Residual occurrence and energy property of proteins in HNP model*

Jiang Zhou-Ting(姜舟婷)[a)†], Dou Wen-Hui(窦文辉)[a)], Shen Yu(沈 瑜)[b)],
Sun Ting-Ting(孙婷婷)[c)], and Xu Peng(徐 鹏)[a)]

[a)] *Department of Applied Physics, China Jiliang University, Hangzhou* 310018, *China*
[b)] *Department of Applied Physics, Zhejiang University of Science and Technology, Hangzhou* 310023, *China*
[c)] *College of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou* 310018, *China*

(Received 4 May 2015; revised manuscript received 3 July 2015; published online 10 October 2015)

Four categories of globular proteins, including all-α, all-β, $\alpha+\beta$, and α/β types, are simplified as the off-lattice HNP model involving the secondary-structural information of each protein. The propensity of three types of residues, i.e., H, N, and P to form a secondary structure is investigated based on 146 protein samples. We find that P residues are easy to form α-helices, whereas H residues have a higher tendency to construct β-sheets. The statistical analysis also indicates that the occurrence of P residues is invariably higher than that of H residues, which is independent of protein category. Changes in bond- and non-bonded potential energies of all protein samples under a wide temperature range are presented by coarse-grained molecular dynamics (MD) simulation. The simulation results clearly show a linear relationship between the bond-stretching/bending potential energy and the reduced temperature. The bond-torsional and non-bonded potential energies show distinct transitions with temperature. The bond-torsional energy increases to the maximum and then decreases with the increase of temperature, which is opposite to the change in non-bonded potential energy. The transition temperature of non-bonded potential energy is independent of the protein category, while that of bond-torsional energy is closely related to the protein secondary structure, i.e., α-helix or β-sheet. The quantitatively bonded- and semi-quantitatively non-bonded potential energy of 24 $\alpha+\beta$ and 23 α/β protein samples are successfully predicted according to the statistical results obtained from MD simulations.

Keywords: HNP model, molecular dynamics simulation, residue hydrophobicity

PACS: 68.35.bm **DOI:** 10.1088/1674-1056/24/11/116802

1. Introduction

One of the most crucial and still unclear problems in biophysics is how a protein folds into its native three-dimensional structure from a linear amino acid sequence. The biological function of a protein strongly depends on its unique structure determined by complicated interactions among residues. Different potentials among residues are given by varying scales of residue hydrophobicity, the presence of partial charges or hydrogen-bonding groups or hydropgen-bonding groups.[1] It is well established, the primary structural information, i.e., the sequence of amino acid is sufficient to determine the tertiary structure according to the global minimum on free-energy landscape of a protein.[2] However, in the presence of additional protein chains or in the crowded cellular environment, the misfolding and aggregation of non-native protein structures are observed.[3,4] Then, determining the three-dimensional structure of a protein molecule from its amino acid sequence remains difficult because the energy that stabilizes the folded conformation is poorly understood and the number of possible structures is extensive.[5,6]

Computational simulation is a powerful method to discover the detailed mechanism of structural information at an atomistic level. Nowadays, fruitful researches have been carried out in the field of structural formation of macromolecular systems by computer simulation to overcome experimental difficulties.[7–12] Several studies have been carried out in the crystallization of polymer chains and the processes of DNA condensation by molecular dynamics (MD) simulations.[13,14] Despite the rapid progress of computer power, numerous basic questions, such as the nature of the interactions involved in protein folding, the thermodynamic states involved as a protein folds, and the structural properties of each thermodynamics state remain unanswered.[15] At present, the statistical analysis of amino acid sequences is regarded as another important method to understand the structure of proteins, which attempts to provide a deep insight into fundamental biology and a new idea for combating protein misfolding diseases.[16–20]

Proteins, with numerous freedom degrees, show considerable structural diversity. Therefore, the simulation on realistic all-atom representations of a protein limits the running time because of computational expense. To resolve this problem, the simplified representations and energy functions of real proteins have been presented.[1,5,19–24] One of the widely used simplified lattice models is the HP model, which was first proposed by Dill.[21] In the HP model, two beads named H (hydrophobic residues) and P (polar residues) exist. The hy-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 21204078, 11304282, and 11202201) and the Natural Science Foundation of Zhejiang Province, China (Grant No. LY12B04003).

†Corresponding author. E-mail: z.jiang@cjlu.edu.cn

© 2015 Chinese Physical Society and IOP Publishing Ltd

drophobic energies are attractive, i.e., H–H contacts mainly contribute to the energetics in the protein folding process and all the other groups are treated as zero. Recently, protein engineering experiments suggest that instead of two, three kinds of residues can be effectively substituted for the twenty amino acids.[25] According to this pattern, Miyazawa–Jernigan (MJ) modified the interaction matrix.[26] Therefore, in the modified HNP off-lattice model, the two kinds of residues in the standard HP lattice model are replaced by three kinds of residues, namely H (hydrophobic residues), N (neutral residues), and P (polar residues).[19,20,24,25] The nonlocal interaction term is residue type dependent. The interaction between H–H residues has a deep primary minimum that gives rise to defining folded structures. The other residue pair interactions are always repulsive with different strength. The protein is composed of residues with C_α at the center. The backbone of the protein constituted by simplified residues in a continuous space instead of a cubic lattice, have been studied in the literatures.[24,27,28] While lacking the chemical detail of all-atom representations, these simplified models are amenable to characterize both kinetics and thermodynamics during the protein folding process.

Globular proteins are categorized into four structural classes, namely all-α, all-β, $\alpha+\beta$, and α/β proteins, based on the topology and content of its secondary structure, i.e., α-helices and β-sheets.[29] The present article reports the simulation results carried out by the MD method on a single protein molecular, which is simplified as a modified HNP model involving its secondary structure and performed in the continuous space. A total of 146 protein samples divided into four types are adopted. Besides the occurrence of residues, especially in the secondary regions, the properties of the potential energy of each protein under its equilibrium states are discussed in detail. These results help us to abstract "essential physical properties" and predict the energy expression according to its sequence of residues.

2. Samples and methods

2.1. Database

A total of 146 globular proteins classified by the SCOP database,[30] including 50 all-α, 49 all-β, 24 $\alpha+\beta$, and 23 α/β proteins, were selected as the sample source for our present work. The amino acid sequence and secondary structure of each protein are obtained from the Protein Date Bank (PDB). Three types of amino acids, namely hydrophobic residues (H) consisting of Phe, Met, Leu, Ile, Trp, Cys, Tyr, and Val, neutral residues (N) consisting of Gly, His, Thr, and Ala, and polar residues (P) consisting of Asn, Lys, Asp, Ser, Glu, Gln, Arg, and Pro, are adopted by the MJ matrix and other literatures.[19,20,24,25,31] Then, the protein is represented by a chain containing 3 types of beads. Each bead in a protein molecule is represented by the position of an α-carbon atom (C_α). The PDB codes, chain length, and the components (i.e., the number of residues categorized by hydrophobic scale and secondary structure) for all protein samples used in this present article are given in Table 1.

