Vol 18 No 10, October 2009
1674-1056/2009/18(10)/4580–11

Chinese Physics B

© 2009 Chin. Phys. Soc.
and IOP Publishing Ltd

Statistical interior properties of globular proteins*

Jiang Zhou-Ting(姜舟婷)[a)†], Zhang Lin-Xi(章林溪)[b)],
Sun Ting-Ting(孙婷婷)[c)], and Wu Tai-Quan(吴太权)[a)]

[a)] *Department of Applied Physics, China Jiliang University, Hangzhou* 310018, *China*
[b)] *Department of Physics, Zhejiang University, Hangzhou* 310027, *China*
[c)] *College of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou* 310018 *China*

(Received 31 July 2008; revised manuscript received 8 May 2009)

The character of forming long-range contacts affects the three-dimensional structure of globular proteins deeply. As the different ability to form long-range contacts between 20 types of amino acids and 4 categories of globular proteins, the statistical properties are thoroughly discussed in this paper. Two parameters N_C and N_D are defined to confine the valid residues in detail. The relationship between hydrophobicity scales and valid residue percentage of each amino acid is given in the present work and the linear functions are shown in our statistical results. It is concluded that the hydrophobicity scale defined by chemical derivatives of the amino acids and nonpolar phase of large unilamellar vesicle membranes is the most effective technique to characterise the hydrophobic behavior of amino acid residues. Meanwhile, residue percentage P_i and sequential residue length L_i of a certain protein i are calculated under different conditions. The statistical results show that the average value of P_i as well as L_i of all-α proteins has a minimum among these 4 classes of globular proteins, indicating that all-α proteins are hardly capable of forming long-range contacts one by one along their linear amino acid sequences. All-β proteins have a higher tendency to construct long-range contacts along their primary sequences related to the secondary configurations, i.e. parallel and anti-parallel configurations of β sheets. The investigation of the interior properties of globular proteins give us the connection between the three-dimensional structure and its primary sequence data or secondary configurations, and help us to understand the structure of protein and its folding process well.

Keywords: globular proteins, long-range contact, hydrophobicity scale
PACC: 8715

1. Introduction

A protein molecule composed of sequences of amino acids has a unique (native) three-dimensional (3D) structure in natural environment, where its shape specifies its biological functions. There are 20 different amino acids in a protein molecule, and their effects may be different in the folding process. The folding of a protein chain into a compact, unique 3D structure is directed and stabilized by intramolecular interactions between the constituent amino acid residues along the chain. From the viewpoint of polymer statistics, every polymer chain may occasionally return to itself. That is, some points of the free chain trajectory may come within a short reach of one another, forming a closed loop, also named contacts. So the amino acid residue–residue contacts play an important role in the protein folding process. It is also useful to predict the 3D structure of a globular protein from the knowledge of its amino acid sequence. When accurate multiple alignments of the members of these families can be made, the evolutionary information in the alignment can be used to identify which residues are of key importance to the folding process and function of the protein.

In the last half a century, researchers tried to explore the mechanism of how proteins fold into their native structures. Since the experiments of Anfinsen *et al*[1] into protein denaturation and renaturation, it has been assumed that all the information required to predict the 3D structure of a protein is inherent within the primary sequence of its amino acids. Despite efforts made by many researchers, no one has completely succeeded in predicting the protein structure based on detailed, all-atom energy functions to date.[2−4] By far the most accurate method of predicting the secondary (and tertiary) structure of a globular protein is by aligning the sequence to a homologue of known 3D structure,[5,6] and the structure-based activity prediction for an enzyme of unknown function

*Project supported by the National Natural Science Foundation of China (Grant No 10747160) and SRF for ROCS, SEM (Grant No [2008]890).
†Corresponding author. E-mail: z.jiang@cjlu.edu.cn, jiangzhouting@hotmail.com

has just published.[7] However, until now the number of 3D structures available in the central repository for this information: the Protein Data Bank (http://www.rcsb.org/pdb), is only a tiny fraction of the number of sequences known. With the rapid growth in protein sequence data available and the requirement to infer protein conformation and function from sequence, a pressing need of more accurate methods has developed for protein structure prediction.

Experiments and theoretical studies have shown that the amino acids are subdivided into three kinds of residues: 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 polar residues (P) consisting of Lys, Asp, Asn, Glu, Gln, Ser, Pro, and Arg.[8−11] There is an effective attraction between hydrophobic amino acids that arises from their aversion to the solvent and leads to such amino acids forming the core in the protein native state. In the HNP model, the hydrophobic energy, or the solvent effects, is a major contributor to the energies of protein folding. This simplified model is required for the study of the protein folding process and simplification can be made in both geometry and potential functions.[12]

In this paper, we proceed with counting long-range contacts between the residues and focus on the signification of the protein structure by statistical methods.[13] The location of the three kinds of residues, which as the composition of globular proteins, is the essential question we will discuss in detail. It might provide the information for a deep analysis on the structure and the function of a protein.

