3SSI	4LZM	4TLN	4TMS	5FD1	7RSA	8TLN	9PAP	9RNT
α/β proteins	1ABA	1ABE	1ALD	1AZL	1C4W	1DHR	1E49	1E6K	1EAF	1ETU
	1EX7	1F4P	1FX1	1GKY	1GOX	1IPD	1OFV	1OVB	1Q21	1RHD
	1RNH	1SBP	1TFD	1THM	1ULA	2DRI	2FCR	2FOX	2FX2	2GBP
	2HAD	2LIV	2PRK	2SBT	2TS1	3ADK	3CHY	3CLA	3CPA	3DFR
	3PGK	3PGM	4CLA	4ICD	4PFK	5ABP	5CPA	5P21	6XIA	8DFR

describes the tendency of forming atom pair contacts between λ-type and μ-type residues. Here we only consider the ability of forming atom pair contacts between two amino acid residues. $N_{\lambda\text{-}\mu}$ represents the statistical properties of 200 protein chains, and we do not discuss the ability of forming atom pair contacts for a special protein chain.

Moreover, the average number of contacts per residue indicates the ability of forming contacts for residues. Because of the different number of atoms for the 20 amino acid residues, we use the parameter of effective atom coefficient p_λ, which shows the different effects of different atoms (such as C, O, N, and S) on the folding process. The reason is that the ability to form atom pair contacts is different for C, O, N and S atoms. The values of p_λ are obtained with a statistical method that compensates the deletion of the atom coordinate in PDB files. Here we also define the average number of short-range contacts per residue C_S and the average number of long-range contacts per residue C_L as

$$C_{\lambda,\eta} = \frac{\sum_{\mu=\mathrm{Ala,Asp,\ldots,Tyr}} n_{\lambda\text{-}\mu,\eta}}{8.5 \cdot n_\lambda \cdot p_\lambda}, \quad (\eta = \mathrm{S\ or\ L},\ \lambda = \mathrm{Ala, Asp, \ldots, Tyr}) \tag{2}$$

These calculations can help us know which residue plays an important role in the protein folding and the stability of protein molecule. Here the coefficient 8.5 is the average number of atoms in the 20 amino acid residues (excluding H atoms).

2.3. Computation the distance of contact and folding rate prediction

Several years ago, it was found that there is a relationship between the folding rate and the average sequence separation between contacting residues [4–9]. In the simple zeroth order model, increasing uniformly the strength of all interactions clearly reduces the free energy barrier to folding (the unfavorable entropy of ordering is better compensated by the formation of more favorable interactions) and so the folding rate would increase. Indeed, there is a nearly linear correlation between folding rate and stability for a given protein [4,5]. The CO is defined as

$$\mathrm{CO} = \frac{1}{n_c n_r} \sum_{|i-j|>l_{cut}}^{n_c} |i-j| \tag{3}$$

Here n_r is number of amino acid residues of a protein (excluding disordered regions), and n_c is the number of non-local atom–atom contacts. In general, n_r is total number of amino acid residues in a protein. If a protein includes disordered regions, the disordered regions are not considered in calculating the CO [5]. A non-local contact is defined as two heavy atoms within a cutoff distance R_C and separated by at least a residue separation cutoff value l_{cut}. In order to compare with previous works [4,5], we choose $R_C = 6.5$, 8.0, 10.0 Å and $l_{cut} = 2$. Recently, TCD has also been used to predict folding rates [5]. The expression for TCD is

$$\mathrm{TCD} = \frac{1}{n_r^2} \sum_{|i-j|>l_{cut}}^{n_c} |i-j| \tag{4}$$

In fact, CO means the average sequence separation per contact, and TCD is the summation over all the contacts. As

a contact consists of two residues, it is important to investigate the values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ between λ-type and μ-type residues. Here we define $CO_{\lambda\text{-}\mu}$and $TCD_{\lambda\text{-}\mu}$ as

$$CO_{\lambda\text{-}\mu} = \frac{1}{n_{\lambda\text{-}\mu}\sqrt{n_\lambda n_\mu}} \sum_{|i-j|>l_{cut}}^{n_{\lambda\text{-}\mu}} |i-j| \quad (5)$$

$$TCD_{\lambda\text{-}\mu} = \frac{1}{n_\lambda n_\mu} \sum_{|i-j|>l_{cut}}^{n_{\lambda\text{-}\mu}} |i-j| \quad (6)$$

$CO_{\lambda\text{-}\mu}$ is the average sequence separation per λ-μ type of residue–residue contact and $TCD_{\lambda\text{-}\mu}$ is the summation over all the λ-μ residue–residue contacts. Through those calculations, we can know the properties of residue–residue contacts clearly based on the atomic pair contacts. Some applications to predicting the folding rate will be discussed next.

3. Results and discussions

3.1. Occurrence of residues in short- and long-range interactions

We first investigate the ability to form atomic pair contacts for all 20 amino acids types. In the case of R_C = 6.5 Å, the maximum number of long-range atom pair contacts is 58,162 between Leu-Leu, and the minimum number is 1606 between Cys-Met. In the meantime, the maximum and minimum numbers of short-range atom pair contacts is 24,902 and 382, which exist in Leu-Leu and Pro-Cys, respectively. But the ability to form contacts cannot be described well by calculating the number of long-range and short-range atom pair contacts only. Because the following three aspects should be considered: different appearance of amino acid residues in the 200 protein molecules, the different number of atoms in each amino acid residue and the distinct ability to forming the contacts for C, O, N, and S atoms. In the following passage, these three aspects will be discussed one by one.

It is important that the probability distribution of amino acid residues and the number of atomic pair contacts should be considered simultaneously. The average number of atom pair contacts for each amino acid residue pair in the cases of R_C = 6.5, 8.0, and 10.0 Å, obtained from Eq. (1), are shown in Tables 2–4. In Tables 2–4, the number of long-range atom pair contacts is larger than for short-range ones. Comparing Tables 2–4, each value in the same location of 20 × 20 matrix of Table 4 is the largest one. The reason is that there are more atom pair coordinates satisfied in the R_C = 10.0 Å condition. In these Tables, most of values are greater than 1. It means that there is more than one atomic pair contact between two amino acid residues.

