ELSEVIER

Polymer 43 (2002) 6037–6047

polymer

www.elsevier.com/locate/polymer

Effect of amino acid on forming residue–residue contacts in proteins

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

[a]*Department of Physics, Zhejiang University, Hangzhou 310028, 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 14 February 2002; received in revised form 15 July 2002; accepted 15 July 2002

Abstract

The long-range contacts contribute more function to the protein folding and play an active role in the stability of protein molecules. In this paper, we calculated the number of short- and long-range contacts from 278 globular proteins and analyzed the effects of amino acids on the long-range contacts by contrasting the average number of the long-range contacts between different amino acid residues in the protein sample. The amino acids of Leu, Val, Ile, Met, Phe, Tyr, Cys, and Trp are easy to form the long-range contacts, and the average number of long-range contacts per residue is 5.008 when $R_c = 0.80$ nm. Here R_c is the minimum distance between two C^α atoms of residues. The amino acids of Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro are difficult to form the long-range contacts, and the average number of long-range contacts per residue is only 3.232 when $R_c = 0.80$ nm. However, the effect of amino acid on the short-range contact is negligible, and the average number of short-range contacts per residue ranges from 3.649 to 3.721 when $R_c = 0.80$ nm. We also find that the highest preference is observed for Cys-Cys contact, and the lowest preference is Gln-His contact. The average number of contacts depends on R_c, and two cases of $R_c = 0.65$ and 0.80 nm are discussed. The average distance of the residue–residue contacts is also concluded. Through these calculations, we can discuss how the amino acids affect the protein folding and how the proteins achieve the stability conformations. © 2002 Published by Elsevier Science Ltd.

Keywords: Amino acid; Short-range and long-range contacts; Protein folding

1. Introduction

Proteins are heterogeneous chain molecules composed of sequences of amino acids. The understanding of protein folding is a long-standing goal in structural biology. There are 20 different amino acids in protein molecules, and their effects may be different in the folding. In the protein folding the amino acid residue–residue contacts play an important role. The folding of a protein chain into a compact, unique three dimension structure is directed and stabilized by intra molecular interactions between the constituent amino acid residues along the chain. And it is useful to predict the three-dimensional structure of a globular protein from the knowledge of its amino acid sequence. During the last two decades, many biologists, chemists, and physicists have attempted to identify and simulate the mechanisms through which a given sequence reaches its stable which has the lowest free energy, native conformation (protein folding) [1–3]. The complexity of the problem is enormous because the calculation is not yet practical using current force-field algorithms and the computer time required is far too great.

Experiments and theoretical studies have shown that the amino acids are subdivided into two kinds of residues: hydrophobic (H) and polar (P) [4–6]. There is an effective attraction between hydrophobic amino acids that arises from their aversion to the solvent and lead to such amino acids forming the core in the protein native state. In the HP model, the hydrophobic energy, or the solvent effects, is a major contributor to the energetics of protein folding. A favorable contact energy, $\varepsilon_{HH} = -1$ (or $-\varepsilon_0$) is assigned to two non-consecutive residues which are one lattice spacing apart, while the other interactions, ε_{PH}, ε_{HP}, ε_{PP} are set equal to zero [4–7]. This simplified model is required for the study of the protein folding process and simplification can be made to both geometry and potential functions. In the recent years, protein engineering experiments also suggest that maybe not two but certainly several amino acids can be effectively substituted for the 20 amino acids and the helical bundles can be built with the set of three amino acids: hydrophobic Leucine (L), polar E (Glutamine), and polar K (Lysine) [8]. Comply with this pattern, Miyazawa–Jernigan

* Corresponding author.
E-mail address: zhanglx@mail.hz.zj.cn (L. Zhang).

0032-3861/02/$ - see front matter © 2002 Published by Elsevier Science Ltd.
PII: S0032-3861(02)00501-3

Table 1
Number of short-range and long-range contacts in different globular proteins. N_S (or N_L) represents number of short (or long)-range contacts, and N is the number of amino acid residues