Table 1. PDB codes and structural information of globular proteins classified as all-α, all-β, $\alpha+\beta$, and α/β types.

	NO.	PDB code	Chain length	N_H	N_N	N_P	$N_{\alpha-helix}$	$N_{\beta-sheet}$	Others
all-α proteins	1	1ALA	316	106	77	133	220	0	96
	2	1BBL	37	10	12	15	23	0	14
	3	1BP2	123	44	18	61	67	10	46
	4	1CCR	111	30	30	51	53	0	58
	5	1EA8	140	46	26	68	127	0	13
	6	1ECA	136	45	41	50	116	0	20
	7	1ECD	136	45	41	50	116	0	20
	8	1ECO	136	45	41	50	116	0	20
	9	1F63	154	49	43	62	133	0	21
	10	**1FCS**	**154**	**50**	**43**	**61**	**121**	**0**	**33**
	11	1FHA	172	57	34	81	125	0	47
	12	1FIP	73	27	14	32	59	0	14
	13	1H96	167	55	37	75	135	0	32
	14	1HBG	147	44	56	47	120	0	27
	15	1IFA	158	66	21	71	91	0	67
	16	1LE4	139	45	25	69	126	0	13
	17	1LH1	153	53	41	59	126	0	27
	18	1LH2	153	53	41	59	126	0	27
	19	1LPE	144	46	29	69	128	0	16
	20	1MBA	146	45	43	58	112	0	34
	21	1MBC	153	48	45	60	121	0	32
	22	1MBD	153	48	45	60	121	0	32
	23	1MBS	153	46	44	63	121	0	32
	24	1PPA	121	44	23	54	66	4	51
	25	1RCB	129	41	31	57	80	9	40
	26	1RRO	108	30	21	57	57	4	47
	27	1UTG	70	25	12	33	53	0	17
	28	1YCC	108	30	31	47	52	7	49
	29	1YEA	112	28	32	52	54	0	58
	30	2C2C	112	31	33	48	59	0	53
	31	2CDV	107	24	37	46	19	6	82
	32	2CTS	437	158	104	175	315	7	115
	33	2CY3	118	27	41	50	26	4	88
	34	2CYP	293	93	62	138	155	16	122
	35	2END	137	49	30	58	61	0	76
	36	2FAL	146	45	43	58	112	0	34
	37	2HBG	147	44	56	47	120	0	27
	38	2LHB	149	50	39	60	113	0	36
	39	2MHR	118	42	24	52	76	0	42
	40	2PDE	43	14	12	17	22	0	21
	41	2WRP	104	35	20	49	78	0	26
	42	3C2C	112	31	33	48	59	0	53
	43	451C	82	23	23	36	42	0	40
	44	4BP2	64	25	11	28	39	0	25
	45	4CPV	108	30	34	44	70	0	38
	46	4ICB	76	24	9	43	56	2	18
	47	4MBN	153	48	45	60	121	0	32
	48	5CPV	108	30	34	44	60	6	42
	49	5CYT	103	30	29	44	52	0	51
	50	5TNC	161	50	32	79	106	10	45

Table 1. (*Continued*).

	NO.	PDB code	Chain length	N_H	N_N	N_P	$N_{\alpha-helix}$	$N_{\beta-sheet}$	Others
all-β proteins	51	1A45	173	63	30	80	7	62	104
	52	1ACX	108	26	45	37	0	49	59
	53	1CA2	256	81	58	117	41	59	156
	54	1CD8	114	42	22	50	5	53	56
	55	1CID	177	57	29	91	4	99	74
	56	1EST	240	86	67	87	19	119	102
	57	1F3G	150	55	33	62	29	74	47
	58	1F53	84	29	16	39	0	32	52
	59	1GPR	158	55	33	70	17	53	88
	60	1H6X	159	53	47	59	0	82	77
	61	1HJC	118	38	32	48	0	50	68
	62	1HOE	74	25	24	25	0	50	24
	63	1IFC	131	44	29	58	16	77	38
	64	1KL9	165	62	21	82	87	33	45
	65	1MPP	357	122	91	144	50	150	157
	66	1NN2	388	130	82	176	11	188	189
	67	1PAZ	120	42	28	50	21	54	45
	68	1PYP	280	87	62	131	52	105	123
	69	1QNY	237	74	59	104	18	108	111
	70	1RBP	174	61	34	79	16	81	77
	71	1SGT	223	68	79	76	28	71	124
	72	1STP	121	31	54	36	13	67	41
	73	1TEN	89	26	21	42	0	48	41
	74	1TIE	166	58	31	77	11	61	94
	75	1TLK	103	35	21	47	4	61	38
	76	1TON	226	82	51	93	29	71	126
	77	2ALP	198	57	75	66	19	83	96
	78	2APR	325	106	99	120	40	150	135
	79	2AYH	214	72	61	81	13	102	99
	80	2BB2	181	56	33	92	22	80	79
	81	2BUK	184	62	49	73	28	82	74
	82	2CA2	256	81	58	117	41	59	156
	83	2CAB	256	75	59	122	48	72	136
	84	2CNA	237	73	60	104	4	121	112
	85	2ER7	330	100	115	115	34	184	112
	86	2GCH	131	44	37	50	0	61	70
	87	**2MCM**	**112**	**31**	**50**	**31**	**0**	**93**	**19**
	88	2PCY	99	33	21	45	5	61	33
	89	2REN	320	120	85	115	32	134	154
	90	2RHE	114	32	30	52	8	51	55
	91	2SGA	181	50	75	56	14	108	59
	92	2SNS	141	40	36	65	34	32	75
	93	2SNV	151	45	46	60	5	77	69
	94	3APP	323	96	97	130	44	160	119
	95	3CNA	237	73	60	104	0	112	125
	96	3EST	240	86	67	87	19	120	101
	97	4FGF	124	43	24	57	17	71	36
	98	4PEP	326	118	78	130	42	156	128
	99	5PTP	222	77	52	93	23	70	129
	100	102L	163	53	41	69	124	14	25
	101	125L	162	53	39	70	107	14	41
	102	190L	162	54	41	67	108	17	37

Table 1. (*Continued*).