2. Statistical method of globular proteins

Globular proteins are grouped into four structural classes based on the content and topology of α-helices and β-strands in the 3D structures;[14] α-helices and β-strands have the different exhibition in the structure. Proteins can be categorized into four different structural classes, namely: all-α, all-β, α/β and $\alpha+\beta$ proteins. We have studied 611 globular protein structures, including 146 all-α, 151 all-β, 145 α/β and 169 $\alpha+\beta$ globular proteins. The coordinates of all globular protein structures are obtained from the Protein Date Bank (PDB). We obtain the information about the structural class from SCOP.[15] The PDB codes for the four types of the representative globular proteins used in the present study are given in Table 1.

Each residue in a protein molecule is represented by the position of α-carbon atom (C^α). If the distance between center C^α atoms of residues is shorter than R_C, these residues are defined to be in contact. This kind of simple method to evaluate residue–residue contacts in proteins has been widely used.[12,13,16−18] By 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 on the surrounding medium extends effectively only up to 0.80 nm.[19,20] The limit of 0.80 nm is sufficient to characterise the hydrophobic behavior of amino acid residues and to accommodate both the local and non-local interactions.[21] In the previous papers, the limit value 0.65 nm has been chosen to estimate the effective inter-residue contact energies by Miyazawa and Jernigan.[12,16] These limits also have been used to calculate the folding rate of protein,[22] protein stability upon mutations[23] and thermal stability of proteins.[24] In this paper, the limit values $R_C = 0.65$ nm and 0.80 nm are both attempted. For a given residue, the composition of surrounding residues is analyzed in terms of the location by its linear sequence, and the contributions within (or equal) 4 residues are treated as a short-range contact, and those beyond 4 residues are treated as a long-range one. This definition is the same as that for the short-range and long-range interactions in the rotational isomeric state (RIS) model.[24]

The attractive interactions between residues lead to the formation of contacts. Sometimes the interaction is so enormous that several residues along its linear sequence are influenced in the field to construct the contact sequentially. There may be several residues surrounding one within the limit value of R_C, so there is more than one contact for each residue. Furthermore, the number of the sequential residues with long-range contacts is needed to take into account, since it may be related to the specific folding length of a protein chain. Therefore, two parameters N_C and N_D must be defined to confine the valid residues. N_C represents the number of contacts, i.e., the number of surrounding residues located round the central one within the sphere of radius R_C. N_D denotes the number of sequential residues, each one has the long-range contacts no less than N_C.

Table 1. PDB codes of globular proteins.

