Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Study on properties of residue-residue contacts in protein*

WANG Xiang-hong (王向红)†, KE Jian-hong (柯见洪), ZHEN Yi-zhuang (郑亦庄),
CHEN Ai (陈 爱), XU Yin-xiang (徐银香)

(*Department of Physics, Wenzhou Normal College, Wenzhou 325000, China*)

†E-mail: wangxh@wznc.zj.cn

Received Oct. 15, 2003; revision accepted Jan. 11, 2004

Abstract: Residue-residue contacts are very important in forming protein structure. In this work, we calculated the average probability of residue-residue contacts in 470 globular proteins and analyzed the distribution of contacts in the different interval of residues using Contacts of Structural Units (CSU) and Structural Classification (SCOP) software. It was found that the relationship between the average probability $\overline{P}_L$ and the residue distance L for four structural classes of proteins could be expressed as $\lg P_L=a+b\times L$, where a and b are coefficients. We also discussed the connection between two aspects of proteins which have equal array residue number and found that the distribution probability was stable (or unstable) if the proteins had the same (or different) compact (for example synthase) in the same structural class.

Key words: Globular protein, Structural class, Residue distance, Long- and short-range contact
Document code: A **CLC number:** O631

INTRODUCTION

Recognition of protein folding from amino acid sequence is a challenging task. The structure and stability of proteins from different fold are mainly dictated by inter-residue interactions. Information on the distribution of contacts shown by 20 amino acid residues can be used to predict the folding type of each protein (Chan and Dill, 1998; Zhang and Kim, 2000; Gromiha, 2001; Gromiha and Selvaraj, 1999; Miyazawa and Jernigan, 1985; 1996; Bahar *et al*., 1997; McDonald and Thornton, 1994). It is very important to analyze the interaction of residue-residue contacts in investigations on the structure and stability of globular proteins.

During the last decades, a large number of biochemists, biophysicists and molecular biologists conducted many researches on protein sequence and structure based on residue-residue contacts (Gromiha, 2001; Gromiha and Selvaraj, 1999; Miyazawa and Jernigan, 1985; 1993; Sobolev *et al*., 1999; Russell and Barton, 1994; Fiser *et al*., 1997; Selbig and Argos, 1998). For instance, Miyazawa and Jernigan (1985) studied the interactions of amino acid contacts in protein molecules; Gromiha and Selvaraj (1999) investigated the importance of long-range interaction in the folding and stability of proteins, and so on. However, in previous works, the residues in a protein molecule were represented by their C_α atoms (Gromiha, 2001; Gromiha and Selvaraj, 1999; Miyazawa and Jernigan, 1985). Residues whose centers are close to R_c, are defined to be in contact. In fact, each of the amino acid residues has different structure. If the distance of the C_α atom of the residue is the

* Project supported by the National Natural Science Foundation of China (Nos. 29874012, 20174036, 20274040), and the Natural Science Foundation of Zhejiang Province (No. 10102), and the Science Technology Development Plan of Wenzhou City (No. S2002A014), China

same, the interactions between two residues may be different for different residues. For example, the side group of amino acid residue Gly is H and the side group of amino acid of Arg is $(CH_2)_3NHC(NH)NH_2$. In the meantime, there is much interaction in proteins, such as hydrogen bonds, hydrophobic-hydrophobic, aromatic-aromatic, aromatic-polar, etc. So it is not enough to only take into account the distance between C_α-C_α in forming contact. In order to determine which type of contact contributes most to stabilizing the structure, contacts of structural units (CSU) software and SCOP database were adopted for our calculation (Miyazawa and Jernigan, 1993; Russell and Barton, 1994). As amino acid residues, which contribute to contacts, need not be sequential neighbors and can exit far from the sequence, it is of interest to reveal the distribution of residue-residue contacts in the different interval of residues and study the statistical properties of residue-residue contacts in protein molecules.

CALCULATION METHOD

Long- and short-range contacts

Crystallographic data on 470 proteins were taken from the Protein Data Bank (PDB) of Brookhaven National Bank (Bernstein *et al*., 1977; Berman *et al.*, 2000). The PDB codes of the protein samples used in the present study are given in Table 1. We obtained the information on the structural class and the contacts of all proteins from SCOP and CSU databases (Hubbard *et al*., 1999; Sobolev *et al*., 1999) and calculated and analyzed the distribution of amino acid contacts in the 470 protein molecules.