In the upper triangle of Table 2, there are three values of atom pair contacts greater than 20.0, i.e. Phe-Phe (31.0), Leu-Phe (24.4) and Leu-Leu (20.8). The other 10 topmost long-range atom pair contacts are Tyr-Tyr, Trp-Phe, Cys-Cys, Ile-Leu, Phe-Ile, Phe-Val, and Ile-Ile, which centralized in the upper left corner of the matrix. In this area are the hydrophobic amino acids, and it means these residue pairs have a greater tendency to form long-range contacts and that the interactions between them can be strong. On the other hand, the minimal average number $N_{\lambda\text{-}\mu}$ is 2.18 between the Cys-Glu residue contacts. The other 10 lowest values are observed between Cys-Lys, Cys-Gln, Met-Gly, Cys-Gly, Met-Asp, Met-Ser, Cys-Met, Cys-Ala, and Cys-Asp. It shows that there is weak ability to form long-range atom pair contact between Cys and Met amino acid residue. The reason for the low tendency to form contacts perhaps lies in the fact that Cys and Met residues have sulfur atoms. In the lower triangle, the maximum value is 8.90 between Leu-Leu, and the other 10 topmost values are Glu-Lys, Glu-Arg, Leu-Phe, Lys-Asp, Arg-Arg, Phe-Phe, Leu-Glu, Leu-Ile, and Glu-Glu, respectively. There is no clear dominant region in the matrix for these large values of the average number of short-range atom pair contacts. We think that the ability to form short-range atom–atom contacts mainly depends on the sequence of amino acid residues. There are 10 values lower than 1, and the minimum is 0.415 for Cys-Pro, and these values often occur with Met, Cys and Pro residues.

We also calculate the average number of short- and long-range atomic pair contacts with R_C = 8.0, and 10.0 Å, and the results are given in Tables 3 and 4, respectively. Although the values of the average number of long-range atomic pair contacts in Tables 3 and 4 are larger than the values in Table 2, the three highest values of long-range atom pair contacts are the same as for R_C = 6.5 Å. Tables 3 and 4 show that the Met and Cys residues may have difficulties in forming contacts, which agrees with Table 2. Almost the same results are found in lower triangle, too. Therefore, the relative ability to form atom pair contacts does not depend on the value of R_C.

Considering single residues, we calculate the average number of short-range atomic pair contacts per residue C_S and the average number of long-range atom pair contacts per residue C_L according to Eq. (2), and the results are given in Table 5. In order to compare with our previous work, we also give our previous results [16]. Considering the different numbers of atoms in the different type of residues and the different probabilities in the proteins, in Eq. (2) we divide the average number of the atoms for all of the 20 amino acid residues into P_λ, in order to scale the atom's capability in forming contacts. The other two aspects of the ability to form contacts discussed above are considered here. The sign $*$ represents our previous works [16]. Although in this article only 200 protein chains are studied, which is fewer than in our previous works, we think that the statistical properties obtained from 200 protein chains are reliable. Using the new way of counting contacts, the values of C_L and C_S are evidently higher than in the previous work. It implies that a residue could form several atom–atom contacts with other amino acid residues. In the present work, the largest value of the average number of long-range

Table 2
Average number $N_{\lambda\text{-}\mu}$ of atomic pair contacts between different amino acid residues λ and μ. The upper triangle counts long-range contacts and the lower triangle counts short-range contacts in protein samples with $R_C = 6.5$ Å

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro
	20.8	17.4	18.9	8.88	24.4	15.9	5.28	16.9	9.72	5.83	9.10	6.80	6.12	6.86	5.55	5.31	7.17	6.23	9.12	5.27
Leu	8.90	17.2	15.7	7.65	18.2	13.3	5.58	12.5	9.14	5.61	8.55	6.05	5.97	6.67	4.86	5.10	6.26	6.65	7.24	5.45
Val	5.04	3.15	17.7	6.98	18.9	15.9	4.42	13.9	8.65	5.19	8.21	5.20	6.35	5.35	5.32	5.16	6.45	6.27	8.58	4.79
Ile	5.51	3.10	3.44	5.47	12.3	8.06	2.62	10.1	4.15	2.51	3.52	3.36	3.50	2.94	2.55	2.89	3.21	2.61	4.69	2.90
Met	3.62	2.57	2.56	3.30	31.0	17.1	6.11	19.6	8.79	6.09	9.90	10.7	6.73	7.20	6.75	6.32	7.50	7.06	9.61	6.73
Phe	6.78	4.13	3.74	3.52	6.03	19.8	6.98	14.8	9.36	6.87	10.1	11.2	8.80	9.66	9.67	8.78	12.5	9.50	14.3	9.27
Tyr	4.49	3.12	2.99	3.23	3.03	3.94	18.9	5.08	2.73	2.54	3.44	2.99	2.18	2.51	2.77	3.64	2.50	3.67	3.28	3.19
Cys	1.69	1.54	1.31	0.836	1.43	1.09	2.70	11.3	7.19	5.95	7.56	9.93	6.17	6.97	6.25	7.58	7.19	5.65	11.4	6.84
Trp	3.58	3.44	2.41	0.823	3.02	2.45	1.85	4.42	4.82	3.39	5.39	3.93	3.47	3.74	3.39	3.57	3.80	4.33	4.20	3.80
Ala	4.59	3.44	3.26	2.17	2.90	2.76	1.42	2.26	4.90	3.67	5.25	2.85	3.18	3.36	4.13	4.03	3.50	4.21	4.59	3.65
Gly	2.54	1.72	1.59	1.14	1.69	1.91	0.750	1.58	1.93	1.44	8.88	5.67	5.23	5.48	6.47	5.87	5.97	6.87	6.41	5.07
Thr	2.71	2.02	2.42	1.66	2.45	2.54	0.95	2.07	2.50	1.65	2.67	6.66	5.46	3.48	5.80	4.59	3.92	5.06	5.84	4.04
His	4.32	2.28	2.32	1.64	4.22	2.36	2.14	1.16	2.18	1.51	2.08	4.86	3.92	4.02	3.76	4.56	8.58	4.37	10.2	3.57
Glu	5.53	3.09	3.70	2.94	3.43	2.62	1.03	2.26	4.43	1.63	3.22	3.08	5.31	3.61	3.73	4.08	4.63	4.46	6.68	4.09
Gln	4.71	2.46	2.90	2.89	2.18	2.94	1.10	2.79	3.53	2.22	2.16	1.51	4.05	3.30	4.11	6.05	7.10	4.69	10.8	3.67
Asp	3.17	2.29	2.64	1.96	2.76	3.17	1.29	1.93	3.75	2.24	2.77	2.54	3.37	3.11	3.42	5.93	5.28	5.16	6.55	3.72
Asn	2.66	2.02	2.29	1.70	2.49	2.88	1.24	2.72	2.81	1.61	2.45	2.03	3.89	3.09	3.29	2.39	4.24	4.56	3.89	3.29
Lys	4.58	3.57	3.32	2.27	3.27	2.93	1.23	1.77	4.43	2.26	2.75	2.55	8.72	4.04	6.68	4.57	4.67	5.82	5.81	3.84
Ser	3.00	2.18	1.97	1.50	2.07	2.96	1.25	1.79	2.76	1.49	2.31	1.42	3.55	3.09	2.92	2.81	2.47	1.98	8.22	4.72
Arg	5.03	3.07	3.01	2.11	2.79	3.80	1.35	2.71	3.67	2.21	3.64	2.63	7.25	5.19	4.93	3.06	2.76	2.69	6.30	4.43
Pro	1.93	1.32	1.36	0.951	1.89	2.33	0.415	0.947	1.65	0.627	1.35	1.39	1.42	1.85	1.59	0.932	2.07	1.32	1.43	0.586