α	1	110M	2	1A6G	3	1A6M	4	1A6N	5	1A7D	6	1ABS	7	1ACP	8	1AIL	9	1ALA	10	1AQE	11	1AVH-A	12	1AVH-B
	13	1B68-A	14	1BAB-A	15	1BAB-B	16	1BAB-C	17	1BAB-D	18	1BBL	19	1BEA	20	1BFA	21	1BP2	22	1BPQ	23	1BVC	24	1BZ6-A
	25	1BZR-A	26	1CCR	27	1CDP	28	1CEH	29	1CHI	30	1CIF	31	1CIH	32	1CIK-A	33	1CPC-A	34	1CPC-B	35	1CPC-K	36	1CPC-L
	37	1CRI	38	1CSU	39	1DO7-A	40	1EA8-A	41	1ECD	42	1ECN	43	1ECO	44	1F63-A	45	1FCS	46	1FDK	47	1FHA	48	1FIA-A
	49	1FIA-B	50	1FIP-A	51	1FIP-B	52	1FOW	53	1FOY	54	1GDI	55	1H96-A	56	1HBG	57	1HJT	58	1I55-A	59	1I55-B	60	1IFA
	61	1IRB	62	1JW8-A	63	1LE4	64	1LH1	65	1LH2	66	1LMB-3	67	1LMB-4	68	1LPE	69	1LRP	70	1LTW	71	1MBA	72	1MBC
	73	1MBD	74	1MBO	75	1MBS	76	1MCY	77	1MKT	78	1MLL	79	1MLO	80	1NFN	81	1NOO	82	1OMD	83	1PHC	84	1PP2-R
	85	1PP2-L	86	1PPA	87	1RAQ	88	1RCB	89	1RMI	90	1RRO	91	1SWM	92	1TRO-A	93	1TRO-C	94	1TRO-E	95	1TRO-G	96	1UNE
	97	1UTG	98	1YCC	99	1YEA	100	1YOI	101	1YTC	102	256B-A	103	256B-B	104	2BPP	105	2C2C	106	2CCY-A	107	2CCY-B	108	2CDV
	109	2CMM	110	2CTS	111	2CY3	112	2CYM	113	2CYP	114	2END	115	2FAL	116	2FAM	117	2GDM	118	2GMF-A	119	2GMF-B	120	2HBG
	121	2HCO-A	122	2HCO-B	123	2LHB	124	2MBW	125	2MGM	126	2MHB-A	127	2MHB-B	128	2MHR	129	2MYE	130	2PAC	131	2PDE	132	2SPM
	133	2WRP-R	134	3C2C	135	3CP4	136	3CYT-O	137	3CYT-I	138	451C	139	4BP2	140	4CPP	141	4CPV	142	4ICB	143	4MBN	144	5CPV
	145	5CYT-R	146	5TNC																				
β	147	1A18	148	1A2T	149	1A45	150	1AAC	151	1AAN	152	1AB0	153	1ABQ	154	1AC0	155	1ACD	156	1ACX	157	1ACZ	158	1ADL
	159	1ALB	160	1AR2	161	1BOV-A	162	1BOV-B	163	1BOV-C	164	1BOV-D	165	1BOV-E	166	1BRP	167	1BTY	168	1BXA-A	169	1BYH	170	1CA2
	171	1CAI	172	1CAM	173	1CD8	174	1CNW	175	1COB-A	176	1COB-B	177	1EAS	178	1ENR	179	1EST	180	1F3G	181	1F53-A	182	1GHJ
	183	1GLH	184	1GRP	185	1H4N	186	1H6X-A	187	1HHP	188	1HIV-A	189	1HIV-B	190	1HJC-A	191	1HJC-D	192	1HOE	193	1INF	194	1KAB
	195	1KL9-A	196	1KUL	197	1KXD	198	1LID	199	1-MAY	200	1MPP	201	1NES-E	202	1NN2	203	1NSB-A	204	1NSB-B	205	1P12-E	206	1PAZ
	207	1PYP	208	1QNY-A	209	1QWE-A	210	1RBP	211	1REI-A	212	1REI-B	213	1RZA	214	1RZB	215	1RZE	216	1SCS	217	1SGT	218	1SHF-A
	219	1SHF-B	220	1SNO	221	1STN	222	1STP	223	1STY	224	1TAL	225	1TEN	226	1TIE	227	1TLK	228	1TON	229	1TPP	230	1UGE
	231	1UPJ	232	1ZSC	233	2ALP	234	2-APR	235	2AVI-A	236	2AVI-B	237	2AYH	238	2AZA-A	239	2AZA-B	240	2BB2	241	2BPA-1	242	2BPA-2
	243	2CA2	244	2CAB	245	2CHA-1	246	2CHA-2	247	2ENR	248	2ER7-E	249	2GCH−1	250	2GCH−2	251	2GMT	252	2HVP	253	2ILA	254	2IZA
	255	2IZB	256	2IZJ	257	2LAL-A	258	2LAL-C	259	2LTN-A	260	2LTN-C	261	2MCM	262	2PAB-A	263	2PAB-B	264	2PCY	265	2RAC-A	266	2REN
	267	2RHE	268	2RSP-A	269	2RSP-B	270	2RTA	271	2SGA	272	2SNS	273	2SNV	274	2SOD-O	275	2SOD-Y	276	2SOD-B	277	2SOD-G	278	2STV
	279	2WEA	280	2WEB	281	3AIT	282	3APP	283	3CNA	284	3EST	285	3HVP	286	3PHV	287	3PTB	288	4CHA-A1	289	4CHA-A2	290	4CHA-B1
	291	4CHA-B2	292	4FAB-L	293	4FAB-H	294	4FGF	295	4PEP	296	4ULL	297	5PTP										

Table 1. (*continued*)