When discussing residue-residue contacts, we takes into consideration (1) the closest distance between their atoms; (2) the solvent-accessible surface of every atom; (3) the hydrogen bonds; (4) the hydrophobic-hydrophobic interaction; (5) the hydrophobic-hydrophobilic interaction, etc. (McDonald and Thornton, 1994; Miyazawa and Jernigan, 1993; Russell and Barton, 1994; Fiser *et al*., 1997; Selbig and Argos, 1998). Using CSU software, we discuss the relationship between distribution of residue-residue contacts and their positions in protein chains. For a given i residue, the composition of surrounding residues j (i, j=1,2,3, …) was analyzed in terms of the location at the sequence level and the contributions from ≤±4 residues (or $L=|j-i|<4$) were treated as short-range contacts, ≥±4 residues (or $L=|j-i|>4$) as long-range contacts (Selbig and Argos, 1998; Gromiha and Selvaraj, 1997).

In the earlier studies (Zhang and Kim, 2000; Gromiha, 2001; Gromiha and Selvaraj, 1999; Miyazawa and Jernigan, 1985; 1996; Bahar *et al*., 1997; McDonald and Thornton, 1994; Fiser *et al*., 1997), the residue intervals for long-range contacts were only classified into 6 intervals (1–3; 4–10; 11–20; 21–30; 3–40; 4–50; >50). If the number of residues of proteins is large, the percentage (or the number) of long-range contacts in the interval of >50 may be large, and the percentage (or the number) of long-range contacts in the other intervals (<50) becomes small. In the same structural class of proteins, the number of residues for different proteins may be different, and this leads to the possibility that the average percentage of long-range contacts in the interval may be incorrect. The long-range contacts (≥±4 residues) are classified into several intervals with a step of 10 in the range of 4–100, and with a step of 20 (or 50) in the range of 100–500 (201–250; 251–300; …; 451–500; >500), respectively.

Distribution probability of contacts in each interval

To know in detail the distribution of the contacts in different residue distances, we define the probability of the contacts of kth interval in all contacts P_L as (Wang *et al*., 2004)

$$P_L = \frac{N_k}{N} \tag{1}$$

Here N_k and N are the number contacts in the kth intervals and the total number of contacts, respectively, and k represents residue intervals, namely 1–3 (short-range contacts); 4–10, 11–20, 21–30, 31–40,…, 91–100; 101–120, 121–140,…, 181–200;