Number	PDB code	N	N_S[a]	N_L[a]	N_S[b]	N_L[b]
1	1ABA	87	166	138	128	66
2	1BBL	51	76	24	55	9
3	1BOV-A	69	106	157	66	86
4	1BOV-B	69	105	152	70	90
5	1BOV-C	69	104	148	68	84
6	1BOV-D	69	104	154	67	88
7	1BOV-E	69	106	154	69	89
8	1CDT-A	60	83	145	39	84
9	1CDT-B	60	84	145	42	87
10	1CRN	46	91	67	69	38
11	1CIF	74	140	117	120	61
12	1CYO	93	170	119	126	70
13	1DUR	55	86	100	59	49
14	1FIA-A	98	194	28	172	16
15	1FIA-B	98	194	27	175	14
16	1FXD	58	101	103	75	47
17	1GCN	29	69	1	54	0
18	1HIP	85	140	163	91	76
19	1HIV-A	99	132	203	69	120
20	1HIV-B	99	132	198	71	116
21	1HOE	74	96	186	47	117
22	1LMB-3	92	214	85	187	33
23	1LMB-4	92	224	77	192	33
24	1LTS-C	41	101	2	93	0
25	1NRC-A	95	148	153	110	83
26	1NRC-B	95	138	151	104	80
27	1PPT	36	75	24	69	7
28	1RDG	53	77	91	60	50
29	1TEN	90	113	214	46	140
30	1TGS-I	56	83	93	48	56
31	1TPA-I	58	88	127	58	71
32	1UTG	70	179	45	158	14
33	2MLT-A	27	69	3	61	1
34	2MLT-B	27	68	1	64	0
35	2OVO	56	92	94	63	49
36	2PCY	99	138	242	78	142
37	2PDE	43	60	88	38	43
38	2SAR-A	96	153	168	97	99
39	2SAR-B	96	152	170	101	100
40	2SN3	65	104	154	67	90
41	3EBX	62	75	146	33	88
42	3ILB	72	117	96	78	51
43	3INS-A	21	44	7	32	2
44	3INS-B	30	53	7	44	2
45	3INS-C	21	44	9	36	5
46	3INS-D	30	54	8	46	2
47	451C	82	179	111	144	54
48	4ICB	76	178	76	143	27
49	5RXN	54	80	89	58	49
50	1A45	173	239	481	108	274
51	1ACX	108	141	268	61	145
52	1BP2	123	267	180	216	75
53	1CCR	122	219	194	178	89
54	1CD8	114	150	258	83	147
55	1CID	177	232	469	112	267
56	1COB-A	151	216	457	124	273
57	1COB-B	151	213	453	117	276
58	1CPC-A	162	420	172	369	67
59	1CPC-B	172	437	157	396	61
60	1CPC-K	162	420	167	368	66
61	1CPC-L	172	438	158	391	62
62	1ECD	136	350	107	316	32
63	1ECO	136	352	109	314	33
64	1FCS	153	401	146	366	44
65	1FHA	183	437	179	396	61
66	1FKF	107	169	264	98	148
67	1FX1	148	284	315	211	150
68	1GKY	187	386	276	299	139
69	1HBG	147	382	185	337	69
70	1I55-A	103	208	180	172	83
71	1I55-B	103	208	175	171	81
72	1IFA	158	376	141	312	33
73	1LE4	144	380	93	343	27
74	1LH1	153	392	159	334	49
75	1LTS-D	103	177	178	117	101
76	1LTS-E	103	176	178	115	98
77	1LTS-F	103	176	177	118	103
78	1LTS-G	103	177	180	121	103
79	1LTS-H	103	179	181	119	97
80	1LTS-A	185	327	389	238	207
81	1LZ1	130	262	235	215	104
82	1MBC	153	400	146	364	47
83	1MBD	153	400	135	365	43
84	1MBS	153	399	169	351	56
85	1MSB-A	115	188	246	128	136
86	1MSB-B	115	188	234	126	125
87	1OFV	169	330	361	250	179
88	1OVB	159	303	370	214	189
89	1P12-E	198	283	619	159	348
90	1PAE	123	186	277	117	159
91	1PP2-R	122	254	187	199	95
92	1PP2-L	122	254	178	192	90
93	1Q21	171	316	333	245	167
94	1RBP	182	260	395	141	232
95	1REI-A	107	145	283	71	161
96	1REI-B	107	144	275	68	164
97	1RNH	155	274	310	196	167
98	1SRX	108	219	196	148	80
99	1TFG	112	178	253	116	145
100	1TIE	172	222	422	125	268
101	1TLK	154	132	247	65	144
102	1YCC	107	215	183	170	83
103	256B-A	106	277	117	247	34
104	256B-B	106	276	121	249	39
105	2ALP	198	281	622	160	353
106	2AVI-A	128	157	290	68	188
107	2AVI-B	128	156	287	71	190
108	2AZA-A	129	202	309	119	181
109	2AZA-B	129	204	308	117	177
110	2C2C	112	230	190	189	87
111	2CCY-A	128	324	140	282	53
112	2CCY-B	128	324	137	285	52
113	2CDV	107	197	142	145	69
114	2CY3	118	212	138	161	65
115	2FCR	173	326	368	248	174
116	2FOX	138	276	262	214	127
117	2GMF-A	127	253	155	209	57
118	2GMF-B	127	249	145	214	58
119	2HCO-A	141	358	149	334	46
120	2HCO-B	146	370	156	327	47
121	2ILA	155	192	384	98	236
122	2LAL-A	181	241	346	127	202

Table 1 (*continued*)

Number	PDB code	N	N_S[a]	N_L[a]	N_S[b]	N_L[b]
123	2LAL-C	181	238	340	125	201
124	2LHB	149	368	155	324	59
125	2LTN-A	181	241	346	126	201
126	2LTN-B	181	241	341	124	202
127	2LYZ	129	265	236	207	104
128	2LZM	164	385	205	332	85
129	2MHB-A	141	359	153	335	45
130	2MHB-B	146	370	171	338	50
131	2MHR	118	292	97	266	28
132	2MS2-A	129	205	212	137	142
133	2MS2-B	129	212	187	147	114
134	2MS2-C	129	206	215	136	141
135	2PAB-A	127	158	262	85	162
136	2PAB-B	127	157	259	83	158
137	2RHE	114	159	294	78	171
138	2RSP-A	124	150	239	78	137
139	2RSP-B	124	150	228	79	132
140	2SNS	149	263	300	181	144
141	2SNV	151	217	370	124	220
142	2SOD-O	152	211	444	108	268
143	2SOD-Y	152	210	449	107	275
144	2SOD-B	152	215	458	113	275
145	2SOD-G	152	213	454	113	268
146	2STV	195	265	484	135	284
147	2TRX-A	108	209	175	156	88
148	2TRX-B	108	209	178	149	99
149	2WRP-R	107	267	39	239	9
150	3ADK	195	438	270	365	112
151	3CHY	128	259	225	212	123
152	3DFR	162	263	315	175	186
153	3LYZ	129	264	236	208	100
154	3RN3	124	207	260	139	148
155	3SGB-E	185	266	567	143	328
156	3SSI	113	176	228	112	138
157	4CPV	109	246	139	204	50
158	4DFR-A	159	255	324	165	179
159	4DFR-B	159	258	323	164	183
160	4MBN	153	402	137	362	46
161	5CPV	109	250	138	205	51
162	5CYT	104	205	178	169	83
163	5FD1	106	186	215	138	98
164	5P21	166	303	315	234	165
165	7RSA	124	204	259	135	145
166	8ATC-B	153	231	289	144	164
167	8ATC-D	153	234	296	143	164
168	9RNT	104	170	196	110	106
169	9WGA-A	171	301	437	189	256
170	9WGA-B	171	303	427	192	238
171	1BKS	268	560	473	465	221
172	1CA2	259	386	652	232	363
173	1COL-A	204	511	264	461	85
174	1COL-B	204	508	268	457	85
175	1CSE-E	274	502	814	367	404
176	1DHR	241	460	494	360	236
177	1EAF	243	461	433	360	215
178	1EST	240	372	657	223	368
179	1FC1-A	224	293	475	155	280
180	1FC1-B	224	295	469	149	279
181	1HIL-A	217	298	525	156	322
182	1HIL-B	220	291	528	157	307
183	1HIL-C	217	297	515	157	322
184	1HIL-D	220	288	528	150	307
185	1HSB-A	270	480	538	321	318

Table 1 (*continued*)