	NO.	PDB code	Chain length	N_H	N_N	N_P	$N_{\alpha-helix}$	$N_{\beta-sheet}$	Others
$\alpha+\beta$ proteins	103	1BKF	107	35	27	45	16	41	50
	104	1E3V	128	46	32	50	45	44	39
	105	1EZM	298	96	84	118	136	35	127
	106	1FDD	106	41	15	50	47	12	47
	107	1FKB	107	34	28	45	7	41	59
	108	1FKD	107	34	28	45	7	42	58
	109	1FKF	107	34	28	45	7	41	59
	110	1FRH	106	41	15	50	44	12	50
	111	1GWD	129	42	32	55	64	8	57
	112	1I1Z	130	45	31	54	63	8	59
	113	1IET	94	25	23	46	9	6	79
	114	1IEU	94	25	23	46	10	0	84
	115	1LHI	130	45	32	53	44	6	80
	116	1LZ1	130	45	31	54	44	6	80
	117	1POP	212	76	52	84	58	47	107
	118	1PPN	212	76	52	84	72	16	124
	119	2LYZ	129	42	32	55	35	20	74
	120	2LZM	164	56	38	70	107	18	39
	121	3SSI	108	36	37	35	18	36	54
	122	9PAP	212	76	52	84	58	42	112
	123	**9RNT**	**104**	**31**	**28**	**45**	**17**	**36**	**51**
α/β proteins	124	1ABA	87	32	17	38	36	16	35
	125	1AZL	147	48	42	57	58	48	41
	126	1C4W	128	47	31	50	65	21	42
	127	1DHR	236	76	72	88	100	65	71
	128	1E49	206	74	60	72	109	35	62
	129	1E6K	130	47	31	52	69	24	37
	130	1EAF	243	87	63	93	85	63	95
	131	1EX7	186	57	39	90	97	39	50
	132	1FX1	147	47	43	57	55	32	60
	133	1Q21	171	60	36	75	70	43	58
	134	1RHD	293	100	69	124	109	43	141
	135	1TFD	294	101	61	132	89	50	155
	136	2FCR	173	56	42	75	58	42	73
	137	2FOX	138	52	25	61	52	27	59
	138	2FX2	147	47	44	56	58	48	41
	139	3ADK	194	64	43	87	124	25	45
	140	**3CHY**	**128**	**46**	**31**	**51**	**50**	**24**	**54**
	141	3CLA	213	87	33	93	66	61	86
	142	3DFR	162	54	46	62	40	60	62
	143	3PGM	230	74	43	113	78	23	129
	144	4CLA	213	87	33	93	65	50	98
	145	5P21	166	58	36	72	61	44	61
	146	8DFR	186	67	31	88	49	54	83

2.2. Simulation model and method

To minimize the running time during simulation processes on numerous protein samples and show the general properties of globular proteins instead of a specific one, the modified off-lattice minimalist HNP model, in which a single

protein chain is described as a sequence of three types of beads H, N, and P, is selected in this article. The beads interact via the non-bonded potentials, and bonded potentials including bond-stretching, bond-bending and bond-torsional potentials. The coarse-grained force field used here is the same as that in the literature.[24,32] The potential energy functions adopted in our work can be thought of as a simple version used in the study of the structure and dynamics of real proteins. The four potential functions of a protein are as follows.

(I) The non-bonded interaction involves beads that are at least four residues apart in the amino acid sequence. For two beads, the non-bonded interaction is given by the truncated 12–6 Lennard–Jones potential

$$E_{\mathrm{LJ}} = \sum_{i=1}^{N}\sum_{j=1}^{N} 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \Lambda\left(\frac{\sigma}{r_{ij}}\right)^{6}\right], \quad (j-i\geq 4), \quad (1)$$

where r_{ij} is the distance between beads i and j, ε is the parameter that defines the energy scale, ε sets the strength of the nonlocal interaction and Λ sets the nonlocal interaction that is attractive or repulsive. Both of these two parameters depend upon the nature of the two beads involved (see Table 2); σ is the length parameter representing the equilibrium bond length in Eq. (2).

Table 2. Parameters ε and Λ in the non-bonded interaction term of the potential function depending on the nature of the two residues involved.

	H	N	P
H	$\varepsilon=\varepsilon_{\mathrm{h}};\Lambda=1$	$\varepsilon=2/3\varepsilon_{\mathrm{h}};\Lambda=0$	$\varepsilon=7/12\varepsilon_{\mathrm{h}};\Lambda=0$
N	$\varepsilon=2/3\varepsilon_{\mathrm{h}};\Lambda=0$	$\varepsilon=1/3\varepsilon_{\mathrm{h}};\Lambda=0$	$\varepsilon=1/4\varepsilon_{\mathrm{h}};\Lambda=0$
P	$\varepsilon=7/12\varepsilon_{\mathrm{h}};\Lambda=0$	$\varepsilon=1/4\varepsilon_{\mathrm{h}};\Lambda=0$	$\varepsilon=1/6\varepsilon_{\mathrm{h}};\Lambda=-1$

(II) The bond-stretching potential energy models the covalent bond connecting adjacent residues as a spring with an equilibrium value of σ rather than as a rigid rod. It takes the form

$$E_{\mathrm{stretch}} = \sum_{i=2}^{N}\frac{1}{2}k_{\mathrm{b}}(r_i-\sigma)^2, \quad (2)$$

where r_i is the bond length between beads $i-1$ and i. The bond-stretching potential energy constant is expressed as $k_{\mathrm{b}} = 100\varepsilon_{\mathrm{h}}/\sigma^2$.

(III) The bond-bending potential energy models the harmonic potential to constrain the bond angle at equilibrium angle θ_0 between three successive beads

$$E_{\mathrm{bend}} = \sum_{i=3}^{N}\frac{1}{2}k_{\theta}(\theta_i-\theta_0)^2, \quad (3)$$

where the equilibrium bond angle $\theta_0 = 1.8326$ rad and θ_i is the bond angle between three beads $i-2$, $i-1$, and i. The bond-bending potential energy constant is $k_\theta = 13.33\varepsilon_{\mathrm{h}}\cdot\mathrm{rad}^{-2}$.

(IV) The bond-torsional potential energy describes the rotation of the three bonds that connect four successive residues in the amino acid sequence

$$E_{\mathrm{torsion}} = \sum_{i=4}^{N}\left[-A\cos^6\left(\frac{\phi_i-\phi_\alpha}{2}\right) - B\cos^6\left(\frac{\phi_i-\phi_\beta}{2}\right)\right], \quad (4)$$

where $\phi_\alpha = 1.0$ rad, and $\phi_\beta = \pi$ rad, ϕ_i is the dihedral angle of two planes consisting of three successive bonds. The parameters A and B vary with different secondary regions of the protein. For α-helices: $A = 6.0\varepsilon_{\mathrm{h}}$, $B = 5.6\varepsilon_{\mathrm{h}}$. For β-sheets: $A = 5.6\varepsilon_{\mathrm{h}}$, $B = 6.0\varepsilon_{\mathrm{h}}$. For other cases: $A = B = 0$.