Table 1 The PDB code and No. of protein samples used in this paper

α class proteins				
1:1ROP-A	2:1RPR-A	3:1RPR-B	4:1RPO	5:1AHD-P
6:1HOM	7:1UTG	8:1C5A	9:3ICB	10:1ACI
11:1ADR	12:2BCA	13:2BCB	14:4ICB	15:1KSM-A
16:2DVH	17:2PAC	18:351C	19:451C	20:1AAB
21:1CC5	22:1LEA	23:1LEB	24:1ACA	25:2ABD
26:1AVS-A	27:1AFH	28:256B-A	29:256B-B	30:1WRP-R
31:1CIF	32:1RRO	33:1YCC	34:1CCR	35:2C2C
36:3C2C	37:2HMQ-A	38:2HMQ-B	39:2HMQ-C	40:2HMQ-D
41:2MHR	42:1BP2	43:4P2P	44:2CCY-A	45:2CCY-B
46:4BP2	47:1BBH-A	48:1BBH-B	49:1ABV	50:155C
51:2HCO-A	52:2HHB-A	53:2HHB-C	54:2ASR	55:1BAB-A
56:1BAB-C	57:1BAB-B	58:1BAB-D	59:2HCO-B	60:2HHB-B
61:2HHB-D	62:1HBG	63:2FAL	64:2FAM	65:2HBG
66:4CLN	67:2LHB	68:1AJH	69:2CMM	70:2MB5
71:2MYA	72:4MBN	73:1ABS	74:1F63-A	75:1IFA
76:1AEP	77:1CPC-A	78:1CPC-K	79:2LIG-A	80:2LIG-B
81:1CPC-B	82:1CPC-L	83:1AEW	84:1FHA	85:1EA8-A
86:2ABK	87:1BJ9	88:1AEK	89:2CYP	90:1UCW-A
91:1UCW-B	92:1ALA			
β class proteins				
93:1BK2	94:1SHF-A	95:1SHF-B	96:1AEY	97:1HD3-A
98: 1ING-A	99:1ING-B	100:1INH-A	101:1INW	102:1INX
103:1INY	104:1IVC	105:1IVD	106: 1IVE	107:1IVF
108:1IVG	109:1F8E-A	110:2QWC	111:2QWG	112:7NN9
113:1A14-N	114:1F8C-A	115:1F8D-A	116:1L7F-A	117:1L7H-A
118:1NNC	119:1BOV-A	120:1BOV-B	121:1BOV-C	122:1BOV-D
123:1BOV-E	124:1C48-C	125:1C48-D	126:1C48-E	127:1C4Q-C
128:1C4Q-D	129:1C4Q-E	130:1CQF-C	131:1CQF-D	132:1CQF-E
133:1CZW-C	134:1CZW-D	135:1CZW-E	136:1CZW-H	137:2PKA-A
138:2PKA-X	139:1F53-A	140:1BBT-4	141:1FMD-4	142:1QQP-4
143:1GVP	144:1VQA	145:1VQB	146:1VQC	147:1VQD
148:1VQE	149:1VQF	150:1VQG	151:1VQH	152:1VQI
153:1YHB	154:2GN5	155:1TEN	156:2FNB-A	157:2CBP
158:1DAZ-C	159:1HIV-A	160:1HIV-B	161:1DZ5-A	162:1DZ5-B
163:1AAC	164:1AAN	165:1REI-A	166:1REI-B	167:1AC0
168:1ACX	169:1ACZ	170:1KUL	171:1KUM	172:1BPV
173:1YEA	174:2MCM	175:2BFV-L	176:1AKP	177:1CD8
178:2CY3	179:2BFV-H	180:1BW4	181:2PAB-A	182:2PAB-B
183:1BYN-A	184:2AVI-A	185:2AVI-B	186:2BFH	187:1A4A-A
188:1A4A-B	189:1AHK	190:2AZA-A	191:2AZA-B	192:1ACD
193:1ADL	194:2EIF	195:1BAR-A	196:1BAR-B	197:2FGF
198:2SNM	199:1COB-B	200:2SNV	201:2SOD-O	202:8I1B
203:2ILA	204:1STP	205:2BPA-2	206:1CID	207:1CDH
208:2LAL-A	209:2LAL-C	210:2SGA	211:1RBP	212:2BVV-A
213:1REE-A	214:1REE-B	215:2DLI-A	216:2ALP	217:1BBT-1
218:2AYH	219:1BBT-2	220:2HFT	221:1BBT-3	222:1SGT
223:5PTP	224:1TON	225:2RHV-3	226:1QNY-A	227:2CNA
228:2ENR	229:3CNA	230:1EST	231:3EST	232:2GCH