contacts is 26.59 for Trp. And the smallest one is 12.83 for Glu when $R_C = 8.0$ Å. The amino acid residues with the largest or smallest number of long-range contacts in the case of $R_C = 6.5$ Å are the same as for $R_C = 8.0$ Å. In the previous work we concluded that the hydrophobic amino acids have a large value of C_L, and the hydrophilic amino acids have a small value of C_L. The same conclusion is obtained with the new count contact method. In view of the statistical properties of amino acid residues which one is the hydrophobic or not, mainly depends on the average number of long-range contacts per residue. As a matter of the fact, our new method can represent this better than the previous one. For example, in the case of $R_C = 6.5$ Å, although Leu is a hydrophobic amino acid, its value of C_L is lower than many of neutral amino acid residues in previous model [16]. This problem is resolved in our present work. The average values of $\bar{C}_L$ and $\bar{C}_S$ for the three types of amino acids (hydrophobic, neutral, and hydrophilic) are also calculated. These values show us clearly the importance of long-range or short-range contacts in protein folding. In Table 5, the remarkable difference between $\bar{C}_L$ and $\bar{C}_S$ is that $\bar{C}_L$ decreases from 11.33 to 6.767 in the case of $R_{C=}6.5$ Å, while $\bar{C}_S$ is almost the same for the three types of amino acid residues. Although the values in case of $R_C = 8.0$ Å are larger than for $R_C = 6.5$ Å, the tendency is the same. The obvious different tendency for H residues and P residues for $\bar{C}_L$ shows that the average number of long-range contacts per residue can provide some insights into protein folding and the stability of protein structure. The hydrophobic residue with large values of C_L is located in the interior of globular protein and easily forms contacts. On the contrary, the hydrophilic residues are often located on the surface of globular proteins, and their ability to form contacts is poor.

We find that the average number of long-range contacts per residue to be in accord with the free energies of transfer of amino acids from water to non-polar environments reported by Fauchere and Pliska. [27] In Fig. 1(a), we plot the average number of long-range contacts per residue C_L vs. the Fachere–Pliska hydrophobicity scale (FPH), and find that the value of C_L increases with increasing FPH value. Here we only show the relationship for the case of $R_C = 6.5$ Å. In fact, a similar relationship can also be found for the other cases. The relationship between C_L and the Fachere–Pliska hydrophobicity scale (FPH) is expressed approximately as

$$C_L = a + b \times \mathrm{FPH} \tag{7}$$

Here $a = 8.02$ and $b = 1.99$ with $R_C = 6.5$ Å.

Except for Pro, the relative deviation between the Fachere–Pliska hydrophobicity scale and our C_L value is very small. We also plot the average number of short-range contact per residue C_S vs. the Fachere–Pliska hydrophobicity scale (FPH) in Fig. 1(b) and do not observe any relationship.

3.2. Folding rate predication from the distance of the atom pair contacts

The contact distance is also important for scaling the amino acid's folding character. The widely used parameters are CO and TCD. In this paper, we discuss these two parameters in detail for the 20 amino acids. Tables 6–8 list the values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ for

Table 3
Average number $N_{\lambda\text{-}\mu}$ of atomic pair contacts between different amino acid residues λ and μ. The upper triangle counts long-range contacts and the lower triangle counts short-range contacts in protein samples with $R_C = 8.0$ Å

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro
	47.0	37.8	40.5	19.3	51.5	34.2	11.7	35.1	21.7	13.1	20.6	15.6	16.1	16.3	14.4	13.0	17.4	15.1	21.2	13.2
Leu	13.4	36.1	34.4	16.5	38.3	28.8	12.3	26.9	20.0	12.1	18.9	14.0	14.0	14.9	12.2	11.9	14.6	15.0	17.2	12.7
Val	8.09	5.38	37.1	15.4	41.6	33.9	10.2	29.7	18.8	11.5	19.1	12.4	14.9	12.6	13.2	12.6	15.5	14.0	18.4	11.5
Ile	8.54	5.10	5.68	11.5	24.4	16.1	6.18	19.4	9.53	6.06	8.26	8.24	8.39	6.72	6.86	6.85	7.59	6.58	10.8	6.80
Met	5.47	3.86	3.97	5.08	62.9	35.7	13.2	39.7	19.9	13.4	22.8	23.3	17.0	15.5	16.8	15.3	17.4	16.7	20.9	14.7
Phe	11.2	7.22	6.39	5.75	10.6	40.9	14.7	30.8	19.9	14.4	22.5	22.6	19.4	19.7	21.5	19.9	25.1	20.1	28.7	19.3
Tyr	8.06	5.89	5.60	5.43	6.15	7.54	29.1	11.8	6.76	5.57	8.02	6.40	5.29	6.00	6.59	7.85	5.98	7.91	7.84	7.02
Cys	2.72	2.51	2.07	1.29	2.43	2.02	4.82	25.6	16.1	12.4	17.6	19.6	14.3	15.6	14.8	16.5	15.1	13.3	23.1	14.5
Trp	6.64	5.72	4.62	1.44	5.40	5.17	2.72	8.35	11.0	7.88	12.0	9.47	8.87	8.76	8.55	8.72	9.57	10.1	10.6	8.70
Ala	7.04	5.35	4.91	3.34	5.32	4.90	2.17	4.45	7.01	7.87	10.9	6.87	7.48	7.66	8.92	8.52	7.86	8.94	10.1	7.95
Gly	4.28	3.20	2.97	1.96	3.40	3.56	1.39	3.13	3.09	2.58	18.0	11.5	11.9	12.1	13.7	12.5	12.9	13.8	14.3	11.1
Thr	4.82	3.69	4.24	2.76	4.67	5.33	1.73	3.88	4.14	3.06	4.79	14.7	11.8	8.50	12.4	10.2	9.41	10.3	12.5	9.41
His	7.30	4.14	4.20	2.73	7.23	4.12	3.46	2.03	3.56	2.79	3.64	8.48	10.4	9.24	9.76	10.2	17.9	10.0	20.9	8.31
Glu	9.09	5.34	6.23	4.65	6.35	5.39	1.84	4.16	7.03	2.82	5.05	5.25	8.66	8.84	9.07	9.30	10.2	10.0	13.84	8.97
Gln	7.63	4.15	3.97	4.35	4.04	5.10	1.78	5.43	5.55	3.51	3.71	2.72	6.68	5.50	10.5	12.2	15.2	10.5	20.7	8.57
Asp	5.82	4.26	5.18	3.33	5.61	6.36	2.16	4.12	5.96	3.75	4.61	4.27	5.91	4.94	5.76	12.3	11.4	11.1	14.1	8.48
Asn	4.81	3.77	4.36	3.03	5.00	5.76	2.30	5.15	4.60	2.95	4.20	3.61	6.19	4.96	5.44	4.10	10.9	10.4	10.2	7.96
Lys	7.76	6.10	5.90	3.90	6.32	5.54	2.35	3.46	6.96	3.77	5.00	4.69	13.6	6.60	10.6	7.22	8.36	12.0	12.6	9.12
Ser	5.19	3.77	3.62	2.40	4.12	5.71	2.14	3.55	4.34	2.59	4.26	2.61	5.43	4.96	4.68	4.53	4.49	3.78	18.1	10.6
Arg	8.41	5.11	5.22	3.71	5.35	6.56	2.33	5.10	5.68	3.53	6.06	4.38	11.2	8.40	8.01	5.23	4.95	4.63	10.5	9.75
Pro	3.84	2.75	2.83	1.70	3.92	4.23	1.01	1.75	3.17	1.68	2.72	2.61	3.08	3.21	3.54	2.08	4.11	2.88	3.02	2.08