The same as the previous work, the simulation results are presented in terms of reduced units.[33] Here, the reduced temperature $T^* = K_{\mathrm{B}}T/\varepsilon$, the reduced energy $E^* = E/\varepsilon$, and the reduced time $t^* = t\sqrt{\varepsilon/m\sigma^2}$. A single protein chain exposed to a vacuum is simulated by the MD method. MD simulations are executed from the random configuration of the protein. The Nose–Hoover method is applied to keep the constant temperature of the system.[34,35]

3. Results and discussion

3.1. Residual occurrence in the secondary regions

Firstly, the ability of the three types of residues to form secondary structures, such as α-helix or β-sheet, is investigated based on 146 protein samples. The statistic results on the relationship between the residue types H, N, and P, and secondary structures such as α-helix, β-sheet or others, are listed in a 3 × 3 matrix (see Table 3(a)). A total of 23992 residues are found in 146 protein samples. The number of H, N, and P residues is 7915, 6026, and 10051, respectively. Then, the frequency of occurrence is easily obtained, i.e., H (33%), N (25%), and P (42%). In the HNP model, 20 amino acids are replaced by three kinds of residues. Hydrophobic residues (H) and hydrophilic residues (P) both consist of 8 types of amino acids, whereas neutral residues (N) have the half number of residue types compared with H or P. Based on the theory of statistics, the frequency of the occurrence of H and P should be equal, which is double that of N residues. It shows that the occurrence of N residues, which consists of only 4 kinds of amino acids, is common among the proteins. In 146 protein samples, 8586 and 5780 residues participate in forming α-helices or β-sheets, respectively. Both numbers are smaller than 9629 that the number of residues excluded from the secondary structure is neither α-helix nor β-sheet. More detailed data are shown in Table 3(a) and Fig. 1. For example, α-helical structures consist of 8586 residues, including 2964 H, 1955 N, and 3667 P residues, respectively. Among 5780 residues, which as the fundamental elements of β-sheets, H type presented a maximum value of 2665. These results conclude that P residues are found more frequently in an α-helix, meanwhile, H ones present the higher tendency to construct β-sheet secondary-structural elements. Previously,

several groups have derived a helix propensity scale based on studies of the protein stability by experimental or simulation methods.[36,37] In contrast, the propensity for β-sheet formation of 20 naturally occurring amino acids is also measured in the literature.[38] According to the helix propensities of amino acids achieved by Pace and Scholtz, the average value of helix propensities of H residues is 0.464 kcal/mol, which is higher than that of P type residues, i.e., 0.443 kcal/mol.[36] Thus, P residues are favorable to the occurrence of α-helix secondary structures.

Table 3. Residual occurrence in the secondary regions based on all protein samples (a), and four types of globular proteins (b), respectively.

(a)			
all protein samples	H	N	P
α-helix	2964	1955	3667
β-sheet	2665	1278	1837
others	2286	2793	4547
(b)			
all-α proteins	H	N	P
α-helix	1639	1099	1988
β-sheet	40	12	33
others	525	637	907
all-β proteins	H	N	P
α-helix	311	199	460
β-sheet	1836	1021	1339
others	955	1246	2126
$\alpha+\beta$ proteins	H	N	P
α-helix	442	281	524
β-sheet	307	74	181
others	392	486	714
α/β proteins	H	N	P
α-helix	572	376	695
β-sheet	482	171	284
others	414	424	800

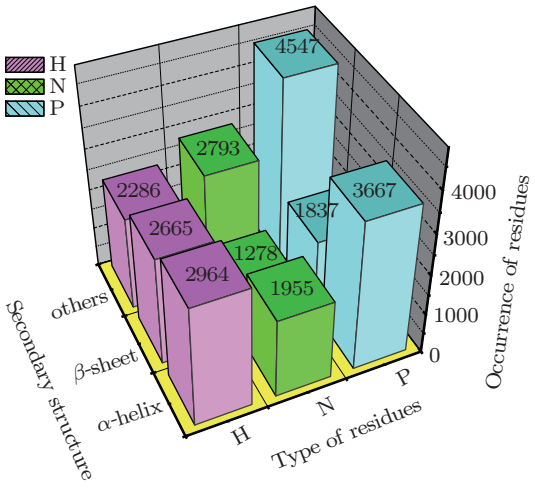


Fig. 1. (color online) Histogram of Residual occurrence. Twenty residues are simplified as three types of beads, i.e., H, N, and P according to the MJ matrix in Ref. [26]. The occurrences of H, N, and P are obtained from the statistical results in 146 protein samples listed in Table 1. The sequence and secondary structures of each protein can be downloaded from http://www.pdb.org.

The data were also analyzed according to four structural classes of globular proteins, i.e., all-α, all-β, $\alpha+\beta$, and α/β proteins, and the results are shown in Table 3(b). For 50 all-α proteins, α-helices dominate at its secondary-structural level with 4726 residues, which is 55.6 times larger than ones that construct β-sheets. For all-β proteins, 4196 residues participate in the β-sheet folds. Although nearly 4.33 times larger than the number of residues forming an α-helix, it is even less than the number of residues, which construct other secondary structures except α-helices and β-sheets. In four such protein categories, except for all-β proteins, the residues forming α-helices always exceed the ones forming β-sheets. In $\alpha+\beta$ and α/β proteins, the difference of residue number between constructing α-helices and β-sheets is within 2.2 times, which is smaller than the condition of all-α or all-β proteins. The all-α and all-β proteins consist almost exclusively of helical and β-sheet secondary structures, respectively, whereas the $\alpha+\beta$ and α/β classes contain varying degrees of both secondary-structural elements. Helics and sheets are spatially segregated in $\alpha+\beta$ proteins. In the case of α/β proteins, α-helices and β-sheets alternate typically, allowing the β-sheets to organize in a parallel fashion. In these four matrices arranged by four protein categories, the number of P residues is always larger than that of H residues, although both of them consist of eight different amino acids. The number of N type residues is the minimum, as it just includes four different amino acids. However, the proportion of occurrence remains higher than 4/20. These results verify that the residual composition of proteins is independent of protein category. It totally conforms to the discussion in a previous paragraph.

3.2. Potential energy property of four protein categories

The temperature dependence of the potential energy of each protein sample is investigated. Figures 2(a)–2(d) show the simulation results of PDB code 1FCS, 2MCM, 9RNT, and 3CHY proteins classified as all-α, all-β, $\alpha+\beta$, and α/β, respectively. The chain length, residual composition, and secondary-structural information are listed in Table 1. These figures show that: (i) the values of bond-stretching/bending potential energies are always larger than 0 under any simulation temperatures. Both energies are linearly increasing with the temperature in the four protein samples belonging to different protein categories. The values of bond-stretching and bond-bending potential energy approximate each other, especially when $T^* = 0.3$. When $T^* > 0.3$, the bond-stretching energy is slightly higher than the value of bond-bending energy, regardless of protein categories. On the contrary, when $T^* < 0.3$ the bond-stretching energy is slightly lower than the one of bond-bending energy. It indicates that the slope of bond-stretching potential energy as a function of the reduced temperature is larger than that of bond-bending energy. (ii)

Bond-torsional potential energy is the lowest one among the four components of potential energy in any simulation case. According to Eq. (4), bond-torsional potential energy is invariably lower than 0. Bond-torsional energies increase initially and then decrease with increasing temperature. The difference of bond-torsional potential energies shown in the four figures is the transition temperature shifting between $T^* = 0.2$ and $T^* = 0.3$. (iii) The trend of non-bonded interaction with the change of temperature is opposite to that of bond-torsional energy, i.e., non-bonded interaction evidently decreases and then increases to a stable value when the reduced temperature remains escalating. In the four figures, the non-bonded potential energies are always lower than 0. According to Eq. (1) and parameter values in Table 2, we can easily conclude that only the interaction between H–H residues present a prospectively negative value. In the other cases the residue pair interactions are always repulsive with different strengths. Therefore, H–H interactions perform the main functions in achieving the equilibrium state of each protein.