233:1CA2	234:1CNB	235:2CA2	236:1CAJ	237:2CAB
238:2RHV-2	239:1PYP	240:2RHV-1	241:2-Apr	242:3APR-E
243:4APR-E	244:1AVF-A	245:1AVF-J	246:1HTR-B	247:1BIL-A
248:1BIL-B	249:1HRN-A	250:1HRN-B	251:2REN	252:1NNB
253:1B9S-A	254:1B9T-A	255:1NSC-A	256:1NSC-B	
		$\alpha+\beta$ class proteins		
257:8TLN-E	258:1FJO-A	259:1FJQ-A	260:1FJT-A	261:1FJU-A
262:1FJV-A	263:1FJW-A	264:1HYT	265:1JMF	266:1JMG
267:1JMH	268:1JMI	269:1L3F-E	270:1LNA-E	271:1LNB-E
272:1LNC-E	273:1LND-E	274:1LNE-E	275:1LNF-E	276:1QF0-A
277:1QF1-A	278:1QF2-A	279:1THL	280:1TLI-A	281:1TLP-Ev
282:1TLX-A	283:1TMN-E	284:1TVU	285:2TDM	286:2TLI-A
287:2TMN-E	288:3TLI-A	289:3TMN-E	290:4TLI-A	291:4TLN
292:4TMN-E	293:4TMS	294:5TLI-A	295:5TLN	296:5TMN-E
297:6TLI-A	298:6TMN-E	299:7TLI-A	300:7TLN	301:1FJ3-A
302:1DUR-A	303:1FCL-A	304:1GB4	305:1F2G	306:1FXD
307:1IGD	308:1QE6-A	309:1QE6-B	310:1QE6-C	311:1QE6-D
312:2IL8-A	313:2IL8-B	314:3IL8	315:1F04-A	316:1B5M
317:1EUE-A	318:1EUE-B	319:1FXI-A	320:1RGE-A	321:1RGE-B
322:1RGF-A	323:RGF-B	324:1RGG-A	325:1RGG-B	326:1RGH-A
327:1RGH-B	328:1RSN-A	329:2SAR-A	330:2SAR-B	331:1ROE
332:1URN-A	333:2CJN	334:2CJO	335:1APS	336:1IET
337:1IEU	338:FXC	339:4FXC	340:1BKF	341:1FKB
342:1FKD	343:1FKF	344:1FRH	345:1ALC	346:1AQP
347:2RNS	348:3RN3	349:7RSA	350:1ACF	351:2PRF
352:2LYM	353:2LYZ	354:3LYZ	355:1LHI	356:2BQM
357:1E3V-A	358:1E3V-B	359:125L	360:190L	361:1D9W-A
362:2LZM	363:4LZM	364:1JTM-A	365:1AVP-A	366:1PPN
367:2PAD	368:9PAP	369:1AIM	370:2ACT	371:1EML
372:2EMD	373:2EMN	374:1CNS-A	375:1CNS-B	376:2BAA
377:1DNK-A	378:2TSC-A	379:2TSC-B	380:1EZM	381:1JMG
382:1QF0-A	383:1THL	384:4TMS		
		α/β class proteins		
385:1AAZ-A	386:1AAZ-B	387:1ABA	388:1DE1-A	389:1DE2-A
390:1AIU	391:1SRX	392:2TRX-A	393:2TRX-B	394:1C4W-A
395:2FOX	396:5NLL	397:1AKT	398:1AZL	399:1BU5-A
400:1BU5-B	401:2FX2	402:1FX1	403:2RN2	404:1GV8
405:1A5V	406:1ASU	407:1CXQ-A	408:3DFR	409:1OFV
410:1Q21	411:2FCR	412:1EX7-A	413:1GKY	414:1AJE
415:3ADK	416:1CLA	417:2CLA	418:4CLA	419:1NFP
420:1DHR	421:3PGM	422:1TPF-A	423:1TPF-B	424:1TRE-A
425:1TRE-B	426:1BKS-A	427:1BKS-A	428:2DRI	429:1HVQ
430:2SBT	431:1THM	432:2PRK	433:1ULA	434:1RHD
435:1CTT	436:1TFD	437:1GYM	438:1ABE	439:1ABF
440:5ABP	441:3CPA	442:5CPA	443:2GBP	444:1SBP
445:2HAD	446:1DOR-A	447:1DOR-B	448:1AZW-A	449:1AZW-B
450:2ACR	451:1AAX	452:1BSL-A	453:1BSL-B	454:1LUC-B
455:1XEL	456:2LIV	457:1IPD	458:1ADD	459:1LUC-A
460:1DOS-A	461:1DOS-B	462:1ALD	463:1GOX	464:1ETU
465:1HPM	466:6XIA	467:1NOY-A	468:1NOY-B	469:1BKS-B
470:1BKS-B				

201–250, 251–300;…; 451–500; >500 (long-range contacts).

We also define the average probability of the contacts in the kth interval $\overline{P_L}$ as

$$\overline{P_L} = \frac{\sum_{n=1}^{M} P_{1,n}}{M} \quad (2)$$

Here M represents the total protein number, and being 470 in this paper. We calculated the average probability of the contacts in the interval of 10 residues $\overline{P_L}$ (i.e. 11–50, 51–100, 101–200, >200). For instance, if the probability of the contacts in the range of 11–50 P_L is 22.0%, the average probability of contact of each 10 residue distances $\overline{P_L}$ in the same range 11–50 will be 5.5% (= 22.0%/4.0) (the same is true of the calculation in Figs.2, 3, 4, 5). We also compared the average probabilities of contacts for four structural classes of proteins.

We classified the protein into three groups based on their size. Proteins with less than 100 residues (S<100) are considered small proteins S_1; with residues between 100 and 200 (100<S<200) are considered to be medium proteins S_2 and with more than 200 residues (S>200) are considered large proteins S_3. S is the total number of residues for each protein molecule. S_1, S_2 and S_3 represent three groups of proteins, respectively (Miyazawa and Jernigan, 1985). The average probability of contacts were calculated and compared between these groups of protein molecules.