Table 4
Average number $N_{\lambda\text{-}\mu}$ of atomic pair contacts between different amino acid residues λ and μ. The upper triangle counts long-range contacts and the lower triangle counts short-range contacts in protein samples with $R_C = 10.0$ Å

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro
	122	91.0	96.7	50.3	122	87.2	29.6	82.9	59.6	39.0	56.4	49.7	59.1	52.4	50.3	42.5	57.2	46.3	65.3	40.2
Leu	22.1	81.4	80.1	41.1	92.3	72.4	30.1	67.1	51.6	33.8	47.9	39.9	45.0	42.3	40.1	36.3	45.4	41.7	49.5	35.3
Val	14.2	9.99	84.7	40.0	96.9	78.4	25.9	69.3	49.7	33.3	50.7	38.6	46.9	36.8	44.5	39.6	47.8	39.4	49.9	34.2
Ile	14.9	9.41	10.5	30.9	57.0	41.9	15.4	40.0	28.4	18.0	25.2	24.6	29.6	22.5	23.8	22.6	26.1	20.9	32.2	18.8
Met	8.89	6.56	6.95	8.84	144	86.1	31.5	86.5	55.4	38.2	60.6	62.0	57.6	43.9	54.9	47.4	56.5	48.4	59.0	41.2
Phe	20.8	14.1	12.2	10.5	20.8	96.0	34.0	72.6	51.2	39.4	60.4	52.9	54.3	51.3	61.1	55.6	60.5	53.1	69.6	47.9
Tyr	15.7	12.3	11.6	10.2	13.1	15.6	94.4	29.8	19.4	15.0	20.8	19.8	16.0	17.8	18.8	21.0	18.8	20.7	22.0	17.3
Cys	4.92	4.59	3.68	2.26	4.62	4.14	9.56	71.1	45.2	33.0	47.2	44.7	42.0	45.3	44.2	46.8	42.0	38.9	56.5	33.6
Trp	13.6	10.6	9.77	2.30	10.5	16.8	4.55	17.4	37.2	24.0	34.1	29.1	33.7	30.1	31.9	29.3	36.2	31.1	35.9	25.5
Ala	11.6	8.97	8.37	5.81	10.8	9.70	3.75	9.63	11.4	21.4	29.0	21.4	23.3	23.1	26.6	24.1	25.2	24.2	29.1	21.0
Gly	8.00	6.57	6.14	3.38	7.34	7.35	2.85	6.82	5.57	5.15	45.9	30.8	36.0	34.5	38.6	34.6	37.8	36.4	41.8	29.1
Thr	9.55	7.48	8.26	5.18	7.53	12.2	3.49	8.07	7.67	6.15	9.55	45.4	35.0	25.8	34.1	29.2	32.4	27.2	36.5	26.4
His	13.8	8.35	8.44	5.09	13.8	8.02	6.29	3.90	6.52	5.68	7.09	16.5	42.3	34.0	36.3	34.4	58.6	31.5	61.0	26.5
Glu	16.5	10.3	11.7	8.28	13.0	12.1	3.68	8.25	12.5	5.45	8.92	10.0	15.8	31.5	30.5	29.5	34.4	31.3	43.8	25.0
Gln	13.7	6.40	7.86	7.41	8.29	9.87	3.23	11.6	9.83	6.23	7.11	5.41	10.9	9.95	36.9	35.3	48.9	32.7	53.7	28.4
Asp	11.2	8.25	11.2	6.31	12.5	14.1	4.04	9.03	10.3	7.04	8.59	7.26	11.5	8.39	10.4	33.7	36.4	31.8	40.5	23.4
Asn	9.63	7.77	9.17	6.01	11.0	12.7	4.69	10.8	8.44	5.98	8.04	6.26	11.1	8.42	10.0	7.76	44.6	31.8	36.8	27.9
Lys	14.7	11.6	11.6	7.49	13.0	11.5	4.96	7.45	12.3	7.05	10.0	8.23	23.8	12.1	18.2	12.2	16.6	33.8	36.0	26.8
Ser	9.94	7.29	7.36	4.33	9.02	12.2	4.09	7.71	7.69	5.01	8.70	5.16	9.38	8.95	12.9	8.21	9.04	7.95	60.3	30.2
Arg	15.7	9.62	10.1	7.35	11.3	12.6	4.49	10.2	9.96	3.98	11.2	7.39	19.5	15.2	14.5	9.98	9.89	8.88	19.1	26.7
Pro	8.09	6.11	6.85	3.38	8.93	8.52	2.52	5.54	6.71	3.89	6.01	5.23	7.01	6.22	8.55	4.95	8.98	6.88	6.98	6.13

Table 5
Average number of contacts per residue. C_S (C_L) is the average number of short-range (long-range) contacts per residue. P_λ is the coefficient measuring the different ability to form contacts for all the 20 amino acids, and $\bar{C}_S$ ($\bar{C}_L$) is the average of C_S (C_L)