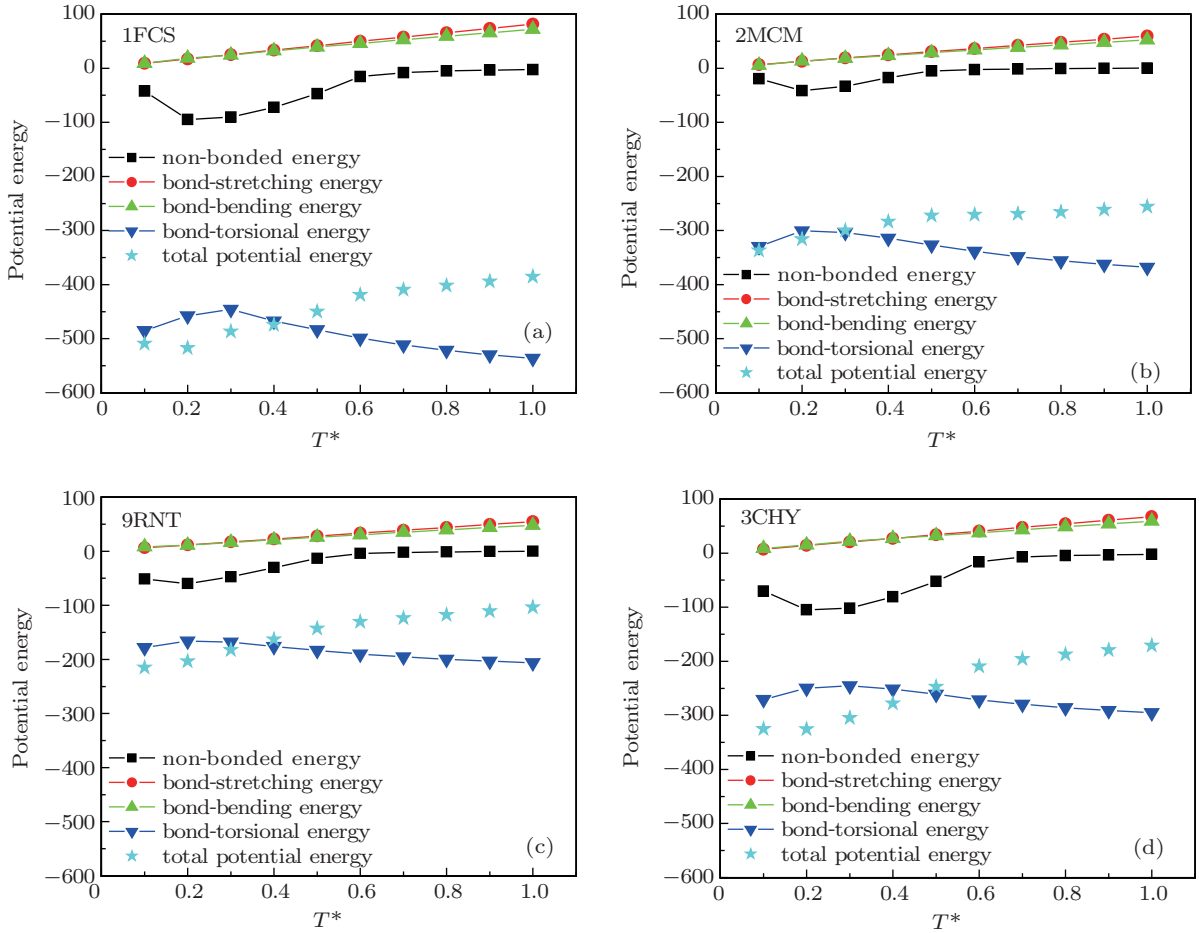


Fig. 2. (color online) The total and components of reduced potential energy versus reduced temperature T^* of four types of globular proteins. (a) 1FCS classified as all-α, (b) 2MCM classified as all-β, (c) 9RNT classified as $\alpha+\beta$, and (d) 3CHY classified as α/β protein, respectively.

From the perspective of protein categories, these figures also show that: (i) four categories of protein indicate the same tendencies of bond- and non-bonded potential energy with variable temperature. (ii) The transition temperature of a non-bonded interaction is the same at $T^* = 0.2$. (iii) The difference among all-α, all-β, $\alpha+\beta$, and α/β proteins is the transition temperature of bond-torsional potential energy. For all-α protein 1FCS and α/β protein 3CHY, the transition temperatures of bond-torsional energy are the same at $T^* = 0.3$. In addition, in the case of all-β protein 2MCM and $\alpha+\beta$ protein 9RNT, the transition temperatures of bond-torsional energy are both at $T^* = 0.2$. (iv) The change in total potential energy as a function of temperature is related to four parts of bond- and non-bonded potentials. Taking 1FCS as an example, figure 2(a) shows that the non-bonded potential energy of 1FCS is the main contributor during energy transition when $0.1 < T^* < 0.2$, because the total potential energy presents the same tendency as that of the non-bonded energy. In the temperature scope $0.2 < T^* < 0.6$, the total potential energy increases evidently with increasing temperature. The increasing

tendencies are dominated by bond-stretching, bond-bending, and non-bonded interaction. When $0.6 < T^* < 1.0$, the total potential energy slightly increases with the temperature, and the major parts in the function of total potential energy are attributed to bond-stretching and bond-bending energies. According to the discussion mentioned above, we can conclude that the change in total potential energy with the system temperature corresponds to the competition between bond- and non-bonded interactions.

3.3. Comparison between simulation results and predicted values

To obtain the general rules on the energy transition of proteins, MD simulations on 146 protein samples have been performed under $0.1 < T^* < 1.0$. According to Eqs. (2) and (3), the bond-stretching and bond-bending potential energies are both linearly proportional to the protein chain length, and independent of secondary structure or protein category. Based on the view of thermodynamics, an amplitude of the variation of bond length/angle is quantified under the equilibrium state, i.e., $\overline{(r-\sigma)^2}$ or $\overline{(\theta-\theta_0)^2}$ is independent of the protein at a certain temperature. Therefore, the total bond-stretching/bending potential energy of the protein chain is the summation of all bonds. In the present study, the bond-stretching/bending energies of each bond at a certain temperature are calculated by averaging simulation results on 50 all-α and 49 all-β proteins (protein No. from 1 to 99 listed in Table 1), and the statistical results are shown in Fig. 3(a). The figure shows a strong linear relationship between the bond-stretching/bending potential energy per bond and the reduced temperature. Based on the theorem of equipartition of energy, the average energy associated with each independent degree of freedom is linearly proportional to the system temperature, which can explain the straight lines in Fig. 3(a). On the other hand, the slope of bond-stretching energy per bond as the function of temperature is 0.53, which is larger than 0.45, which is the slope of bond-bending energy. At $T^* = 0.3$, the values of bond-stretching and bond-bending energy per bond are exactly the same. This result is in good agreement with the previous discussion in Section 3.2, in which the bond-stretching and bond-bending potential energies of 1FCS, 2MCM, 9RNT, and 3CHY proteins are all the same at $T^* = 0.3$ in Fig. 2.