RESULTS AND DISCUSSIONS

Average probability of contacts per protein in different intervals

We calculated the average probability of contacts $\overline{P_L}$ per protein in different residue intervals. Fig.1 illustrates the contact probabilities of protein molecules P_L in four structural classes. Fig.1a shows the probability distribution of long-range contacts (L≥4); Fig.1b the probability distribution of short-range contacts (1–3 residues distance); Fig.1c the probability distribution of contacts in the range of 4–10 and 11–50. Fig.1 shows that the average value of $\overline{P_L}$ for all-α class of protein molecule differs rather greatly from that for all-β class of protein molecule in the different residue intervals. Fig.1a shows that the average probabilities of long-range contacts all have larger values than those of short-range contacts for each protein structure. The average probabilities of long-range contacts have the lowest average value 53.21% for all-α protein molecule, and the highest average value 60.99% for all-β class. Contrary to the average probability of long-range contacts, short-range contacts for all-α class of proteins was the largest one. Fig.1c shows that the average value of $\overline{P_L}$ for all-α class decreases from 22.2% in the range of 4–10 to 4.08% in the range of 11–50, while the average value of $\overline{P_L}$ for all-β class decreases slowly from 13.1% to 7.9%. In the meantime, the average values of $\overline{P_L}$ for all-α class was much greater than that for all-β class in the range of 4–10, but $\overline{P_L}$ for all-α class was lower than that for all-β class in the range of 11–50. The capacity of amino acid forming contact was different for different type protein.

Fig.2 also shows that in the range of 4–10, the average value of $\overline{P_L}$ for all-α class proteins (23.2%) was nearly twice as many as that for all-β class proteins (13.0%), and that the situation of α+β and α/β proteins was almost the same, both being close to 16.4%. Nevertheless, in the range of 11–50, the all-α class of proteins had the smaller value of 4.0% and the all-β class of protein had the larger value of 7.9%. It is important to know that the average probability of contacts in each interval varied with the residue interval.

Above all, through studying all sample proteins, we can get the conclusion that the bigger the residue distance is, the smaller the average value of $\overline{P_L}$ is, and the less difference there is for the average probabilities between four structural classes of proteins. For instance, in the range of 51–100, the average probability of all-α class

contacts is 2.18%, and that of $\alpha+\beta$ class is 1.94%. These results can help us know the globular protein structure. For example, there are different hydrogen-bonding patterns between all-α and all-β. α-helix is formed between i and i+4 residues (short-range contacts), and the β-strands between distant residues, especially in parallel β-strands (long-range contacts) (Gromiha and Selvaraj, 1997; 1998).

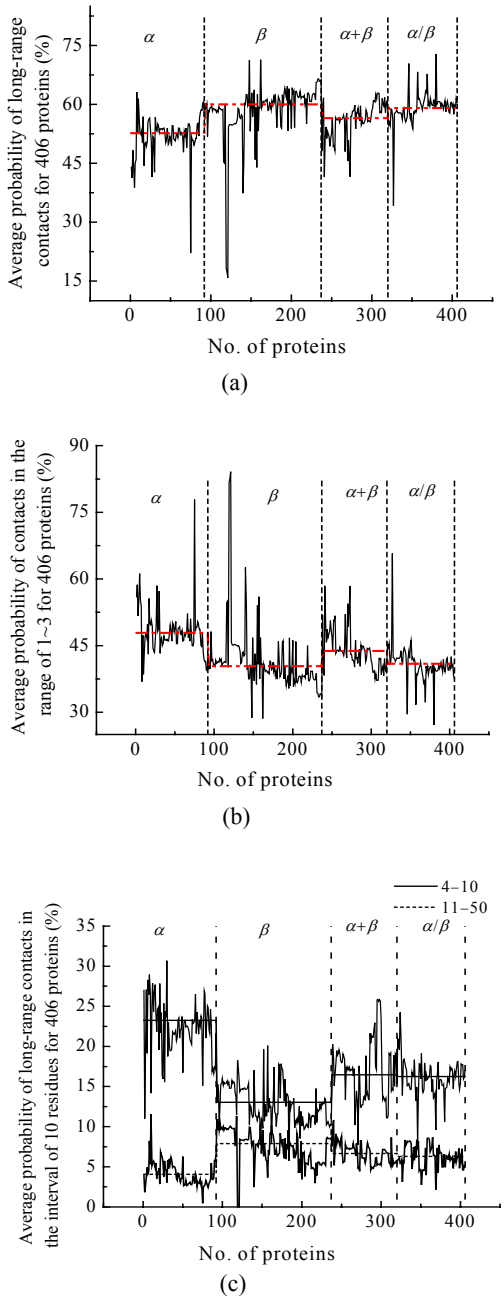


Fig.1 Probability distribution of (a) long-range contacts, (b) short-range contacts, and (c) contacts in the range of 4–10, and 11–50

Distribution of long-range contacts of proteins with the same number of residues

In order to study the relationship between the statistical property of residue-residue contacts and residue distances, two groups of proteins with the same sequence length were chosen from Table 1. One group was made up of 45 protein molecules whose total number of residues was 316. The number of proteins in Table 1 is from 258 to 301. Another group was comprised of 21 protein molecules with the same number of residues (388).