20 Amino acids	Three types of amino acids	FPH scale values	P_α	$R_C = 8.0$ Å								$R_C = 6.5$ Å							
				C_S *	$\bar{C}_S$ *	C_L *	$\bar{C}_L$ *	C_S	$\bar{C}_S$	C_L	$\bar{C}_L$	C_S *	$\bar{C}_S$ *	C_L *	$\bar{C}_L$ *	C_S	$\bar{C}_S$	C_L	$\bar{C}_L$
Leu	Hydrophobic (H)	1.700	1.815	3.995	3.717	4.484	5.008	7.192	6.145	24.08	24.76	2.913	2.674	1.999	2.579	4.352	3.535	10.67	11.327
Val		1.220	1.581	3.510		5.533		5.919		25.25		2.433		2.877		3.447		11.43	
Ile		1.800	1.815	3.696		5.347		6.071		26.17		2.599		2.669		3.527		11.79	
Met		1.230	1.936	3.968		4.392		7.137		21.93		3.060		2.150		4.420		9.897	
Phe		1.790	2.517	3.718		4.726		6.006		26.36		2.653		2.460		3.278		12.11	
Tyr		0.960	2.747	3.667		4.637		5.319		23.56		2.562		2.441		2.818		11.14	
Cys		1.540	1.468	3.518		6.312		5.856		24.12		2.362		3.586		3.439		11.16	
Trp		2.250	3.166	3.870		4.632		5.659		26.59		2.806		2.446		3.003		12.41	
Ala	Neutral (N)	0.310	1.113	4.068	3.649	4.068	4.068	7.884	6.581	19.25	19.39	3.189	2.671	2.102	2.159	4.951	3.859	8.388	8.732
Gly		0.000	0.879	3.364		4.153		6.114		19.05		2.349		2.271		3.461		8.689	
Thr		0.260	1.577	3.311		4.126		5.525		19.48		2.299		2.310		3.118		8.883	
His		0.130	2.177	3.851		3.924		6.801		19.79		2.847		1.954		3.906		8.970	
Glu	Hydrophilic (P)	−0.640	2.041	3.931	3.721	2.640	3.232	6.662	6.147	12.83	15.25	3.031	2.721	1.290	1.610	4.027	3.559	5.507	6.767
Gln		−0.220	1.992	3.942		3.194		6.650		15.01		3.008		1.571		4.025		6.663	
Asp		−0.770	1.807	3.770		2.858		6.051		13.87		2.813		1.401		3.495		6.028	
Asn		−0.600	1.758	3.629		3.341		6.183		15.80		2.658		1.683		3.555		7.057	
Lys		−0.990	1.996	3.881		2.895		6.398		12.85		2.938		1.452		3.757		5.654	
Ser		−0.040	1.343	3.557		3.698		6.027		17.71		2.438		1.967		3.463		7.962	
Arg		−1.010	2.358	3.858		3.435		6.450		16.96		2.897		1.753		3.850		7.782	
Pro		0.720	1.581	3.200		3.794		4.754		17.00		1.982		1.761		2.300		7.483	

Here * refer to results from our previous work [16], and the FPH scale was obtained from Ref. [27].

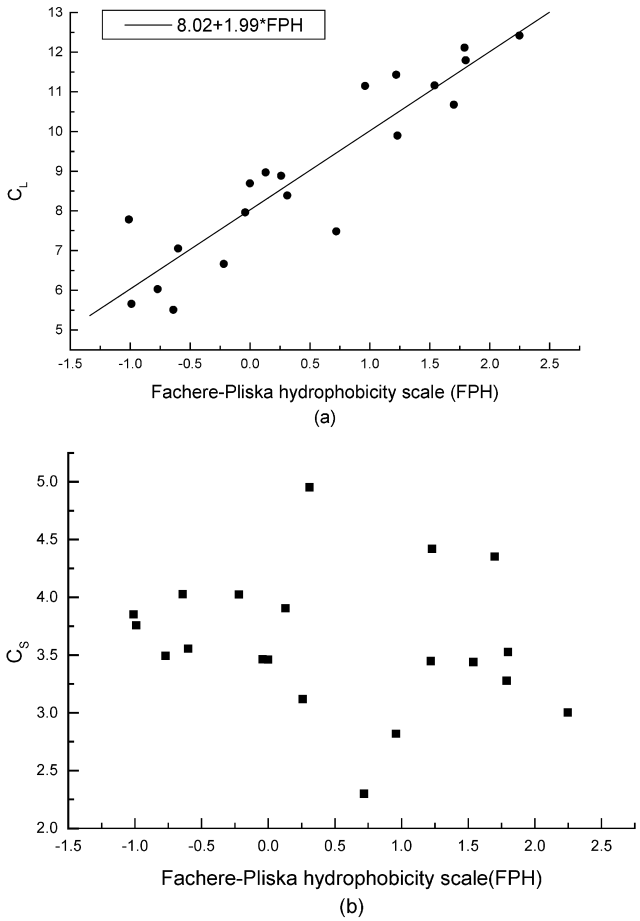

Fig. 1. The relationship between the average number of contacts per residue and the Fachere–Pliska hydrophobicity scale (FPH). (a) C_L and (b) C_S vs. the scale values of FPH. Here $R_C = 6.5$ Å.

pairs of residues according to Eqs. (5) and (6), with $R_C = 6.5$, 8.0 and 10.0 Å, respectively. In Table 6, there are 11 residue pairs with values of $CO_{\lambda\text{-}\mu}$ greater than 0.10 in the upper triangle, for residue pairs Trp-His, Trp-Met, Trp-Trp, Trp-Pro, His-Met, Trp-Cys, Trp-Phe, Cys-Cys, Trp-Asn, Cys-Met and Trp-Ile. The five lowermost values are the pairs Ala-Ala, Ala-Asp, Lys-Ala, Glu-Glu, and Ala-Glu. In the lower triangle of Table 6, there are 9 residue pairs with values of $TCD_{\lambda\text{-}\mu}$ higher than 0.80 and 8 residue pairs with values lower than or equal to 0.08. With increasing values of R_C, most values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ also increase. But the residue pairs with high values or low values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ are almost the same in Tables 6–8. In Table 7, there are 4 residue pairs with $CO_{\lambda\text{-}\mu}$ greater than 0.150, and 6 residue pairs with $CO_{\lambda\text{-}\mu}$ less than 0.025. In the lower triangle, there are 3 residue pairs with $TCD_{\lambda\text{-}\mu}$ greater than 2.00, and 12 residue pairs with $CO_{\lambda\text{-}\mu}$ less than 0.20. Here the more interesting result is that the high values of $CO_{\lambda\text{-}\mu}$ often occur for Trp although it does not have a high value of $N_{\lambda\text{-}\mu}$, while the low values often occur for Ala. We can conclude that if there is a Trp amino acid in the residue contact, the relative distance between two residues is far away. The reason may be that Trp amino acid residue itself consists of 14 atoms and has a large side chain. This means that Trp has a large value of $CO_{\lambda\text{-}\mu}$. Comparing with Tables 7 and 8, we find that the values of $CO_{\lambda\text{-}\mu}$ only increase for a few cases, while the values of $TCD_{\lambda\text{-}\mu}$ increase a lot. For example, the largest value of $TCD_{\lambda\text{-}\mu}$ increases from 2.98 to 7.46 with R_C increasing from 8.0 to 10.0 Å, nearly 2.50 times, while the largest value of $CO_{\lambda\text{-}\mu}$ increases from 0.176 to 0.210. This means that the summation over all contacts depends strongly on the limiting value of R_C, while the average sequence separation per contact depends only weakly on the limiting value of R_C. However, the relative values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ do not depend on the limiting value of R_C. The values $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$can aid us in knowing the structures of globular proteins and the effects of amino acid residues in protein folding.