α/β	298	121P	299	1A2Q	300	1A5Y	301	1AA9	302	1AAM	303	1AAT	304	1AAX	305	1AB5-A	306	1AB5-B	307	1ABA
	308	1ABE	309	1ABF	310	1ABN	311	1ADD	312	1ADS	313	1AKQ	314	1AKR	315	1AMS	316	1APB	317	1ASB
	318	1AZ1	319	1AZL	320	1BAP	321	1BEE	322	1CHN	323	1CIA	324	1DHR	325	1DIR-A	326	1DIR-B	327	1DIR-C
	328	1DIR-D	329	1DIS	330	1DPB	331	1DRK	332	1E49-P	333	1E6K-A	334	1EAA	335	1EAE	336	1EAF	337	1EHC
	338	1ETU	339	1FKX	340	1FVX	341	1FX1	342	1GKY	343	1GNR	344	1GOA	345	1GOX	346	1HEY	347	1I52-A
	348	1INI-A	349	1IPD	350	1JAI	351	1KVA	352	1LAV	353	1LAW	354	1OFV	355	1OVB	356	1PEK-E	357	1PFK-A
	358	1PFK-B	359	1PLJ	360	1PLL	361	1PTY	362	1Q21	363	1RA9	364	1RBT	365	1RDD	366	1RF7	367	1RH3
	368	1RHD	369	1RNH	370	1RVE-A	371	1RVE-B	372	1S01	373	1SBP	374	1ST2	375	1SUP	376	1TFD	377	1THM
	378	1THO	379	1TIM-A	380	1TIM-B	381	1TPU-A	382	1TPU-B	383	1TRE-A	384	1TRE-B	385	1TYC	386	1UDR-A	387	1UDR-B
	388	1UDR-C	389	1UDR-D	390	1ULA	391	1URP-A	392	1URP-B	393	1URP-C	394	1URP-D	395	1YMV	396	2AAT	397	2ACS
	398	2CAN-A	399	2CAN-B	400	2CAN-C	401	2DHE	402	2DRI	403	2FCR	404	2FDX	405	2FLV	406	2FOX	407	2FX2
	408	2GBP	409	2HAD	410	2LIV	411	2PRK	412	2RN2	413	2SBT	414	2TIR	415	2TRX-A	416	2TRX-B	417	2TS1
	418	3ADK	419	3CHY	420	3CPA	421	3DFR	422	3FX2	423	3GBP	424	3PGK	425	3PGM	426	4CLA	427	4DFR-A
	428	4DFR-B	429	4FX2	430	4ICD	431	4NLL	432	4PFK	433	5ABP	434	5CPA	435	5NUL	436	5P21	437	5TIM-A
	438	5TIM-B	439	6XIA	440	8ABP	441	8DFR	442	9ABP										
$\alpha+\beta$	443	125L	444	190L	445	1AAR-A	446	1AAR-B	447	1ACF	448	1AEC	449	1AIQ-A	450	1AIQ-B	451	1AN5-A	452	1AN5-B
	453	1AOB	454	1AQP	455	1AXQ	456	1AXW-A	457	1AXW-B	458	1BKF	459	1BO8-A	460	1BP0-A	461	1BP4	462	1BPJ-A
	463	1BQI	464	1BU4	465	1BYQ-A	466	1CH0-A	467	1CH0-B	468	1CH0-C	469	1CTF	470	1CVZ-A	471	1CYO	472	1D3Z-A
	473	1D9W-A	474	1DAW-A	475	1DDU-A	476	1DDU-B	477	1DET	478	1DNA-A	479	1DNA-B	480	1DUR-A	481	1DYE	482	1E3V-A
	483	1E3V-B	484	1EZM	485	1F0Q-A	486	1F4C-A	487	1F4C-B	488	1FD2	489	1FDD	490	1FER	491	1FIK	492	1FKB
	493	1FKF	494	1FRH	495	1FRI	496	1GAO-A	497	1GAO-B	498	1GAO-C	499	1GAO-D	500	1GWD-A	501	1I0X-A	502	1I0X-B
	503	1I0X-C	504	1I0X-D	505	1I1Z-A	506	1I20-A	507	1ICW-A	508	1ICW-B	509	1IET	510	1IEU	511	1JG0-A	512	1JG0-B
	513	1JTU-A	514	1JTU-B	515	1JU6-A	516	1JU6-B	517	1JU6-C	518	1JU6-D	519	1JUJ-A	520	1JUJ-B	521	1JUJ-C	522	1JUJ-D
	523	1KCE-A	524	1KCE-B	525	1L00	526	1L3F-E	527	1L66	528	1L77	529	1LCA	530	1LHH	531	1LHI	532	1LRA
	533	1LYD	534	1LYH	535	1LZ1	536	1MSB-A	537	1MSB-B	538	1NJA	539	1NRC-A	540	1NRC-B	541	1PE6	542	1POP-A
	543	1PPN	544	1PPP	545	1PRQ	546	1QE6-A	547	1QE6-B	548	1QE6-C	549	1QE6-D	550	1RGA	551	1RGL	552	1RLS
	553	1RN1-A	554	1RN1-B	555	1RN1-C	556	1ROB	557	1SHA-A	558	1TBE-A	559	1TBE-B	560	1TDA	561	1TSY	562	1TVV
	563	1UBI	564	1WDB	565	1YER	566	1YET	567	1YVS	568	1ZPR-A	569	1ZPR-B	570	217L	571	2AAD-A	572	2AAD-B
	573	2AAE	574	2AAK	575	2ACG	576	2ACT	577	2BBQ-A	578	2BBQ-B	579	2HOH-A	580	2HOH-B	581	2HOH-C	582	2HOH-D
	583	2LYZ	584	2LZM	585	2MS2-A	586	2MS2-B	587	2MS2-C	588	2PAD	589	2PRF	590	2SAR-A	591	2SAR-B	592	2TDD
	593	2TDM	594	2TSC-A	595	2TSC-B	596	3BIR	597	3IL8	598	3LYZ	599	3RAT	600	3RN3	601	3SSI	602	4GSP
	603	4LZM	604	4TLN	605	4TMS	606	5FD1	607	5PAD	608	7RSA	609	8TLN-E	610	9PAP	611	9RNT		

The concept of a stem is next introduced. A stem, as a part of the protein chain, is composed of sequential amino acids with long-range contacts no less than N_C per residue. It takes the count of the valid residues once they have the long-range contacts no less than N_C, and also serves as the composition of a stem with the length no less than N_D. We have changed the values of N_C and N_D separately in order to discover the structural characters of the globular proteins. In the following discussion, we will find out that such a simple model is helpful to distinguish the category of globular proteins and show the structure information in detail.