It was found that for the protein molecule with the same length of amino acids sequence and different structural classes, the distribution of the long-range contacts conforms to an obvious law in different residue distances. Fig.3 shows the probability distribution of the amino acid contacts in 4 interval (4–10; 11–50; 51–100; 101–200). Fig.3a is the probability distribution of proteins with the total number (316) of residues and different component (synthases); Fig.3b is the probability distribution of proteins with residue numbers (388) and the same component (synthases).

Fig.3b shows that the probability distribution of amino acid contacts in the range of 4–10 and 51–100 is clearly divided into mth section and nth section. This reason perhaps was that protein molecules in these two sections contained different component (synthases). The protein molecules in mth section contained compound thermolysin, and

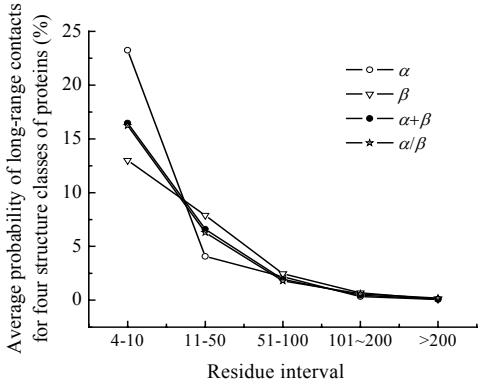
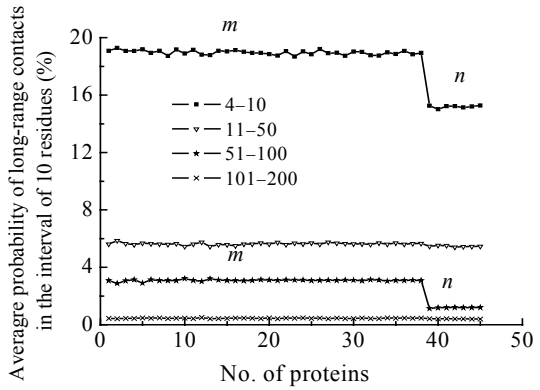


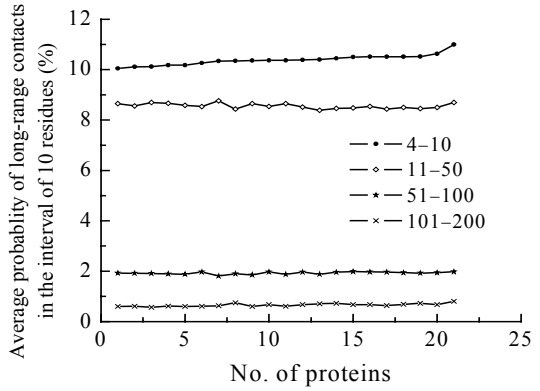
Fig.2 Average probability of long-range contacts for four structural classes proteins

those in *n*th section contained compound thymidylate synthase. In the range of 4–10, the average probability of the *m*th section contacts was 18.97%, while that of *n*th section contacts was 15.19%. Fig.3b shows the distribution of the contacts of protein molecules containing the same component (compound neuraminidase), being a distribution line with small ups and downs. For example, the probability is close to 10.39% in the range of 4–10. This suggests that the distribution of contacts had something to do with the array residue number as well as with the synthase that the protein molecules contained.

Therefore, it is not enough to analyze the distribution of contacts only in light of the sequence length and different structural classes protein molecules. The synthase, which protein molecules themselves contained, cannot be neglected with regard to the distribution of contacts due to the sythase's importance to the formation and stability of protein molecules



(a)



(b)

Fig.3 Probability distribution of the proteins (a) whose residue number is 316 and have different synthase with compound thermolysin and compound thymidylate synthase, and (b) whose residue number is 388 and have same synthase with compound neuraminidase

Average probability of long-range contacts in different residue distances

To know in more detail the distribution of contacts, we investigated the distribution of $\overline{P_L}$ in the interval of 10 residues for all sample proteins. Fig.4 shows the average probability of the long-range contacts of all the sample proteins in different interval. Here the average probabilities of long-range contacts were calculated in the intervals of 10 residues. The calculation showed that about 57% of the total contacts have formed long-range contacts. Fig.4 shows that about 15.9% of long-range contacts was in the interval of 4–10, and that about 1.5% of the long-range contacts was in the range of >200. It is obvious that the average probability of long-range contacts decrease with increase of reside distance.