In Ref. [5], the results of linear regressions of the experimental ln k_f values against CO or TCD are plotted. It shows that TCD is more accurate in folding rate prediction.

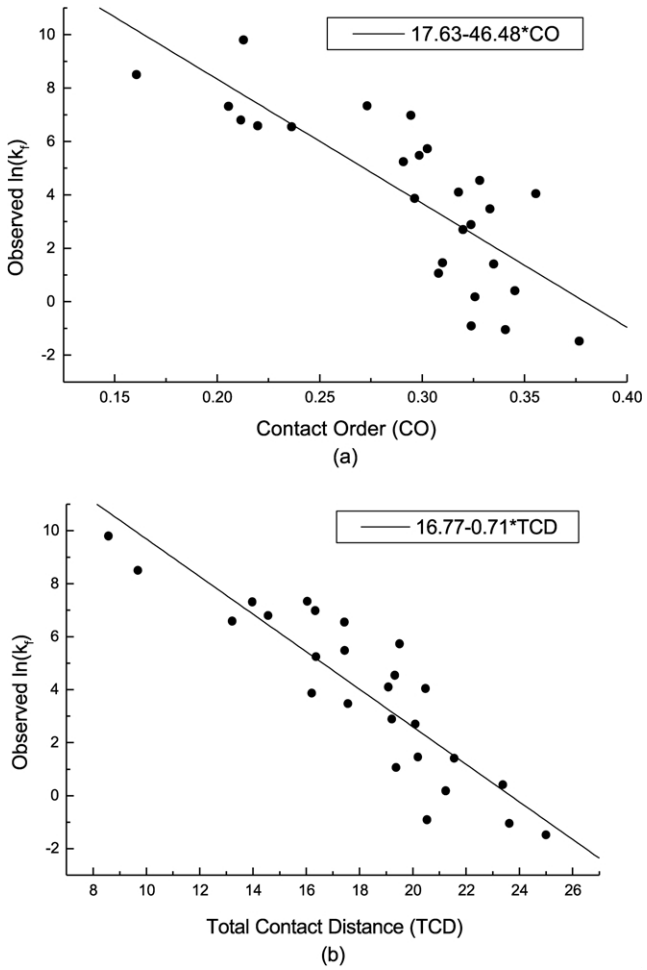

Fig. 2. The experimental (observed) values of ln k_f plotted as functions of parameters CO (a) and TCD (b).

Table 6
The values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ between different amino acid residues λ and μ. The upper triangle show $CO_{\lambda\text{-}\mu}$ and the lower triangle show $TCD_{\lambda\text{-}\mu}$ calculated for protein samples with $R_C = 6.5$ Å

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro
	0.029	0.030	0.036	0.048	0.039	0.042	0.054	0.068	0.026	0.024	0.033	0.049	0.023	0.035	0.029	0.032	0.025	0.029	0.032	0.046
Leu	0.425	0.034	0.039	0.068	0.046	0.048	0.058	0.076	0.028	0.029	0.037	0.056	0.027	0.036	0.033	0.037	0.028	0.033	0.039	0.052
Val	0.336	0.342	0.050	0.078	0.055	0.054	0.066	0.101	0.033	0.032	0.042	0.057	0.028	0.048	0.035	0.045	0.032	0.036	0.052	0.056
Ile	0.439	0.361	0.531	0.097	0.077	0.098	0.110	0.169	0.044	0.050	0.056	0.135	0.046	0.050	0.047	0.067	0.050	0.048	0.074	0.095
Met	0.301	0.346	0.374	0.423	0.065	0.075	0.087	0.116	0.040	0.042	0.049	0.071	0.040	0.051	0.044	0.047	0.039	0.045	0.064	0.067
Phe	0.600	0.509	0.618	0.608	1.21	0.078	0.069	0.098	0.041	0.040	0.044	0.095	0.047	0.059	0.050	0.052	0.049	0.041	0.067	0.063
Tyr	0.428	0.389	0.514	0.553	0.750	0.932	0.116	0.132	0.048	0.063	0.062	0.082	0.051	0.059	0.051	0.061	0.048	0.060	0.095	0.084
Cys	0.187	0.205	0.186	0.190	0.314	0.279	1.26	0.160	0.058	0.056	0.085	0.190	0.081	0.095	0.060	0.112	0.087	0.078	0.099	0.152
Trp	0.696	0.590	0.820	0.921	1.31	0.843	0.462	1.26	0.018	0.023	0.029	0.041	0.020	0.028	0.018	0.025	0.019	0.025	0.030	0.037
Ala	0.185	0.174	0.199	0.140	0.233	0.248	0.101	0.275	0.086	0.028	0.032	0.052	0.026	0.030	0.028	0.034	0.023	0.031	0.038	0.046
Gly	0.101	0.105	0.106	0.091	0.160	0.176	0.103	0.212	0.061	0.071	0.034	0.067	0.031	0.046	0.037	0.035	0.026	0.033	0.037	0.063
Thr	0.199	0.196	0.224	0.145	0.307	0.283	0.137	0.396	0.115	0.111	0.197	0.088	0.054	0.065	0.057	0.066	0.047	0.064	0.082	0.081
His	0.269	0.234	0.213	0.337	0.522	0.646	0.215	1.04	0.122	0.113	0.263	0.509	0.019	0.032	0.025	0.034	0.024	0.025	0.035	0.054
Glu	0.136	0.122	0.142	0.149	0.205	0.266	0.082	0.333	0.077	0.061	0.126	0.228	0.088	0.045	0.031	0.041	0.031	0.038	0.040	0.067
Gln	0.200	0.163	0.184	0.144	0.241	0.368	0.106	0.460	0.100	0.082	0.176	0.165	0.131	0.155	0.033	0.035	0.027	0.029	0.046	0.039
Asp	0.126	0.116	0.137	0.106	0.210	0.318	0.102	0.244	0.064	0.087	0.169	0.237	0.087	0.105	0.125	0.044	0.028	0.035	0.057	0.060
Asn	0.125	0.130	0.169	0.153	0.205	0.305	0.148	0.578	0.080	0.095	0.147	0.221	0.142	0.142	0.164	0.183	0.025	0.028	0.036	0.039
Lys	0.146	0.136	0.155	0.132	0.210	0.382	0.087	0.390	0.072	0.064	0.115	0.152	0.208	0.132	0.186	0.136	0.109	0.032	0.044	0.046
Ser	0.136	0.147	0.150	0.100	0.206	0.252	0.138	0.291	0.090	0.089	0.150	0.209	0.098	0.135	0.112	0.136	0.097	0.124	0.046	0.082
Arg	0.223	0.200	0.301	0.249	0.399	0.607	0.208	0.700	0.116	0.130	0.188	0.345	0.307	0.240	0.357	0.271	0.120	0.185	0.333	0.080
Pro	0.165	0.175	0.173	0.182	0.286	0.370	0.151	0.599	0.099	0.099	0.202	0.221	0.132	0.189	0.103	0.139	0.106	0.118	0.246	0.201