3. Results and discussions

In our calculation, the total number of protein chains is 611, their PDB codes are listed in Table 1. The samples used are more than that given in the database in Gromiha and Selvaraj's work (150 proteins) and thus are sufficient for carrying out statistical studies. The characteristics of forming short- and long-range contacts for 20 amino acids are investigated first. Average number of contacts per residue can indicate the ability of forming contacts. Here the average number of short-range contacts per residue C_S and long-range ones C_L for 20 amino acid residues are defined as

$$C_{\alpha,\eta} = \frac{\sum\limits_{\beta=\mathrm{Ala,Asp,Cys,Glu,...,Tyr}} N_{\alpha-\beta,\eta}}{N_\alpha}, \quad (\eta = S \quad \text{or} \quad L; \quad \alpha = \mathrm{Ala, Asp, \ldots, Tyr}), \quad (1)$$

where $N_{\alpha-\beta}$ represents the number of the contacts between residues of types α and β, and N_α is the total number of type α residues in the 611 globular protein samples.

The probability of forming the short- and long-range contacts for 20 amino acid residues under the condition of N_C and N_D from 2 to 4 is calculated. The results are in good agreement with those in previous work where hydrophobic residues (H) have a higher tendency to form long-range contacts than hydrophilic residues (P).[13,25] Comparing the average value of C_L or C_S for the three types of amino acids, i.e. hydrophobic (H), neutral (N), and hydrophilic (P) residues, we can find the obvious difference between $\bar{C}_L$ and $\bar{C}_S$; $\bar{C}_L$ has the tendency of decreasing with the amino type from H to P, while such a decrease is not shown in $\bar{C}_S$ with the change of residue type. It is clear that the H type of the residue has a higher tendency to form the long-range contacts. Such a high tendency in H type residues indicates that the hydrophobic residues are almost located in the core of the globular protein, while hydrophilic residues are in the exterior of it. The visual frame of a 2D protein chain is shown in Fig.1. The globular protein is composed of solid and open dots, which denote respectively H and P residues. For instance, most solid dots have more than one long-range contacts per residue and open dots have only one or none. From Fig.1, it is clear that the solid dots are often located in the interior of the protein, so as to construct long-range contacts as many as possible. From the point of view of chemical quality, these residues have a higher tendency of forming hydrophobic clusters and disulfide bridges due to long-range contacts. While in the short-range interactions, the amino acid's effect is mostly related to the linear sequence of the protein regardless of what type they are in. From the conclusions in Refs.[13] and [25], we have already known that the long-range contacts play an important role on the structure of a protein and its folding process. So, in this article, we discuss mainly the long-range contacts.

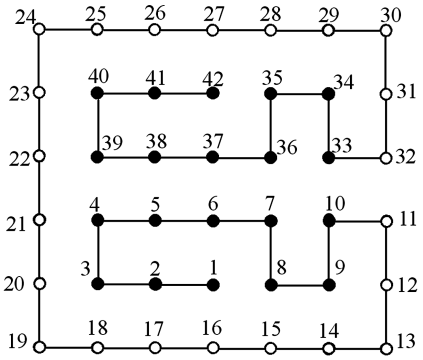


Fig.1. A visual frame of a 2D protein chain in lattice model. Here • represents interior residue and ○ represents exterior residue in a globular protein. Residue • always has 2 long-range contacts and residue ○ only has one or none contact.

3.1. The relationship between the effect of amino acids and hydrophobicity scales

Since the central importance is to understand what determines the stability and structure of proteins, many research groups have given numerous

amino acid hydrophobicity scales during the last few decades. Hydrophobicity is a term used to describe the unfavorable energetics of dissolving nonpolar (e.g. carbon) compounds in water. While a variety of nonpolar compounds readily dissolve in many nonpolar solvents, they have very low solubilities in water. Another characteristic of hydrophobic compounds is their tendency to self-associate in aqueous solutions. Hydrophobicity scales have been determined from free energies of transferring amino acids or their chemical derivatives from water to various nonpolar phases. In general, there has been a lack of quantitative agreement between hydrophobicity scales. The different nonpolar phases and different techniques may give different transfer free energies of amino acids. Here we choose three series of the hydrophobicity scales based on different chemical derivatives of the amino acids and the following different nonpolar phases: (a) alkyl chains in a reversed-phase liquid chromatography stationary phase,[26] (b) bulk octanol,[27] and (c) large unilamellar vesicle membranes.[28]