Analyzing of the average probability distribution of the long-range contacts in the different structural classes of protein molecule, we found the existence of a relationship between the average probability of the long-range contacts and the residue distances. Fig.5a–5d are plots of the average probability of residue-residue contact $\overline{P_L}$ in the

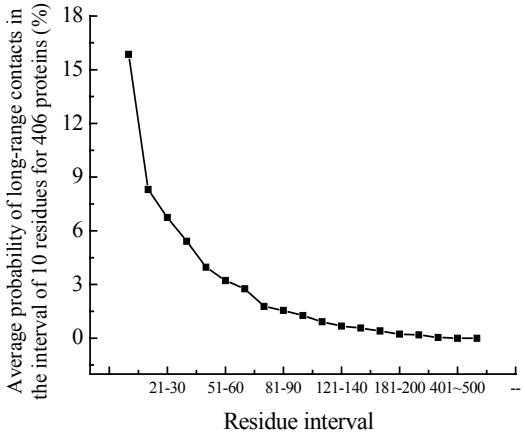


Fig.4 Average probability of contacts in the interval of 10 in all proteins

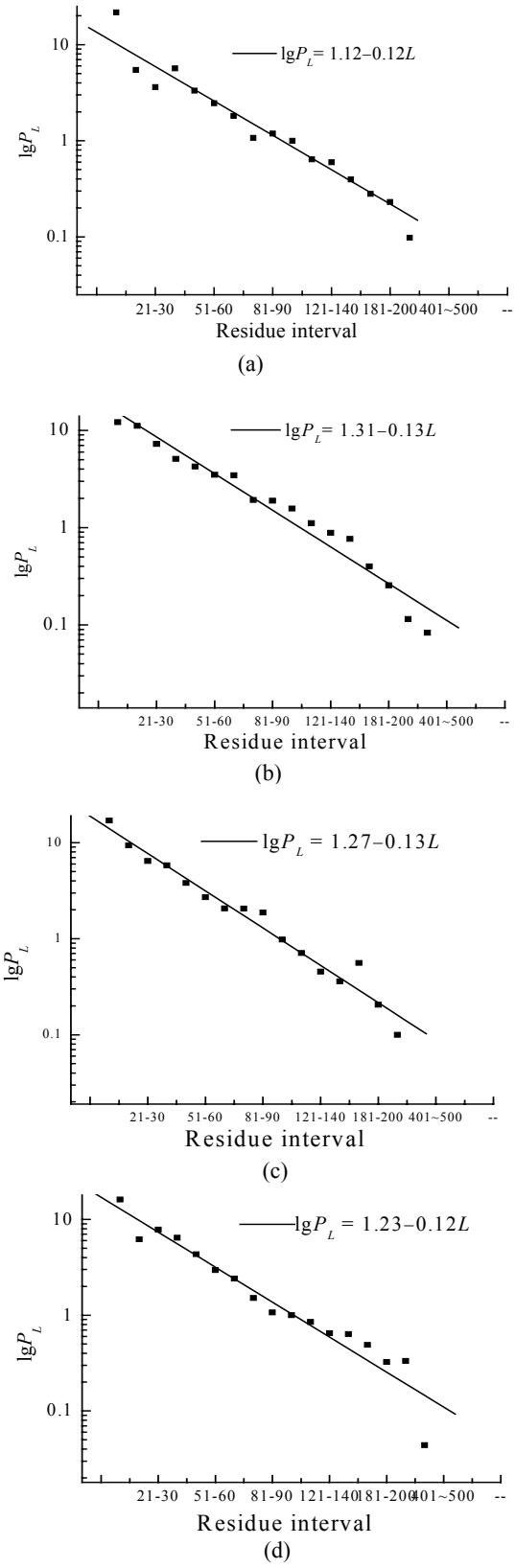


Fig.5 lg $\overline{P}_L$ –L for (a) All-α proteins; (b) All-β proteins; (c) α+β proteins; (d) α/β proteins

intervals of 10 for four structural proteins vs residue intervals L. The relationship between the average probability of the long-range contacts $\overline{P}_L$ and the residue distances L can be expresses approximately as

$$\lg P_L = a + b \times L \tag{3}$$

Here a and b are coefficient, respectively. Fig.5 show that the value of a and b are different in different structure. For all-α protein and α/β protein molecules b have the same value (0.12), and for all-β and α+β protein molecules have the value (0.13). The relative deviation between the average probability of long-range contacts and our expression for $\overline{P}_L$ is very small.