Table 7
The values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ between different amino acid residues λ and μ. The upper triangle show $CO_{\lambda\text{-}\mu}$ and the lower triangle show $TCD_{\lambda\text{-}\mu}$ calculated for protein samples with $R_C = 8.0$ Å

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro
	0.032	0.033	0.039	0.057	0.042	0.045	0.058	0.073	0.028	0.027	0.035	0.053	0.026	0.039	0.032	0.035	0.028	0.032	0.036	0.046
Leu	0.956	0.035	0.041	0.072	0.048	0.048	0.066	0.080	0.030	0.030	0.038	0.061	0.029	0.040	0.035	0.039	0.030	0.035	0.043	0.053
Val	0.747	0.730	0.051	0.083	0.057	0.056	0.071	0.105	0.036	0.035	0.046	0.064	0.032	0.051	0.037	0.048	0.036	0.038	0.054	0.059
Ile	0.946	0.819	1.096	0.113	0.085	0.098	0.135	0.166	0.050	0.054	0.065	0.141	0.054	0.059	0.056	0.072	0.052	0.058	0.078	0.103
Met	0.706	0.733	0.808	0.933	0.066	0.074	0.090	0.118	0.040	0.042	0.052	0.077	0.042	0.055	0.046	0.050	0.040	0.047	0.066	0.065
Phe	1.308	1.098	1.379	1.271	2.417	0.079	0.073	0.098	0.043	0.042	0.048	0.097	0.049	0.062	0.051	0.053	0.050	0.042	0.069	0.068
Tyr	0.949	0.836	1.106	1.053	1.537	1.911	0.112	0.150	0.051	0.062	0.064	0.086	0.059	0.070	0.056	0.064	0.051	0.061	0.098	0.083
Cys	0.417	0.489	0.434	0.502	0.687	0.607	1.908	0.176	0.063	0.060	0.087	0.186	0.089	0.094	0.067	0.107	0.084	0.082	0.099	0.153
Trp	1.537	1.287	1.804	1.721	2.663	1.755	1.101	2.984	0.021	0.025	0.031	0.046	0.023	0.032	0.021	0.028	0.022	0.028	0.034	0.039
Ala	0.402	0.376	0.423	0.322	0.508	0.526	0.227	0.651	0.190	0.029	0.034	0.057	0.029	0.035	0.030	0.035	0.024	0.032	0.041	0.045
Gly	0.232	0.232	0.250	0.217	0.349	0.379	0.218	0.465	0.136	0.150	0.036	0.070	0.035	0.049	0.038	0.038	0.028	0.034	0.041	0.061
Thr	0.454	0.430	0.531	0.359	0.715	0.664	0.311	0.914	0.250	0.235	0.406	0.097	0.058	0.072	0.063	0.067	0.049	0.066	0.086	0.091
His	0.596	0.553	0.526	0.772	1.174	1.303	0.431	2.010	0.300	0.275	0.535	1.123	0.023	0.036	0.029	0.037	0.026	0.030	0.040	0.053
Glu	0.333	0.280	0.340	0.356	0.496	0.605	0.209	0.810	0.179	0.146	0.288	0.493	0.218	0.050	0.038	0.046	0.036	0.040	0.045	0.067
Gln	0.459	0.376	0.425	0.323	0.537	0.759	0.271	0.983	0.232	0.195	0.386	0.405	0.286	0.356	0.038	0.038	0.030	0.031	0.047	0.044
Asp	0.319	0.284	0.339	0.280	0.510	0.702	0.243	0.636	0.155	0.186	0.346	0.526	0.226	0.269	0.305	0.046	0.031	0.038	0.057	0.061
Asn	0.311	0.306	0.409	0.358	0.511	0.677	0.321	1.163	0.186	0.202	0.316	0.462	0.302	0.319	0.338	0.377	0.027	0.030	0.040	0.041
Lys	0.348	0.311	0.376	0.294	0.480	0.761	0.212	0.777	0.171	0.141	0.249	0.346	0.413	0.303	0.379	0.289	0.259	0.033	0.045	0.048
Ser	0.328	0.331	0.341	0.261	0.488	0.546	0.299	0.686	0.199	0.184	0.305	0.432	0.229	0.285	0.239	0.298	0.218	0.257	0.050	0.076
Arg	0.527	0.476	0.634	0.566	0.869	1.214	0.480	1.401	0.276	0.275	0.422	0.723	0.634	0.503	0.666	0.544	0.303	0.390	0.718	0.077
Pro	0.389	0.409	0.427	0.436	0.606	0.798	0.328	1.252	0.232	0.218	0.424	0.544	0.302	0.401	0.264	0.321	0.247	0.291	0.515	0.456

Table 8
The values of $CO_{\lambda\text{-}\mu}$ and $TCD_{\lambda\text{-}\mu}$ between different amino acid residues λ and μ. The upper triangle show $CO_{\lambda\text{-}\mu}$ and the lower triangle show $TCD_{\lambda\text{-}\mu}$ calculated for protein samples with $R_C = 10.0$ Å