In this work, we further study the relationship between the ability in forming long-range contacts and the hydrophobicity scales of 20 types of amino acids in the 611 globular proteins. Once a stem is given with its length no less than N_D, then the residues as the components of it are all taken into account. The definitions of abscissa in Figs.2(a), 2(b) and 2(c) are 3 different hydrophobicity scales chosen from Refs.[26]–[28], respectively. The evidently linear relationships between the average percentage of forming long-range contacts $\bar{C}_{\alpha,L}$ for 20 residues and three series of the hydrophobicity scales are shown in Fig.2. The fitted functions are also marked in the figures. Comparing the correlation coefficients R, we find that the scale (c) is more effective to describe the ability of the amino acids in forming long-range contacts for the absolute value of R is closer to 1. From the three figures, it is found that the slope of the linear function is negative value between –1.5 to –2.5, and the maximum absolute value of slopes is that in Fig.2(c). The negative slope indicates that the hydrophobic residues have a higher tendency to form long-range contacts than hydrophilic ones. The average percentages of constructing long-range contacts per residue $\bar{C}_{\alpha,L}$ for 20 types of amino acids obtained from the statistical results are exactly

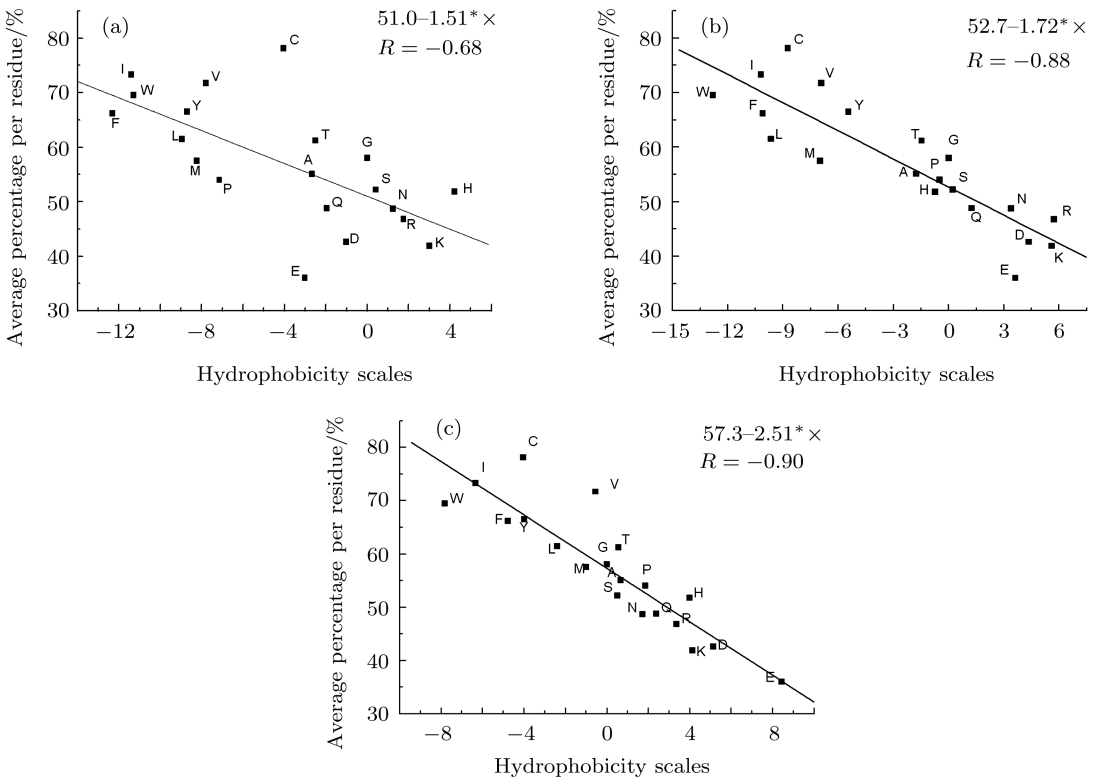


Fig.2. The average percentage per residue $\bar{C}_{\alpha,L}$ for 20 residues vs. hydrophobicity scales of amino acid residues by (a) DeVido *et al*,[26] (b) Fauchere and Pliska,[27] (c) Wimley and White.[28] Here $N_C = 3$, $N_D = 2$ and $R_C = 0.80$ nm.

the same in the three figures. So the difference between the slopes is due to the range of hydrophobicity scale of the three techniques. The maximum absolute value of slope in Fig.2(c) indicates the short range of scale value, but it is sufficient to distinguish the types of residue and the best linear relationship is obtained according to the value of correlation coefficient R.

3.2. The calculation for the different categories of globular proteins

The residue percentage to form long-range contacts of a certain globular protein i is defined as

$$P_i = \frac{\sum_{\alpha,\beta=\text{Ala,Asp,Cys,Glu,...,Tyr}} N_{\alpha-\beta,\eta}}{2N_i}, \quad (\eta = L;\ i = 1,2,\ldots,611), \tag{2}$$

where subscript i is the serial No. of the protein listed in Table 1, and N_i is the total number of amino acid residues of the protein i.