The bond-stretching and bond-bending potential energies of each bond as a function of temperature are obtained by the statistical results on 99 protein samples, which are independent of residue type and protein category. Therefore, the values of bond-stretching/bending potential energy of any protein can be predicted with a certain chain length. Figures 3(b) and 3(c) show the bond-stretching and bond-bending potential energy of 24 $\alpha+\beta$ and 23 α/β protein samples (protein No. from 100 to 146 listed in Table 1), respectively, by comparing simulated and predicted results at $T^* = 0.5$. The extremely close values of bond-stretching/bending potential energy obtained by simulation or the prediction method are shown in these two figures. The correlation coefficient R is higher than 0.999, which indicates the nearly perfect agreement is achieved.

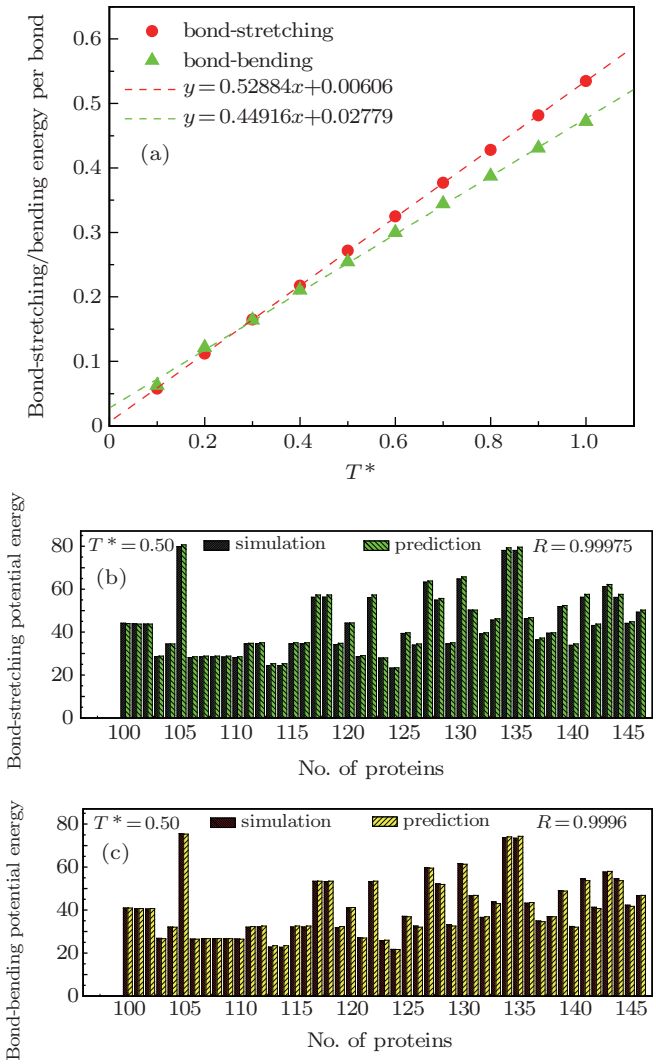


Fig. 3. (color online) Statistical and predicted values of bond-stretching and bond-bending potential energy. (a) Bond-stretching (•) and bond-bending (▲) potential energy of each bond versus reduced temperature T^* based on simulation results from 99 all-α and all-β proteins, (b) simulated and predicted values of bond-stretching potential energy of 47 other proteins at $T^* = 0.5$, (c) simulated and predicted values on bond-bending potential energy of 47 other proteins at $T^* = 0.5$.

Similar to the analysis of bond-stretching/bending potential energy, the statistical result of bond-torsional potential energy per bond as a function of temperature is shown in Fig. 4(a). As the expression of bond-torsional potential energy is not the square of the independent degree of freedom, then the curve of bond-torsional energy per bond shows a nonlinear relationship with the temperature. Equation (4) also shows that the bond-torsional potential energy involves the addition to the residues in its secondary regions. The parameters in Eq. (4) are different according to the different secondary structures

which the residues participate in. Here, the statistical results in Fig. 4(a) are obtained from 38 all-α and 9 all-β protein samples that their secondary-structural regions exclusively contain α-helices and β-sheets, respectively. In Fig. 4(a), the change of two curves corresponding to the α-helix and β-sheet with variable temperature is similar, i.e., the value of bond-torsional potential energy of each bond gradually increases to the maximum and then decreases with the increasing temperature. The differences between the bond-torsional energy per bond of α-helix and β-sheet include two aspects. First, bond-torsional energy per bond of the α-helix is lower than that of the β-sheet at any temperature, indicating that the α-helical structure is more stable than the β-sheet. Second, in case of the α-helix, the transition temperature corresponding to the maximum value of bond-torsional potential energy occurs at $T^* = 0.3$, which is higher than that of the β-sheet at $T^* = 0.2$. According to the statistical results shown in Fig. 4(a), the value of bond-torsional potential energy of any protein that even contains varying degrees of both secondary-structural elements can be predicted. For example, when $T^* = 0.5$, the statistical values of bond-torsional energy per bond of the α-helix and β-sheet are -3.939 and -3.372, respectively. Then, the bond-torsional potential energy of each protein chain at $T^* = 0.5$ is calculated as $-3.939N_{\alpha-\text{helix}} - 0.372N_{\beta-\text{sheet}}$. The values of $N_{\alpha-\text{helix}}$ and $N_{\beta-\text{sheet}}$ are the number of residues as the elements of the α-helix and β-sheet, respectively. Figure 4(b) shows the bond-torsional potential energy of 24 $\alpha+\beta$ and 23 α/β protein samples, which include both helices and sheets at its secondary structure level at $T^* = 0.5$. The values of bond-torsional potential energy by simulated and predicted methods are closely approximated to each other, and the correlation coefficient is $R = 0.995$. This result also verified the prediction is reasonable and effective in the study of bond-potential energy of proteins. The correlation coefficient R increases with the increasing temperature, which indicates the relatively small error between simulated and predicated bond-potential energy at high temperature.