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro
	0.035	0.036	0.042	0.067	0.046	0.049	0.063	0.082	0.030	0.030	0.038	0.059	0.032	0.043	0.034	0.039	0.032	0.035	0.041	0.048
Leu	2.15	0.038	0.046	0.080	0.051	0.051	0.070	0.084	0.032	0.033	0.040	0.067	0.033	0.045	0.037	0.042	0.034	0.038	0.046	0.056
Val	1.62	1.54	0.054	0.091	0.060	0.059	0.077	0.111	0.038	0.039	0.049	0.074	0.039	0.056	0.041	0.052	0.040	0.042	0.059	0.064
Ile	2.03	1.83	2.29	0.136	0.096	0.105	0.151	0.178	0.059	0.062	0.077	0.140	0.065	0.075	0.066	0.082	0.060	0.066	0.086	0.111
Met	1.68	1.63	1.82	2.11	0.070	0.075	0.094	0.125	0.043	0.044	0.055	0.087	0.046	0.063	0.049	0.056	0.044	0.050	0.068	0.066
Phe	2.84	2.36	2.94	2.74	5.01	0.082	0.082	0.107	0.046	0.046	0.051	0.098	0.052	0.068	0.053	0.057	0.052	0.047	0.071	0.071
Tyr	2.15	1.85	2.32	2.20	3.25	3.93	0.115	0.164	0.053	0.063	0.068	0.099	0.064	0.084	0.059	0.066	0.057	0.063	0.094	0.084
Cys	0.942	1.06	1.00	1.16	1.47	1.40	2.83	0.210	0.072	0.067	0.089	0.187	0.092	0.103	0.079	0.109	0.080	0.088	0.102	0.159
Trp	3.41	2.80	3.86	3.54	5.39	3.90	2.45	7.46	0.025	0.028	0.034	0.054	0.027	0.038	0.026	0.032	0.026	0.030	0.038	0.044
Ala	0.908	0.835	0.953	0.846	1.18	1.17	0.513	1.63	0.472	0.030	0.036	0.064	0.032	0.041	0.032	0.038	0.027	0.034	0.043	0.047
Gly	0.591	0.560	0.642	0.558	0.847	0.911	0.475	1.11	0.332	0.326	0.039	0.077	0.038	0.052	0.041	0.043	0.031	0.037	0.047	0.062
Thr	1.08	0.955	1.24	0.965	1.68	1.55	0.709	2.09	0.578	0.523	0.905	0.113	0.062	0.081	0.067	0.071	0.055	0.069	0.094	0.104
His	1.46	1.34	1.43	1.74	2.71	2.60	0.992	4.18	0.792	0.684	1.19	2.58	0.028	0.042	0.034	0.041	0.030	0.035	0.045	0.055
Glu	0.933	0.737	0.906	0.926	1.31	1.42	0.512	1.93	0.454	0.368	0.682	1.09	0.602	0.057	0.045	0.053	0.043	0.043	0.052	0.070
Gln	1.12	0.944	1.03	0.850	1.39	1.73	0.746	2.33	0.566	0.471	0.895	1.05	0.715	0.899	0.039	0.043	0.033	0.034	0.047	0.050
Asp	0.853	0.741	0.904	0.779	1.36	1.60	0.560	1.75	0.416	0.430	0.800	1.15	0.622	0.689	0.712	0.050	0.036	0.041	0.057	0.064
Asn	0.835	0.757	1.02	0.931	1.34	1.59	0.687	2.54	0.472	0.450	0.745	1.03	0.698	0.771	0.763	0.846	0.030	0.033	0.044	0.047
Lys	0.914	0.762	0.945	0.768	1.25	1.56	0.531	1.73	0.463	0.345	0.589	0.896	0.886	0.731	0.809	0.658	0.678	0.035	0.048	0.053
Ser	0.820	0.787	0.829	0.687	1.23	1.25	0.647	1.72	0.474	0.414	0.676	0.948	0.550	0.658	0.552	0.651	0.519	0.588	0.057	0.073
Arg	1.34	1.14	1.46	1.39	2.01	2.47	1.02	2.88	0.684	0.629	0.992	1.70	1.37	1.12	1.26	1.15	0.805	0.853	1.72	0.080
Pro	0.960	0.981	1.09	1.04	1.36	1.73	0.718	2.51	0.555	0.500	0.902	1.38	0.729	0.870	0.704	0.742	0.650	0.709	1.11	1.07

Table 9
The standard error of the linear regression. The linear function is $Y = A + B \times X$

	CO		TCD	
Standard error	Previous work [5]	Present work	Previous work [5]	Present work
A	2.725	2.122	1.579	1.381
B	8.985	7.102	1.317	0.075

We also do the same work based on the atom pair contacts, and the results are shown in Fig. 2. In order to compare, $R_C = 6.5$ Å and the protein samples were the same as in a previous work [5]. The standard error of the linear regression is also given in Table 9. Our method using counts of the contacts provides the better results as seen in the correlation with the logarithms of folding rates. The knowledge of long-range and short-range atomic pair contacts can help improve protein structure and property predictions.

Acknowledgements

This research was financially supported by NSFC (Nos. 29874012, 20174036, 20274040), Natural Science Foundation of Zhejiang Province (No. 101002) and the Special Funds for Major State Basic Research Projects (G1999064800). We also thank the referees for their critical reading of the manuscript and their good ideas.

References

[1] Chan HS, Dill KA. Macromolecules 1989;22:4559–73.
[2] Anfinsen CB, Haber E, Sela M, White FH. Proc Natl Acad Sci USA 1961;47:1309–14.
[3] Anfinsen CB. Science 1973;181:223–30.
[4] Plaxco KW, Simons KT, Baker D. J Mol Biol 1998;277:985–94.
[5] Zhou HY, Zhou YQ. Biophys J 2002;82:458–63.
[6] Debe DA, Goddard WA. J Mol Biol 1999;294:619–25.
[7] Dinner AR, Karplus M. Nat Struct Biol 2001;8:21–2.
[8] Zhang LX, Li J, Jiang ZT, Zhao DL. Polymer 2003;44:1751–6.
[9] Miller EJ, Fischer KF, Marqusee S. Proc Natl Acad Sci USA 2002;99:10359–63.
[10] Vladimir S, Anatoli S, Jaime P, Enrique EA, Marvin E. Bioinformatics 1999;15:327–32.
[11] McDonald IK, Thornton JM. J Mol Biol 1994;238:777–93.
[12] Miyazawa S, Jernigan RL. Protein Engng 1993;6:267–78.
[13] Russell RB, Barton GJ. J Mol Biol 1994;244:332–50.
[14] Levitt M. J Mol Biol 1976;104:59–107.
[15] Sippl MJ. Biophys Mol Biol Suppl 1996;65:3–10.
[16] Jiang ZT, Zhang LX, Chen J, Xia AG, Zhao DL. Polymer 2002;43:6037–47.
[17] Hubbard TJP, Ailey B, Brenner SE, Murzin AG, Chothia C. Nucl Acids Res 1999;272:254–6.
[18] Miyazawa S, Jernigan RL. Proteins 1999;34:49–68.
[19] Miyazawa S, Jernigan RL. Proteins 2003;38:35–43.
[20] Kumarevel TS, Gromiha MM, Ponnuswamy MN. Biophys Chem 1998;75:105–13.
[21] Kumarevel TS, Gromiha MM, Selvaraj S, Gayatri K, Kumar PKR. Biophys Chem 2002;99:189–98.
[22] Gromiha MM. Biophys Chem 2001;91:71–7.
[23] Gromiha MM, Selvaraj S. Int J Biol Macromol 2001;29:25–34.
[24] Miyazawa S, Jernigan RL. J Mol Biol 1996;256:623–44.
[25] Manavalan P, Ponnuswamy PK. Nature 1987;275:673–4.
[26] Chen J, Zhang LX, Li J, Wang YX, Jiang ZT, Zhao DL. Biophys Chem 2003;105:11–22.
[27] Fauchere JL, Pliska V. Eur J Med Chem 1983;18:369–75.