The definition of abscissa in Fig.3 represents the serial number of a protein given in Table 1. For example, the PDB code of No.1 is 110-M. The boundary condition has been changed, i.e., the values of N_C and N_D are attempted from 2 to 4. $N_C = N_D = 3$, $R_C = 0.65$ nm in Fig.3(a) and $N_C = 4$, $N_D = 2$, $R_C = 0.80$ nm in Fig.3(b) are chosen as examples. Figure 3(a) shows clearly that all-α proteins have a lower values than those of all-β proteins, indicating that all-β proteins have a higher tendency to construct long-range contacts under the condition of $N_C = N_D = 3$, $R_C = 0.65$ nm. Comparing the average residue percentage of each protein category, $\bar{P}_i$ of all-β proteins is the maximum (0.448) among these four types of globular proteins. It is about 10 times higher than the one of all-α proteins (0.035). Comparing the ability to form long-range contacts between α/β proteins and $\alpha+\beta$ proteins, the average values $\bar{P}_i$ are almost the same. It is due to both the effects of α-helices and β-strands conformation in proteins; such as the approximately alternation between α-helices and β-strands for α/β globular proteins, or the non-mixing of α-helices and β-strands but segregating into different domains in $\alpha+\beta$ globular proteins.

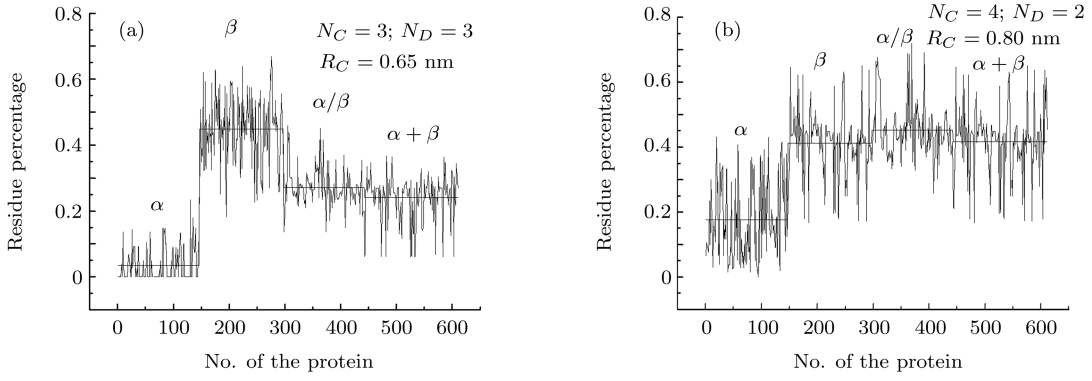


Fig.3. Distribution of the residues percentage P_i for all protein samples. Here No. of the protein is listed in Table 1, and the total number of proteins is 611.

In Fig.3(b), $N_C = 4$, $N_D = 2$ and $R_C = 0.80$ nm are chosen as the constraint condition to keep enough protein samples with valid residues in each category. The minimum value of $\bar{P}_i$ is also related to all-α proteins, and the values of all-β, α/β and $\alpha+\beta$ proteins are almost the same. Comparing Fig.3(a) with Fig.3(b), except for all-β proteins, the residue percentages of the other 3 types are increasing from 3(a) to 3(b). As the positive factors R_C and negative factor N_D lead to the increase of P_i. Figure 3(b) shows that the statistical results for all-β proteins are influenced mainly by the constraint N_C rather than N_D or R_C. One reason is that the distance between two residues constructing a long-range contact is in the range from 0.65 nm to 0.80 nm. The other one is related to the secondary configuration, i.e. parallel

and anti-parallel configuration of β sheets. Since the number of valid residues is increased with decreasing N_D (or N_C), N_D has the dominative influence on the residue percentage P_i of α/β, $\alpha+\beta$ and especially of all-α proteins. This means that these proteins are hardly to form long-range contacts one by one along their primary sequence. For all-β proteins, instead of N_D, N_C plays the main role on the statistical result of valid residues percentage P_i, which indicates the ability to form long-range contacts sequentially.

In addition to the residue percentages P_i of forming long-range contacts for different types of globular proteins, the number of sequential residues also has been calculated in the present work. Once the residues m till n of a given protein i all have the long-range contacts no less than N_C, the number of sequential residues, which also called sequential residue length L_i, is defined as