Number	PDB code	N	N_S[a]	N_L[a]	N_S[b]	N_L[b]
186	1MAM-L	214	297	517	154	309
187	1MAM-H	217	291	510	137	303
188	1PPF-E	218	320	583	176	317
189	1PPN	212	374	538	258	257
190	1PRC-L	273	659	338	560	134
191	1PRC-H	259	451	406	304	213
192	1RHD	293	561	585	379	270
193	1RVE-A	245	441	441	302	217
194	1RVE-B	245	435	435	311	212
195	1TGS-Z	229	344	610	203	326
196	1TIM-A	247	505	465	396	221
197	1TIM-B	247	509	444	390	211
198	1TPA	223	336	606	202	331
199	1TRE-A	255	537	505	426	225
200	1TRE-B	255	528	494	422	221
201	1ULA	289	523	550	370	251
202	2ACT	220	384	551	262	259
203	2AYH	214	285	604	137	355
204	2CAB	261	397	682	246	377
205	2CNA	237	324	685	142	376
206	2DRI	271	556	651	443	326
207	2GCH	245	355	644	214	346
208	2PTC-E	223	337	602	201	333
209	2SBT	275	496	784	336	363
210	2TSC-A	264	484	499	353	245
211	2TSC-B	264	480	501	353	251
212	3CNA	237	320	668	150	381
213	3EST	240	365	652	218	362
214	3PGM	241	457	466	317	216
215	4BLM-A	265	517	530	402	248
216	4BLM-B	265	520	538	397	251
217	4CHA-A	245	358	668	215	347
218	4CHA-B	245	355	657	213	339
219	4CLA	213	385	365	278	180
220	4FAB-L	219	308	537	152	307
221	4FAB-H	216	293	549	138	325
222	5PTP	223	332	599	194	327
223	5TIM-A	250	507	469	394	216
224	5TIM-B	250	512	480	399	221
225	9PAP	212	367	531	256	244
226	1ABE	306	631	662	500	321
227	1ALD	363	744	797	592	371
228	1AVH-A	320	785	393	669	121
229	1AVH-B	320	783	399	660	124
230	1CD1-A	315	469	502	322	290
231	1CD1-C	315	471	509	320	288
232	1ETU	394	359	340	277	142
233	1GOX	369	707	702	528	324
234	1IPD	345	719	768	557	347
235	1MNS	357	707	800	556	397
236	1NSB-A	390	513	1198	243	678
237	1NSB-B	390	515	1196	250	678
238	1PAX	361	668	669	502	332
239	1PFK-A	320	658	687	494	325
240	1PFK-B	320	658	680	497	321
241	1PHH	394	771	807	537	408
242	1PRC-C	336	698	481	542	189
243	1PRC-M	323	745	383	639	181
244	1SBP	310	655	606	514	297
245	2ACH-A	360	620	690	442	370
246	2APR	325	489	880	287	491
247	2ER7-E	330	491	890	288	513

(*continued on next page*)

Table 1 (*continued*)

Number	PDB code	N	N_S[a]	N_L[a]	N_S[b]	N_L[b]
248	2GBP	309	641	643	508	316
249	2HAD	310	633	610	479	270
250	2LIV	344	710	731	556	343
251	2PIA	321	524	697	348	375
252	2POR	301	420	868	193	526
253	3APP	323	488	875	285	483
254	3CPA	307	605	693	466	331
255	3LDH	330	659	624	465	284
256	4CPPA	307	611	691	459	329
257	4PFK	319	661	736	504	346
258	4TLN	316	644	742	507	376
259	4TMS	316	618	541	481	265
260	5ABP	306	630	654	494	316
261	5ADH	374	674	962	447	473
262	5CPA	307	604	688	473	328
263	6LDH	330	669	610	507	299
264	6XIA	387	829	643	669	289
265	8ADH	374	669	965	457	464
266	8TLN-E	316	643	724	501	674
267	1GLA-G	501	944	1172	703	581
268	2AAA	484	883	1066	641	510
269	2BPA-1	426	704	842	459	450
270	2CTS	437	1026	622	865	226
271	2PGD	482	1062	820	881	368
272	2TAA-A	478	861	1057	579	487
273	2TS1	419	700	493	557	202
274	3PGK	416	807	908	697	383
275	4ENL	436	891	1023	656	482
276	4ICD	416	828	895	656	416
277	8CAT-A	506	929	877	672	431
278	8CAT-B	506	924	875	661	428

[a] The case of $R_c = 0.80$ nm.
[b] The case of $R_c = 0.65$ nm.

(MJ) modified the interaction matrix [9]. They put forward that there are three types of amino acids named hydrophobic residues (H) consisting of Leu, Phe, Met, Ile, Trp, Val, Cys, and Tyr, neutral residues (N) consisting of His, Ala, Gly, and Thr, and hydrophilic residues (P) consisting Lys, Asp, Asn, Glu, Gln, Ser, Pro, and Arg [10]. As a matter of fact, the two kinds of residues (hydrophobic and polar) in standard HP lattice model are replaced by three kinds of residues (hydrophobic, neutral, and hydrophilic) in the modified HP lattice model [11]. In this modified model, the interactions between those residue pairs may be more close to the interactions in the real proteins. On the other hand, there are different roles in forming contacts for different amino acids. Hydrophobic amino acid residues have a strong attractive in forming a long-range contact. In the meantime, there may be no difference for hydrophobic and hydrophilic amino acid residues in forming short-range contacts. In this paper, we will discuss the effects of amino acid residues on forming short-range and long-range contacts through calculating the number of the short-range and long-range residue–residue contacts. Our aim is to study the important role of amino acids residue–residue contacts in the protein folding.

2. Method of the calculation

We study 278 globular protein structures. A database of these proteins is derived from the information about their three-dimensional structures available in the literature. We use a set of representative globular proteins obtained from http://www.rcsb.org/. The coordinates of all the globular protein structures are obtained from the protein date bank (PDB). The PDB codes for all the proteins used in the present study along with the protein length and the number of the contacts are given in Table 1.

Each residue in a protein molecule is represented by the center of its side chain atom positions, and the position of C^{α} atom is used for glycine residue. Residues whose centers are closer than R_c are defined to be in contact. This kind of simple method to evaluate the number of residue–residue contacts in proteins has often been used [9–14]. The limiting values $R_c = 0.65$ and 0.80 nm for contacts are chosen. Using the C^{α} coordinates, a sphere of radius R_c is fixed around each residue, and the composition of surrounding residues associated with all the residues is calculated. It has been shown that the influence of each residue over the surrounding medium extends effectively only up to 0.80 nm [15–17]. The limit of 0.80 nm is sufficient to characterize the hydrophobic behavior of amino acid residues [16] and to accommodate both the local and non-local interactions [18]. This limit also has been used to understand the folding rate of protein [19], protein stability upon mutations [20] and thermal stability of proteins [21]. In the previous papers, $R_c = 0.65$ nm is also chosen in estimation of effective interresidue contact energies by Miyazawa and Jernigan [9–11].

For a given residue, the composition of surrounding residues is analyzed in terms of the location at the sequence level, and the contributions from $\leq \pm 4$ residues are treated as a short-range contact, and $> \pm 4$ residues as long-range contact [21–23], which is the same as the short-range and the long-range interactions in the rotational-isomeric-state model [24].