Influence of the residue numbers of proteins

Fig.6 on the average probability of long- and short-range contacts shows that for big protein molecules with long array it is 60.92%; that for moderate protein molecules it is 56.25%; that for small protein molecules it is 54.45%. The conclusion is that, the bigger protein array residue number is, the more important role the long-range action plays, and vice versa for short-range action. Thus, both protein molecule array number and structural classes of protein molecules exert some in fluency on long-range and short-range contacts, and in turn have influence on the formation and stability of protein molecules.

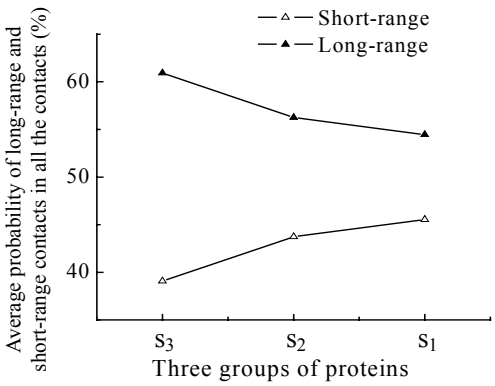


Fig.6 Average probability of long and short-range contacts for three groups of proteins

SUMMARY

The protein molecule is a stable structure formed under the combined influence of long-range and short-range actions. Study of the protein molecular structure and its function will be helpful for research on the thermodynamic nature of residue-residue contacts between amino-acids in protein molecules with respect to the distribution of contacts average probabilities, the probability distribution of residues forming contacts and so on, and on the influence that different structures of protein molecules and different array residue numbers have on the stability and folding action of protein molecules.

References

Bahar, I., Kaplan, M., Jernigan, R.L., 1997. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins*, **29**:292-308.

Berman, H.M., Westbrook, J., Feng, Z., 2000. The protein data bank. *Nucleic Acids Res.*, **28**:235-242.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, Q., Shimanouchi, T., Tasumi, M., 1977. The protein data bank: a computer-based archival file for molecular structures. *J Mol. Biol.*, **112**:535-542.

Chan, H.S., Dill, K.A., 1998. Compact polymers. *Macro-molecules*, **22**:4559-4573.

Fiser, A., Dosztanyi, Z., Simon, I., 1997. The role of long-range interactions in defining the secondary structure of proteins is overestimated. *Computer Applications in the Biosciences*, **13**:297-301.

Gromiha, M.M., Selvaraj, S., 1997. Influence of medium and long-range interactions in different structural classes of globular proteins. *J. Biol. Phys.*, **23**:151-162.

Gromiha, M.M., Selvaraj, S., 1998. Protein secondary structure prediction in different structural classes. *Prot. Eng.*, **11**:249-251.

Gromiha, M.M., Selvaraj, S., 1999. Important of long-range interactions in protein folding. *Biophysical Chemistry*, **77**:49-68.

Gromiha, M.M., 2001. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophysical Chemistry*, **91**:71-77.

Hubbard, T.J.P., Ailey, B., Brenner, S.E., Murzin A.G., Chothia, C., 1999. SCOP: A structural classification of proteins database. *Nucleic Acids Res.*, **27**:535-542.

McDonald, K., Thornton, J.M., 1994, Satisfying hydrogen bonding potential in proteins. *J Mol Biol.*, **238**:777-793.

Miyazawa, S., Jernigan, R.L., 1985. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**:543-552.

Miyazawa, S., Jernigan, R.L., 1993. A new subatitution matrix for protein sequence searches based *n* contact frequencies in protein structures. *Protein Eng.*, **6**:267-278.

Miyazawa, S., Jernigan, R.L., 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**:623-644.

Russell, R.B., Barton, G.J., 1994. Structural features can be umconservated in proteins with similar folds. *J Mol Biol.*, **244**:332-350.

Selbig, J., Argos, P., 1998. Relationships between protein sequence and structure patterns based on residue contacts. *Proteins*, **31**:172-185.

Sobolev, V., Sorokine, A., Prilusky, J., 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**:327-332.

Wang, X.H., Ke, J.H., Hu, M.X., 2004. Statistical properties of long-range contacts in globular proteins. *Chinese of Journal of Polymer Science*, **22**(4):187-194 (in Chinese).

Zhang, C., Kim, S.H., 2000. Environment-dependent residue contacts energies for proteins. *Proc. Natl. Acda. Sci.*, **97**(6):2550-2555.