The non-bonded potential energy involves two body interactions that are at least four residues apart in the protein sequence. In the HNP model, 20 amino acids are classified as three types by its hydrophobic property. According to the discussion mentioned in Section 3.2, the interactions between H–H residues contribute significantly in stabilizing the configuration of a protein. Then, the non-bonded potential energy of each protein sample under variable temperature is simply divided by possible H–H pairwise number $N_{\text{H}}(N_{\text{H}}-1)/2$, and the results are shown in Fig. 5(a). Here, N_{H} is the number of H type residues of the protein. The curve of non-bonded potential energy per H–H pair as a function of temperature shows the same tendency in Fig. 2. The non-bonded potential energy of each H–H pair decreases when T^* increases from 0.1 to 0.2. In the medium temperature range $0.2 < T^* < 0.6$, the values increase rapidly to –0.01 with the increasing temperature. When $T^* > 0.6$, the non-bonded potential energy of each H–H pair reaches a plateau with almost constant value. Figure 5(a) also shows that the non-bonded potential energy per H–H pair of all-α proteins is lower than that of all-β proteins, especially under the low temperature, which illustrated that all-α proteins present higher structural stability than all-β proteins. The non-bonded potential energy of 24 $\alpha+\beta$ and 23 α/β protein chains with the number of H components listed in Table 1 are also evaluated at $T^* = 0.5$. The statistical results of non-bonded potential energy per H–H pair of all-α and all-β proteins under $T^* = 0.5$ are -0.0312 and -0.0262, respectively. The total non-bonded potential energy of each chain at $T^* = 0.5$ is calculated as

$$-0.0312\frac{N_{\alpha-\text{helix}}}{N_{\alpha-\text{helix}}+N_{\beta-\text{sheet}}}\frac{N_{\text{H}}(N_{\text{H}}-1)}{2},$$
$$-0.0262\frac{N_{\beta-\text{sheet}}}{N_{\alpha-\text{helix}}+N_{\beta-\text{sheet}}}\frac{N_{\text{H}}(N_{\text{H}}-1)}{2}.$$

The comparative results are shown in Fig. 5(b). The value of correlation coefficient, $R = 0.757$, is consequently obtained. Comparing Figs. 3(b), 3(c), and 4(b), the numerical gap of the non-bonded potential energies between simulated and

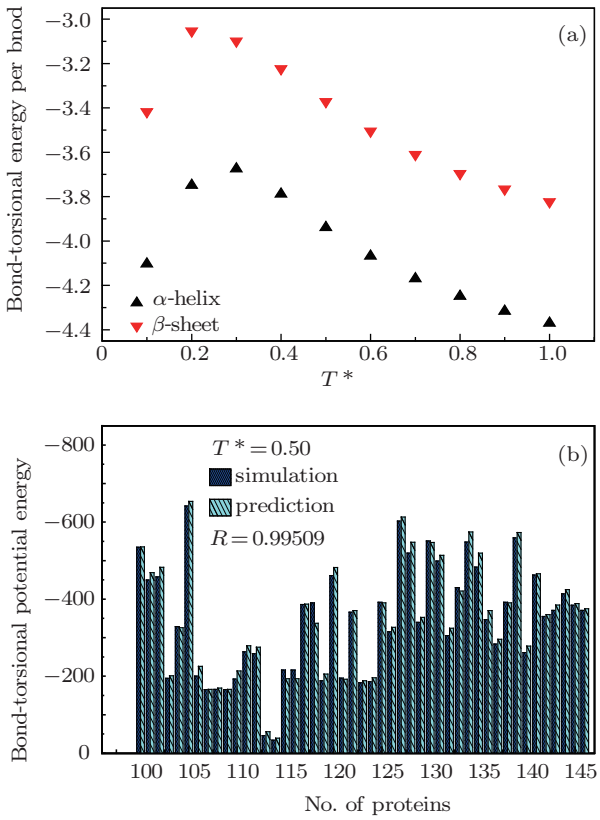


Fig. 4. (color online) Statistical and predicted values of bond-torsional potential energy. (a) Bond-torsional potential energy of each bond with the residues constructing α-helices (▲) or β-sheets (▼) versus reduced temperature T^* based on simulation results from 99 all-α and all-β proteins, (b) simulated and predicted values of bond-torsional potential energy of other 47 proteins at $T^* = 0.5$.

predicted results in Fig. 5(b) is relatively large. The correlation coefficient R as a function of temperature is also shown in Fig. 5(b). A lower value of non-bonded potential energy corresponds to a higher value of R obtained. The reason for the comparatively low value of correlation coefficient R mainly lies in disregarding other residue–residue interactions, such as H–N, H–P, N–P, N–N, and P–P. Besides H–H pairwise interactions, interactions between other residue types also partly contribute to the non-bonded potential energy, especially with the large value of potential energy under high temperature. Therefore, the higher non-bonded potential energy leads to the relatively large error.

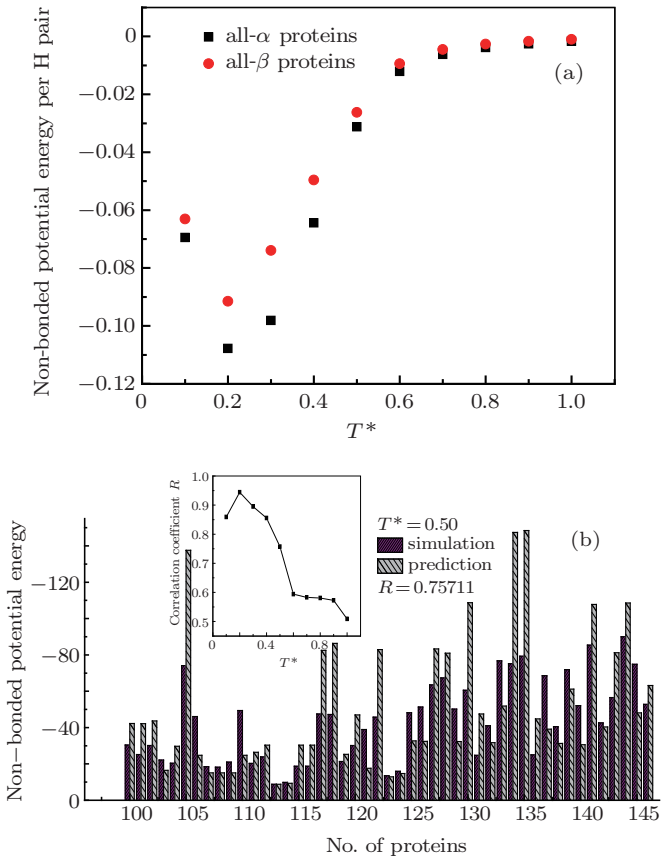


Fig. 5. (color online) Statistical and predicted values of non-bonded potential energy. (a) Non-bonded potential energy of each H–H pair versus reduced temperature T^* based on simulation results from 50 all-α proteins (■) and 49 all-β proteins (•), (b) simulated and predicted values of non-bonded potential energy of 47 other proteins at T^*=0.5 and the correlation coefficient R versus reduced temperature T^* in the inserted figure.

4. Conclusions

The biological function of a protein is related to its unique structure determined by complex interactions among protein residues. In the present article, we studied the potential energy of 146 protein samples, including bond-stretching, bond-bending, bond-torsional, and non-bonded potential energy, by a coarse-grained HNP model. MD simulations are carried out under a broad temperature range. The predicted values of the four parts of potential energy are also compared with simulated data obtained directly by the MD method. These results presented in our work are as follows.