A contact consists of two residues A and B. Sometime, residue A and residue B may be the same. If residue A and residue B have the high tendency of appearance in the proteins, there maybe a large probability of forming contacts. Here the preference of all the 20 amino acid residues to form contacts is computed. The average number P_{A-B} of residue–residue contacts for different amino acids A and B is defined as

$$P_{A-B} = \frac{N_{A-B}}{\sqrt{N_A N_B}} \quad (1)$$

Here N_{A-B} represents the number of the contacts between residue A and residue B, and N_A and N_B are the total numbers of the residues of type A and B in the 278 protein chains, respectively.

Average number of contacts per residue indicates the ability of forming contacts. Here we define the average

Table 2
Number of residue–residue contacts N_{A-B} for different amino acids A and B. Upper triangle counts the long-range contacts, and lower triangle counts the short-range contacts in protein samples. Here $R_c = 0.80$ nm

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
	1862	2128	1542	412	979	742	423	381	1944	1513	1176	329	582	547	653	519	782	1045	530	674	Leu
Leu	1560	2642	1695	398	958	718	494	321	2049	1541	1326	387	713	514	627	511	810	1199	618	685	Val
Val	1113	948	1282	304	717	651	365	254	1522	1066	954	258	487	390	508	410	655	818	462	412	Ile
Ile	909	677	586	146	212	147	114	87	363	367	271	88	151	132	143	134	177	227	155	153	Met
Met	379	242	231	84	452	370	237	181	825	687	636	218	278	280	342	301	350	569	278	332	Phe
Phe	625	512	401	123	370	448	295	222	709	719	516	185	278	291	340	372	460	516	376	376	Tyr
Tyr	573	426	388	163	236	296	996	111	376	578	371	116	181	180	215	188	245	428	153	214	Cys
Cys	291	238	164	67	102	138	160	66	300	300	203	84	104	132	134	141	169	218	170	169	Trp
Trp	256	221	155	51	100	93	52	88	1752	1517	1104	365	616	523	708	642	764	1113	523	729	Ala
Ala	1522	1343	937	352	644	560	321	270	2182	2204	1338	371	642	599	924	763	805	1335	681	825	Gly
Gly	1293	1019	767	245	547	563	348	268	1459	1298	1006	306	486	422	681	579	598	1008	499	623	Thr
Thr	827	729	535	189	390	417	207	179	997	1027	762	168	172	95	254	167	208	286	163	230	His
His	395	319	208	90	189	147	94	72	376	393	222	132	292	217	259	259	542	490	359	323	Glu
Glu	1044	668	598	225	438	377	166	152	1017	732	640	285	714	170	213	221	305	383	215	304	Gln
Gln	759	513	371	143	248	282	134	137	718	559	428	145	444	274	388	436	553	612	398	355	Asp
Asp	946	794	619	269	448	401	173	163	1100	943	634	250	585	452	598	402	365	573	278	337	Asn
Asn	680	528	494	151	302	372	171	161	721	574	552	169	448	344	551	410	442	625	213	344	Lys
Lys	1131	826	677	226	484	379	200	194	1325	961	702	268	947	459	887	542	850	1022	445	509	Ser
Ser	1081	807	675	225	473	487	255	247	1126	1070	908	257	651	536	759	504	712	1104	254	309	Arg
Arg	785	468	449	199	291	320	111	156	687	572	523	182	571	345	522	315	396	486	320	368	Pro
Pro	553	509	355	124	316	287	108	116	571	560	448	174	363	278	437	263	450	547	254	262	
	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	

number of short-range contacts per residue C_S and the average number of long-range contacts per residue C_L as

$$C_{\alpha,\eta} = \frac{\sum_{\beta=\mathrm{Ala,Asp,Cys,Glu,\ldots,Tyr}} N_{\alpha\text{-}\beta,\eta}}{N_{\alpha,\eta}} \quad (2)$$

$(\eta = \mathrm{S},\ \mathrm{or},\ \mathrm{L};\ \alpha = \mathrm{Ala, Asp, \ldots, Tyr})$

If residue A has a large value of C_L, it means that residue A has a high tendency of forming long-range contacts. These calculations can help us indicate the mechanism of the globular protein folding, and what plays an important role in the protein folding.

3. Results and discussions

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

The numbers of short-range contacts (N_S) and long-range contacts (N_L) of all the 20 amino acid residues in a set of 278 globular proteins are listed in Table 1, here two cases of $R_c = 0.80$ and 0.65 nm are considered, respectively. These globular proteins are the source for our present study. Some results are almost accord with the Gromiha and Selvaraj's work [22,23]. In our calculation, the total number of globular proteins is 278, and is almost twice as large as Gromiha and Selvaraj's work (150). It shows that N_S is greater than N_L for 37.8% proteins, N_L is greater than N_S for 53.2% proteins, and N_L is almost the same as N_S for 9.0% proteins when $R_c = 0.80$ nm. For short proteins with the number of amino acid residues (chain length) $N < 100$ and long proteins with the number of amino acid residues (chain length) $N > 200$, the average number of long-range contacts per protein is greater than the average number of short-range contacts per protein. However, the average number of long-range contacts per protein equals to the average number of short-range contacts per protein, for proteins with $200 > N > 100$. For example, the ratio of N_L/N_S is 1.173 for $N > 200$, and is 0.99 for $200 > N > 100$. It implies the long-range contacts have more advantage to attain the stable tertiary structure in protein, and long-range contacts for each amino acid residues are important for maintaining their foldings.

We also calculate the distribution of the short-range and long-range contacts with $R_c = 0.65$ nm, and the results are also given in Table 1. When the radius R_c decreases, both N_S and N_L also decrease, especially for N_L. The percentage of proteins with $N_S > N_L$ increases from 37.8 to 60.8%, the percentage of proteins with $N_L > N_S$ decreases from 53.2 to 36.0%, and the percentage of proteins with $N_L \approx N_S$ decreases from 8.0 to 3.2% when R_c decreases from 0.80 to 0.65 nm. The reason is that the number of long-range contacts depends more strongly on R_c. In the meantime, the average ratio of N_L/N_S become large and the value is 1.312, which is greater than the case of $R_c = 0.80$ nm.

We also investigate the ability to form a residue–residue contact in all the 20 amino acids. The number of residue–residue contacts for different amino acid is listed in Table 2.

Here $R_c = 0.80$ nm. Upper triangle counts the long-range contacts, and lower triangle counts the short-range contacts in protein samples. In the upper triangle, the maximum number of the residue–residue contacts is 2642, and it occurs to the Val-Val residue–residue contact. In the other hand, the minimum number is 66, and it occurs to the Trp-Trp residue–residue contacts. Although the number of Trp-Trp long-range contacts is only 66, the average number of long-range contacts per residue pair is not minimum because the total number of Trp amino acid is only 1.56% of all the 20 amino acids in 278 protein molecules. Contrast to the lower triangle, the maximum and minimum numbers of short-range contacts is 2182 and 51, respectively, and they exist in the Ala-Ala and Trp-Met residue–residue contacts, respectively. This shows it is important that the probability distribution of the amino acids and the number of contacts are considered simultaneously.

3.2. Preference of amino acid residues in the short- and long-range contacts

It is difficult to conclude which of two residues easy to form a contact only from Table 2 because the percentage of the amino acids is different in all the 20 amino acids. We count all the 20 amino acids and there have 52,667 residues in the 278 different globular proteins. The probability distribution of the amino acids is shown in Fig. 1. In Fig. 1, the Ala amino acid occupies the maximal proportion 8.62% in all the 20 amino acids. The minimum is 1.56%, and this is the Trp amino acid. Those results can help us know distributions of the amino acids clearly. We use the average number P_{A-B} of residue–residue contacts for different amino acids A and B to scale the ability of forming the contacts. Considering the same contacts between A–B and B–A, we only show the results of A–B contact in Tables 3 and 4. Table 3 is the case of $R_c = 0.80$ nm, and we find that Cys-Cys has the largest value of average number of long-range contacts. Val-Val, Val-Leu, Ile-Val, and Ala-Val also have the high value of long-range contacts, and the values are greater than 0.50. The 10 topmost long-range contacts are Cys-Cys, Val-Val, Val-Leu, Ile-Val, Ala-Val, Gly-Gly, Ile-Ile, Ile-Leu, Leu-Leu, and Ala-Ile. Our results agree well with the Gromiha and Selvaraj's work [22,23]. Those residue pairs have the higher tendency of forming long-range contacts. Without these long-range contacts, it is difficult to become globular proteins. On the other hand, the minimum is Gln-His contact, and the average number of long-range contacts P_{A-B} of Gln-His contact is 0.064, and Glu-Trp, Arg-Lys, Asp-Met, Asp-Trp, His-Met, His-Trp, Asp-Glu, Asp-Gln, Gln-Gln, Glu-Met, Gln-Glu, Glu-His, Gln-Met, Trp-Met contacts are all less than 0.1. It shows that Cys-Cys is the easiest one to form the long-range contact and Gln-His is the more difficulty to form the long-range contact. In other words, the Cys-Cys, Val-Val, Ile-Val, Val-Leu and Ala-Val contacts play an important role on the protein folding.

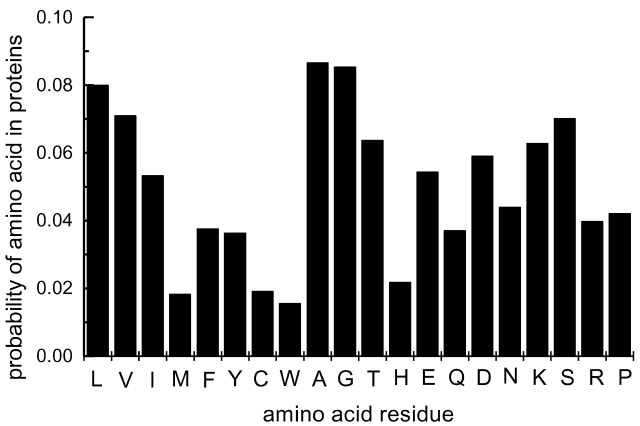

Fig. 1. The probability of different amino acid residues in all the 20 amino acid obtained from 278 globular proteins. Here L,V,…,R, and P represent amino acids in one letter symbol

We also discuss the short-range contacts, and the results are given in Table 3. The average number of short-range contacts P_{A-B} ranges from 0.058 to 0.487. Except Ala-Ala Leu-Leu, Ala-Leu, Lys-Ala, Ala-Val and Gly-Ala short-range contacts, most of them have a value 0.1–0.3. We think that forming a short-range contact depends mainly on the sequence of amino acids. If the amino acid has a large probability in all the 20 amino acids, it may have a possible large value of the short-range contacts.

We also calculate the average number P_{A-B} of residue–residue contacts for different amino acids with $R_c = 0.65$ nm, and the results are given in Table 4. We find that although the values of the number of long-range contacts and short-range contacts become small, the 10 topmost long-range contacts are the same as the case of $R_c = 0.80$ nm. Therefore, the relative ability to form a residue–residue contact does not depend on the value of R_c.

Considering the single residue, we calculate the average number of short-range contacts per residue C_S and the average number of long-range contacts per residue C_L according to Eq. (2), and the results are given in Table 5. Here we discussed the contact numbers of per residue for all the 20 amino acids with $R_c = 0.80$ and 0.65 nm. In $R_c = 0.80$ nm condition, we obtain that the Cys amino acid has the largest value of the average number of long-range contacts, and the Glu amino acid has the smallest one. We also find that the amino acid of Leu, Val, Ile, Met, Phe, Tyr, Cys, and Trp has a large value of C_L, and the amino acid of Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro has a small value of C_L. Our results agree well with previous calculations [10,11]. In our another paper, we concluded that all the 20 residues can be classified into three groups: hydrophobic residues (H) consisting of Phe, Met, Ile, Leu, Trp, Val, Cys, and Tyr; neutral residues (N) consisting of His, Ala, Gly, and Thr; and hydrophilic residues (P) consisting of Lys, Asp, Asn, Glu, Gln, Ser, Pro, and Arg [11]. If one is the hydrophobic residue, it must have a large value of the average number of long-range contacts per

Table 3
Average number P_{A-B} of residue–residue contacts for different amino acids A and B. Upper triangle counts the long-range contacts and lower triangle counts the short-range contacts in protein samples. Here $R_c = 0.80$ nm