(I) The ratio of residues participating in forming helices or sheets as secondary-structural elements to total residues of a protein are strongly relevant to protein categories, i.e., all-α, all-β, $\alpha+\beta$, and α/β classes. On the contrary, the frequency of the occurrence of P type residues is higher than that of H type according to the compositional analysis of 146 protein samples. It is completely unrelated to the protein categories classified by the SCOP database.

(II) From the perspective of the propensity that three types of residues H, N, and P to construct secondary structures, we have concluded that P residues easily form α-helices, whereas H ones present a higher tendency to construct β-sheets. This result indicates that α-helices are mostly located on the surface of the protein, while β-sheets aggregate inside of it. Such a kind of configuration makes a globular protein consisting of a hydrophobic core and a hydrophilic exterior. This spatial distribution of secondary structure by amino acid hydrophobicity is useful in discriminating native from non-native protein structures.

(III) The bond-stretching/bending potential energy increases linearly as a function of temperature. We attempted to explain this trend by the theorem of equipartition of energy according to the average value of bond-stretching/bending potential energy associated with each independent degree of freedom r or θ. The slope of the bond-stretching potential energy as the function of temperature is larger than that of the linear curve of bond-bending potential energy. The values of bond-stretching and bending energy are the same at $T^* = 0.3$, regardless of protein samples.

(IV) The bond-torsional potential energy gradually increases to the maximum value and then decreases with increasing temperature. The transition temperature corresponding to the maximum value of bond-torsional energy is related to the secondary structure of the protein. The bond-torsional potential energy per bond of α-helices is lower than that of β-sheets under a wide temperature range, which indicates that the α-helix is more stable than the β-sheet in the secondary-structural level. For the α-helix, the transition temperature is $T^* = 0.3$, which is higher than that of the β-sheet at $T^* = 0.2$. Both numbers of residues constructing α-helices or β-sheets would be taken into account in evaluating the total bond-torsional energy of each protein.

(V) In contrast to the tendency of the bond-torsional potential energy as the function of temperature, the non-bonded potential energy decreases until $T^* = 0.2$. Then the values increase rapidly to -0.01 in the medium temperature range $0.2 < T^* < 0.6$ and slowly increase when $T^* > 0.6$. The curve tendency of non-bonded potential energy per H–H pair is the

same for all-α and all-β proteins. However, the average value of all-α samples is lower than that of all-β proteins under a certain temperature, which indicates the higher stability of the α-helical structure. The change in the total potential energy of each protein is attributed to the competition between bond- and non-bonded interactions.

The HNP model presented in this article simplifies a single protein chain to the sequence of three types of beads. In considering the transition of potential energy, the HNP model effectively yields the essentials and successfully predicts the bonded potentials. Semi-quantitative prediction of non-bonded potential energy is also achieved. These results provide us with a natural realization of the energy system of proteins.

Acknowledgment

The computations for this research were performed using the Shuguang HPC supercomputer of China Jiliang University.

References

[1] Bratko D and Blanch H W 2001 *J. Chem. Phys.* **114** 561
[2] Anfinsen C B 1973 *Science* **181** 223
[3] Harrison P M, Chan H S, Prusiner S B and Cohen F E 1999 *J. Mol. Biol.* **286** 593
[4] Asherie N, Pande J, Lomakin A, Ogun O, Hanson S R A, Smith J B and Benedek G B 1998 *Biophys. Chem.* **75** 213
[5] Jiang T Z, Cui Q H, Shi G H and Ma S D 2003 *J. Chem. Phys.* **119** 4592
[6] Ou-Yang Z C, Su Z B and Wang C L 1997 *Phys. Rev. Lett.* **78** 4055
[7] Li T, Jiang Z T, Yan D D and Nies E 2010 *Polymer* **51** 5612
[8] Chen J X, Mao J W, Thakur S, Xu J R and Liu F Y 2011 *J. Chem. Phys.* **135** 094504
[9] Chen J X, Zhu J X, Ma Y Q and Cao J S 2014 *Europhys. Lett.* **106** 18003
[10] Sakae Y and Okamoto Y 2010 *Mol. Simulat.* **36** 302
[11] Bond P J and Sansom M S P 2006 *J. Am. Chem. Soc.* **128** 2697
[12] Li C and Strachan A 2014 *J. Polym. Sci. B* **53** 103
[13] Chai A H, Ran S Y, Zhang D, Jiang Y W, Yang G C and Zhang L X 2013 *Chin. Phys. B* **22** 098701
[14] Yang J S, Huang D H, Cao Q L, Li Q, Wang L Z and Wang F H 2013 *Chin. Phys. B* **22** 098101
[15] Ferreon A C M and Deniz A A 2011 *BBA-Proteins Proteom.* **1814** 1021
[16] O'Connell T M, Wang L, Tropsha A and Hermans J 1999 *Proteins* **36** 407
[17] Vettorel T, Grosberg A Y and Kremer K 2009 *Phys. Biol.* **6** 025013
[18] Faccioli P, Sega M, Pederiva F and Orland H 2006 *Phys. Rev. Lett.* **97** 108101
[19] Jiang Z T, Zhang L X, Sun T T and Wu T Q 2009 *Chin. Phys. B* **18** 4580
[20] Jiang Z T, Zhang L X, Chen J, Xia A and Zhao D 2002 *Polymer* **43** 6037
[21] Dill K A 1985 *Biochemistry* **24** 1501
[22] Yue K, Fiebig K M, Thomas P D, Chan H S, Shackhnovich E I and Dill K A 1995 *Proc. Natl. Acad. Sci. USA* **92** 325
[23] Wust T and Landau D P 2012 *J. Chem. Phys.* **137** 064903
[24] Sangha A K and Keyes T 2010 *J. Phys. Chem. B* **114** 16908
[25] Regan L and DeGrado W F 1988 *Science* **241** 976
[26] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623
[27] Gutin A M, Abkevich V I and Shakhnovich E I 1996 *Phys. Rev. Lett.* **77** 5433
[28] Taneri S 2008 *Int. J. Mod. Phys. C* **19** 749
[29] Levitt M and Chothia C 1976 *Nature* **261** 552
[30] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 *J. Mol. Biol.* **247** 536
[31] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534
[32] Friedel M, Sheeler D J and Shea J E 2003 *J. Chem. Phys.* **118** 8106
[33] Jiang Z T, Xu P and Sun T T 2012 *Chinese J. Polym. Sci.* **30** 45
[34] Nose S 1984 *J. Chem. Phys.* **81** 511
[35] Hoover W G 1985 *Phys. Rev. A* **31** 1695
[36] Pace C N and Scholtz J M 1998 *Biophys. J.* **75** 422
[37] Best R B, de Sancho D and Mittal J 2012 *Biophys. J.* **102** 1462
[38] Minor D L and Kim P S 1993 *Nature* **367** 660