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
	0.421	0.528	0.442	0.201	0.334	0.257	0.202	0.202	0.437	0.343	0.308	0.147	0.165	0.188	0.178	0.164	0.206	0.261	0.176	0.217	Leu
Leu	0.358	0.719	0.532	0.213	0.358	0.273	0.258	0.186	0.505	0.383	0.381	0.190	0.222	0.193	0.187	0.177	0.234	0.328	0.225	0.242	Val
Val	0.276	0.258	0.465	0.188	0.309	0.286	0.220	0.170	0.433	0.305	0.316	0.146	0.175	0.169	0.175	0.164	0.219	0.259	0.194	0.168	Ile
Ile	0.260	0.213	0.212	0.153	0.156	0.110	0.117	0.099	0.176	0.179	0.153	0.085	0.092	0.098	0.084	0.091	0.101	0.122	0.111	0.106	Met
Met	0.185	0.129	0.143	0.088	0.232	0.193	0.170	0.144	0.279	0.234	0.251	0.147	0.119	0.145	0.140	0.143	0.139	0.214	0.139	0.161	Phe
Phe	0.213	0.191	0.173	0.090	0.190	0.238	0.216	0.180	0.244	0.249	0.207	0.127	0.121	0.153	0.142	0.180	0.186	0.197	0.191	0.186	Tyr
Tyr	0.199	0.162	0.170	0.122	0.123	0.157	1.001	0.124	0.178	0.276	0.205	0.109	0.108	0.130	0.123	0.125	0.136	0.225	0.107	0.145	Cys
Cys	0.139	0.124	0.099	0.069	0.073	0.101	0.161	0.082	0.158	0.159	0.124	0.088	0.069	0.106	0.085	0.104	0.104	0.127	0.132	0.127	Trp
Trp	0.135	0.128	0.104	0.058	0.080	0.075	0.058	0.109	0.391	0.341	0.287	0.162	0.173	0.178	0.191	0.201	0.200	0.276	0.172	0.233	Ala
Ala	0.342	0.331	0.266	0.170	0.218	0.193	0.152	0.142	0.487	0.499	0.351	0.166	0.182	0.206	0.251	0.241	0.212	0.334	0.226	0.266	Gly
Gly	0.293	0.253	0.220	0.120	0.187	0.195	0.166	0.142	0.328	0.294	0.305	0.158	0.160	0.168	0.214	0.211	0.183	0.291	0.191	0.232	Thr
Thr	0.217	0.209	0.177	0.107	0.154	0.167	0.114	0.110	0.259	0.269	0.231	0.148	0.096	0.064	0.136	0.104	0.108	0.141	0.107	0.146	His
His	0.176	0.156	0.118	0.087	0.127	0.101	0.088	0.076	0.167	0.176	0.115	0.116	0.104	0.093	0.088	0.102	0.179	0.153	0.149	0.130	Glu
Glu	0.296	0.208	0.215	0.137	0.187	0.164	0.099	0.101	0.286	0.208	0.210	0.160	0.254	0.089	0.088	0.106	0.122	0.145	0.108	0.149	Gln
Gln	0.261	0.193	0.161	0.106	0.128	0.148	0.097	0.110	0.245	0.192	0.170	0.098	0.191	0.143	0.127	0.165	0.175	0.184	0.159	0.137	Asp
Asp	0.257	0.237	0.213	0.158	0.184	0.167	0.099	0.104	0.297	0.257	0.200	0.134	0.199	0.187	0.195	0.177	0.134	0.199	0.128	0.151	Asn
Asn	0.215	0.183	0.197	0.103	0.144	0.180	0.114	0.119	0.226	0.181	0.202	0.105	0.177	0.165	0.209	0.180	0.136	0.182	0.082	0.129	Lys
Lys	0.298	0.239	0.226	0.128	0.192	0.153	0.111	0.120	0.347	0.254	0.214	0.140	0.313	0.184	0.281	0.199	0.261	0.282	0.163	0.181	Ser
Ser	0.270	0.221	0.213	0.121	0.178	0.186	0.134	0.144	0.279	0.267	0.262	0.127	0.204	0.203	0.228	0.175	0.207	0.304	0.123	0.146	Arg
Arg	0.260	0.170	0.188	0.142	0.145	0.162	0.078	0.121	0.226	0.190	0.201	0.119	0.237	0.173	0.208	0.146	0.153	0.178	0.155	0.169	Pro
Pro	0.178	0.180	0.145	0.086	0.153	0.142	0.073	0.087	0.183	0.181	0.167	0.111	0.147	0.136	0.169	0.118	0.169	0.194	0.120	0.120	
	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	

Table 4
Average number P_{A-B} of residue–residue contacts for different amino acids A and B. Upper triangle counts the long-range contacts and lower triangle counts the short-range contracts in protein samples. Here $R_c = 0.65$ nm

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
	0.200	0.249	0.204	0.091	0.157	0.122	0.110	0.096	0.214	0.195	0.165	0.066	0.083	0.093	0.083	0.076	0.096	0.129	0.085	0.093	Leu
Leu	0.350	0.377	0.276	0.117	0.181	0.137	0.138	0.089	0.256	0.226	0.209	0.096	0.122	0.105	0.108	0.084	0.118	0.189	0.116	0.117	Val
Val	0.217	0.179	0.218	0.083	0.163	0.148	0.109	0.078	0.213	0.154	0.161	0.063	0.095	0.091	0.094	0.081	0.110	0.138	0.092	0.079	Ile
Ile	0.226	0.152	0.162	0.055	0.090	0.057	0.053	0.044	0.090	0.088	0.077	0.045	0.046	0.048	0.041	0.048	0.043	0.065	0.071	0.043	Met
Met	0.171	0.106	0.103	0.071	0.136	0.102	0.091	0.067	0.137	0.129	0.133	0.067	0.065	0.066	0.065	0.078	0.074	0.124	0.063	0.067	Phe
Phe	0.191	0.129	0.125	0.079	0.149	0.144	0.127	0.093	0.128	0.139	0.111	0.051	0.053	0.079	0.072	0.103	0.096	0.107	0.093	0.080	Tyr
Tyr	0.163	0.112	0.119	0.095	0.093	0.105	0.678	0.079	0.088	0.138	0.123	0.075	0.044	0.071	0.065	0.067	0.064	0.123	0.048	0.086	Cys
Cys	0.093	0.081	0.077	0.050	0.055	0.053	0.166	0.042	0.082	0.083	0.064	0.047	0.044	0.044	0.043	0.051	0.063	0.078	0.079	0.055	Trp
Trp	0.110	0.090	0.071	0.043	0.062	0.066	0.038	0.086	0.229	0.175	0.150	0.077	0.086	0.085	0.098	0.103	0.104	0.151	0.089	0.110	Ala
Ala	0.283	0.240	0.196	0.146	0.169	0.153	0.102	0.109	0.418	0.275	0.208	0.088	0.077	0.113	0.130	0.124	0.097	0.174	0.116	0.138	Gly
Gly	0.218	0.174	0.150	0.092	0.114	0.133	0.114	0.097	0.241	0.216	0.163	0.101	0.083	0.089	0.099	0.115	0.107	0.163	0.104	0.110	Thr
Thr	0.159	0.131	0.128	0.087	0.101	0.110	0.089	0.073	0.193	0.155	0.141	0.095	0.054	0.030	0.062	0.047	0.060	0.071	0.061	0.052	His
His	0.138	0.109	0.087	0.075	0.086	0.075	0.079	0.044	0.112	0.098	0.086	0.101	0.045	0.037	0.044	0.047	0.099	0.078	0.072	0.064	Glu
Glu	0.240	0.149	0.145	0.100	0.136	0.120	0.064	0.077	0.242	0.149	0.162	0.109	0.184	0.049	0.041	0.049	0.054	0.076	0.057	0.066	Gln
Gln	0.208	0.133	0.112	0.080	0.073	0.089	0.051	0.103	0.217	0.128	0.113	0.078	0.165	0.122	0.061	0.085	0.075	0.094	0.075	0.068	Asp
Asp	0.188	0.149	0.139	0.113	0.134	0.116	0.075	0.070	0.231	0.194	0.137	0.100	0.159	0.147	0.155	0.082	0.071	0.113	0.066	0.057	Asn
Asn	0.153	0.112	0.138	0.073	0.095	0.119	0.086	0.094	0.183	0.129	0.147	0.077	0.132	0.128	0.164	0.140	0.071	0.097	0.047	0.057	Lys
Lys	0.224	0.177	0.156	0.114	0.142	0.118	0.070	0.092	0.261	0.177	0.156	0.098	0.275	0.150	0.244	0.146	0.206	0.153	0.091	0.057	Ser
Ser	0.191	0.161	0.133	0.084	0.123	0.118	0.081	0.093	0.213	0.177	0.155	0.101	0.153	0.153	0.169	0.136	0.144	0.182	0.061	0.057	Arg
Arg	0.207	0.137	0.132	0.098	0.101	0.116	0.054	0.081	0.188	0.131	0.142	0.095	0.197	0.142	0.174	0.109	0.112	0.120	0.132	0.057	Pro
Pro	0.123	0.110	0.088	0.061	0.090	0.074	0.042	0.060	0.133	0.085	0.094	0.068	0.101	0.099	0.086	0.069	0.126	0.096	0.072	0.057	
	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	

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 is the probability of the residue in all 20 amino acids, and $\bar{C}_S(\bar{C}_L)$ is the average of $C_S(C_L)$

			$R_c = 0.80$ nm				$R_c = 0.65$ nm			
20 Amino acids	Three types of amino acids	P (%)	C_S	$\bar{C}_S$	C_L	$\bar{C}_L$	C_S	$\bar{C}_S$	C_L	$\bar{C}_L$
Leu	Hydrophobic (H)	8.05	3.995	3.717	4.484	5.008	2.913	2.674	1.999	2.579
Val		7.07	3.510		5.533		2.433		2.877	
Ile		5.31	3.696		5.347		2.599		2.669	
Met		1.75	3.968		4.392		3.060		2.150	
Phe		3.75	3.718		4.726		2.653		2.460	
Tyr		3.62	3.667		4.637		2.562		2.441	
Cys		1.92	3.518		6.312		2.362		3.586	
Trp		1.56	3.870		4.632		2.806		2.446	
Ala	Neutral (N)	8.72	4.068	3.649	4.068	4.068	3.189	2.671	2.102	2.159
Gly		8.70	3.364		4.153		2.349		2.271	
Thr		6.58	3.311		4.126		2.299		2.310	
His		2.18	3.851		3.924		2.847		1.954	
Glu	Hydrophilic (P)	5.42	3.931	3.721	2.640	3.232	3.031	2.721	1.290	1.610
Gln		3.70	3.942		3.194		3.008		1.571	
Asp		5.89	3.770		2.858		2.813		1.401	
Asn		4.38	3.629		3.341		2.658		1.683	
Lys		6.26	3.881		2.895		2.938		1.452	
Ser		6.98	3.557		3.698		2.438		1.967	
Arg		3.97	3.858		3.435		2.897		1.753	
Pro		4.20	3.200		3.794		1.982		1.761	

residue. Which one is the hydrophobic residue or not depends on the average number of long-range contacts per residue. If there are not any long-range contacts, it cannot form the globular protein structure. Of course, the attraction interactions between atoms lead to form long-range residue–residue contacts. We also calculate the average number ($\bar{C}_L$ of long-range contacts per hydrophobic residue (H), or per neutral residue (N), or per hydrophilic residue (P). Here $\bar{C}_L$ is defined as

$$\bar{C}_L = \frac{\sum_{\alpha=\text{Leu,Val,...,Trp; or Ala,...His; or, Glu,Gln,...,Pro}} C_{L,\alpha}}{m}$$

$$\begin{pmatrix} \alpha = \text{Leu, Val, ...Trp} & m = 8 \\ \alpha = \text{Ala, Gly, ..., His} & m = 4 \\ \alpha = \text{Glu, Gln, ..., Pro} & m = 8 \end{pmatrix} \quad (3)$$

The results are given in Table 5 and Fig. 2. In Fig. 2, we find that $\bar{C}_L$ of hydrophobic residue, neutral residue, and hydrophilic residue is 5.008, 4.068, and 3.232, respectively. We also discuss the case of $R_c = 0.65$ nm, and similar results are obtained.

In order to study whether the long-range contact or the short-range contact plays an important in protein folding, we calculate the average number C_S of short-range contacts per residue. The results are given in Table 5. We observe that the Ala residue has the largest value of the short-range contacts, and the Pro residue has the smallest one. However, the difference of two values is small. The important result is that C_S ranges from 3.5 to 4.0 for 80% of all the 20 amino acids, and the average numbers of short-range contacts per residue of hydrophobic, neutral, and hydrophilic residues are almost the same. This means that hydrophobic residue plays an equally important role in forming short-range contacts, which is different from the forming of long-range contacts. When R_c decreases, the average numbers of short-range contacts and long-range contacts per residue also decrease, see Table 5. However, the residues with the maximum and the minimum of C_L and C_S are the same, and relative position of all the 20 amino acids unchange (see Table 5). It is worthy of considering the average of C_L and C_S for three types of amino acids, i.e. hydrophobic, neutral, and hydrophilic residues. The results are shown in Fig. 2. The obvious distinguish between $\bar{C}_L$ and $\bar{C}_S$ is that $\bar{C}_L$ has

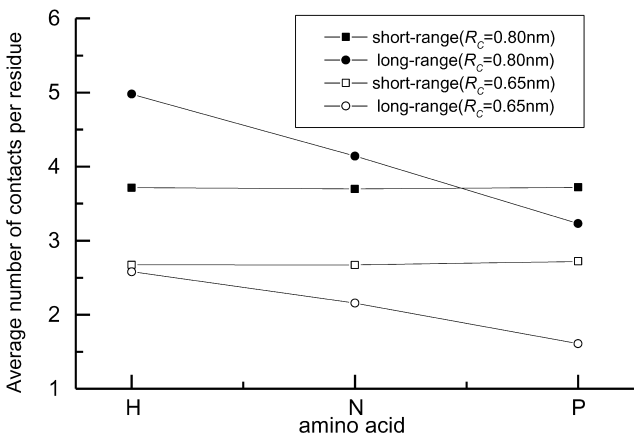

Fig. 2. Average number of contacts per residue. Here H, N, and P represent hydrophobic, neutral, and hydrophilic residues, respectively. $\bar{C}_S(\bar{C}_L)$ is the average number of short-range (long-range) contacts per residue. R_c is the limiting value of $C^\alpha - C^\alpha$ atoms forming a contact.

Table 6
The average distance of residue–residue contacts with $R_c = 0.80$ nm in unit of average length of protein chain $\bar{N}$. Here $\bar{N} = 189.45$

Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
0.283	0.264	0.288	0.285	0.286	0.258	0.246	0.240	0.322	0.304	0.295	0.249	0.264	0.306	0.279	0.279	0.272	0.300	0.271	0.278	Leu
	0.255	0.254	0.261	0.278	0.250	0.265	0.224	0.287	0.283	0.261	0.304	0.237	0.233	0.295	0.263	0.229	0.279	0.266	0.280	Val
		0.284	0.314	0.287	0.243	0.249	0.268	0.326	0.291	0.279	0.287	0.230	0.234	0.259	0.255	0.270	0.280	0.294	0.274	Ile
			0.244	0.292	0.282	0.230	0.221	0.269	0.292	0.304	0.294	0.239	0.285	0.271	0.354	0.229	0.280	0.271	0.255	Met
				0.261	0.247	0.221	0.271	0.302	0.294	0.261	0.288	0.277	0.229	0.326	0.284	0.248	0.267	0.280	0.282	Phe
					0.256	0.171	0.247	0.279	0.253	0.244	0.236	0.268	0.235	0.227	0.228	0.256	0.207	0.252	0.233	Tyr
						0.179	0.245	0.264	0.281	0.227	0.183	0.250	0.195	0.256	0.196	0.240	0.239	0.303	0.225	Cys
							0.279	0.282	0.320	0.242	0.246	0.195	0.288	0.220	0.270	0.198	0.245	0.241	0.210	Trp
								0.325	0.332	0.300	0.270	0.283	0.273	0.280	0.283	0.273	0.321	0.319	0.286	Ala
									0.337	0.291	0.314	0.262	0.280	0.283	0.294	0.270	0.307	0.325	0.316	Gly
										0.252	0.278	0.291	0.259	0.304	0.235	0.227	0.278	0.274	0.343	Thr
											0.210	0.250	0.253	0.245	0.271	0.223	0.238	0.288	0.315	His
												0.254	0.262	0.287	0.259	0.208	0.267	0.293	0.310	Glu
													0.258	0.288	0.250	0.226	0.270	0.247	0.283	Gln
														0.352	0.246	0.264	0.269	0.313	0.316	Asp
															0.251	0.253	0.297	0.289	0.318	Asn
																0.230	0.254	0.282	0.296	Lys
																	0.258	0.308	0.291	Ser
																		0.380	0.327	Arg
																			0.286	Pro

the tendency to a decrease with the amino type from H to P, while the change of $\bar{C}_S$ without this disciplinarian. It clearly indicates that the H type of the residue has higher tendency to forming the long-range contacts. From the point of view of chemic quality, these residues have the higher tendency of forming hydrophobic clusters and disulfide bridges due to long-range contacts. But in the short-range interactions, the amino acid's effect mostly relates to the coordinate regardless what type they are. The value of $\bar{C}_L$ proves that dividing the amino acids into three types is valid [10,11].

3.3. The average distance of each amino acid

After we know what amino acid has the preference to effect the contacts, we consider the action range of the 20 amino acids. Here *distance* means residues interval between the contacts pairs. The results are shown in Table 6. Here the average distance of amino acid action range is in the unit of the average length of protein, $\bar{N}$. The reason is that the average distance of action range is relative and depends on the size of protein. The longest distance is 0.380 $\bar{N}$, which occurs on Arg-Arg residue–residue contact and the shortest one is 0.171 $\bar{N}$, which occurs on Cys-Tyr residue–residue contact. The average distance ranges from 0.25 $\bar{N}$ to 0.30 $\bar{N}$ in general, and the difference is small. This investigation may help to improve the secondary structure predictions and provide some insights into the protein folding.

Acknowledgements

This research was financially supported by NSFC (Nos. 29874012 and 20174036), Natural Science Foundation of Zhejiang Province (No. 10102) 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] Branden C, Tooze J. Introduction to protein structure. New York: Garland; 1991.
[2] Creighton TE. Proteins: structures and molecular properties. New York: Freeman; 1983.
[3] Fletterick RJ, Schroer T, Matela RJ. Molecular structure: macromolecular in three dimensions. Oxford: Blackwell Scientific; 1985.
[4] Kamtekar S, Schiffer JM, Xiongm H, Babik JM, Hecht MH. Science 1993;262:1680–5.
[5] Lau KF, Dill KA. Macromolecules 1989;22:3986–97.
[6] Dill KA, Bromberg S, Yue K, Fiebig KM, Yee KM, Thomas PD, Chan HS. Protein Sci 1995;4:561–602.
[7] Yue K, Fiebig KM, Thomas PD, Chan HS, Shackhnovich EI, Dill KA. Proc Natl Acad Sci USA 1995;92:325–9.
[8] Regan L, DeGrado WF. Science 1988;241:976–9.
[9] Miyazawa S, Jernigan RL. J Mol Biol 1996;256:623–44.
[10] Miyazawa S, Jernigan RL. Macromolecules 1985;18:543–52.
[11] Linxi Z, Delu Z. Submitted for publication.
[12] Crippen GM. Biopolymers 1977;16:2189–96.
[13] Miyazawa S, Jernigan RL. Biopolymers 1982;21:1333–63.
[14] Jernigan RL, Miyazawa S. Biopolymers 1983;22:79–85.
[15] Manavalan P, Ponnuswamy PK. Arch Biochem Biophys 1977;184:476–87.
[16] Manavalan P, Ponnuswamy PK. Nature 1978;275:673–4.
[17] Ponnuswamy PK. Prog Biophys Mol Biol 1993;59:57–103.
[18] Gromiha MM, Selvaraj S. Recent Res Dev Biophys Chem 2000;1:1–14.
[19] Debe DA, Goddard WA. J Mol Biol 1999;294:619–27.
[20] Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Protein Eng 1999;12:549–55.
[21] Gromiha MM. Biophys Chem 2001;91:71–7.
[22] Gromiha MM, Selvaraj S. Biophys Chem 1999;77:49–68.
[23] Gromiha MM, Selvaraj S. Int J Biol Macromol 2001;29:25–34.
[24] Flory PJ. Statistical mechanics of chain molecules. New York: Wiley; 1969.