$$L_i = n - m. \tag{3}$$

Such a part of the protein can be treated as a stem, and several stems may be present in a protein chain. The sequential residue length represents the character of contacts aggregated along the primary structure. The statistical results of average value $\bar{L}_i$ for each protein category are shown in Fig.4, in which the conditions are the same as those in Fig.3, i.e., $N_C = N_D = 3$, $R_C = 0.65$ nm in Fig.4(a), and $N_C = 4$, $N_D = 2$, $R_C = 0.80$ nm in Fig.4(b). In Fig.4(a), the average value $\bar{L}_i$ of all-α proteins is a minimum 1.77 among the four types of globular proteins. It is even lower than the constraint $N_D = 3$, since a lot of all-α proteins have no valid residues with the long-range contacts more than N_C or no stem that is longer than N_D. The values of $\bar{L}_i$ for all-β, α/β and $\alpha+\beta$ proteins lie closely from 5.82 to 6.94. In 611 protein samples, the maximum of L_i 18.5 occurs at 2PAB-B classified as all-β protein, which is much higher than the constraint $N_D = 3$. The same tendency is also found in Fig.4(b). Except all-α proteins, the other three protein categories have the values of $\bar{L}_i$ close to one another. Focus on the variation of $\bar{L}_i$ from Figs.4(a) to 4(b), the values of all-α proteins are increased from 1.77 to 2.74, while those of the other three categories of proteins are decreased a little with decreasing N_D. Only one protein chain coded 1TRO-E has no valid residue fitting the boundary condition in Fig.4(b). The results also show that N_D has the dominating influence on all-α proteins, which have the low ability to form long-range contacts sequentially.

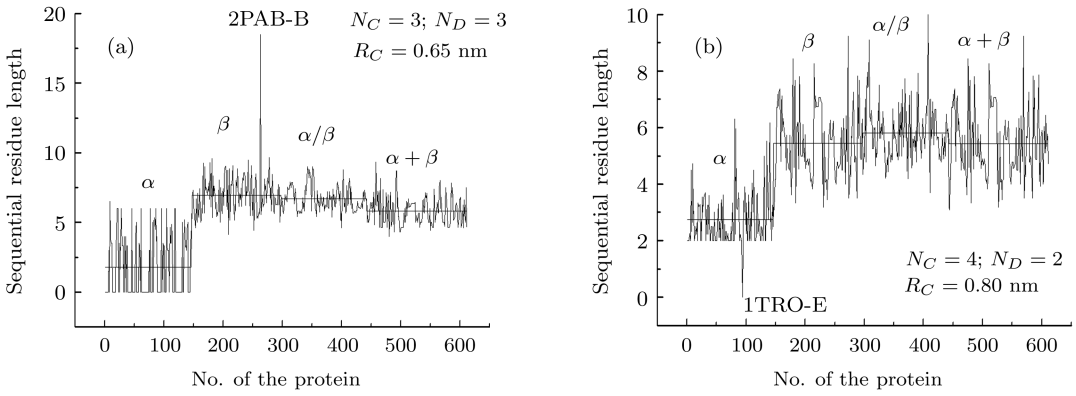


Fig.4. Distribution of sequential residue length L_i for all protein samples. Here No. of the protein is listed in Table 1, and the total number of proteins is 611.

The same conclusions as we obtained in Fig.3 give the connection between the 3D structure and its primary sequence data or secondary configuration. According to our discussions, the parameters N_C and N_D are helpful to explore the residue–residue contacts in detail, which have many affiliations with the 3D structure of proteins. The other structural information of globular proteins could also be analysed by such a model.

3.3. The calculation under the different boundary conditions

The average residue percentage $\bar{P}$ over all protein samples under the condition of N_C from 2 to 5 and N_D from 2 to 10 is calculated. The results calculated under the limited radius $R_C = 0.65$ nm in Fig.5(a) and

$R_C = 0.80$ nm in Fig.5(b) are plotted. In Fig.5, the value of $\bar{P}$ is decreased with the increase of N_D. The general conclusion is as follows: (i) all-α proteins have the minimum average residue percentage among these four categories of globular proteins at the certain value of N_D. (ii) All-β proteins have the maximum of $\bar{P}$ at the certain value of N_D in Fig.5(a). But it does not change much from $R_C = 0.65$ nm to $R_C = 0.80$ nm. Except for all-β proteins, $\bar{P}$ of the other three types of proteins are dramatically increased with increasing R_C. When R_C is increased from 0.65 nm to 0.80 nm, more valid residues with long-range contacts are taken to count, leading to the increase of the average residue percentage. So, such tendency in all-β proteins indicates that the distances between two residues constructing a long-range contact are close to the distance between β sheets. Actually, our statistical results do show an agreement with the experiment results, i.e., the distance between parallel β sheets is about 0.65 nm and anti-parallel ones are 0.80 nm. (iii) The average residue percentage $\bar{P}$ of all-α proteins is decreased from 0.086 to 0.025 with increasing N_D in Fig.5(a) and 0.33 to 0.12 in Fig.5(b), while $\bar{P}$ of all-β proteins is decreased from 0.49 to 0.19 for $R_C = 0.65$ nm and 0.53 to 0.27 for $R_C = 0.80$ nm. It can be seen that all-α proteins decrease more dramatically than all-β proteins with the increase of N_D from 2 to 10. So, under a certain value of N_C, N_D affects the structure of all-α proteins deeply. This is the same conclusion as obtained from Fig.3 and the reason is mentioned in the above paragraphs.

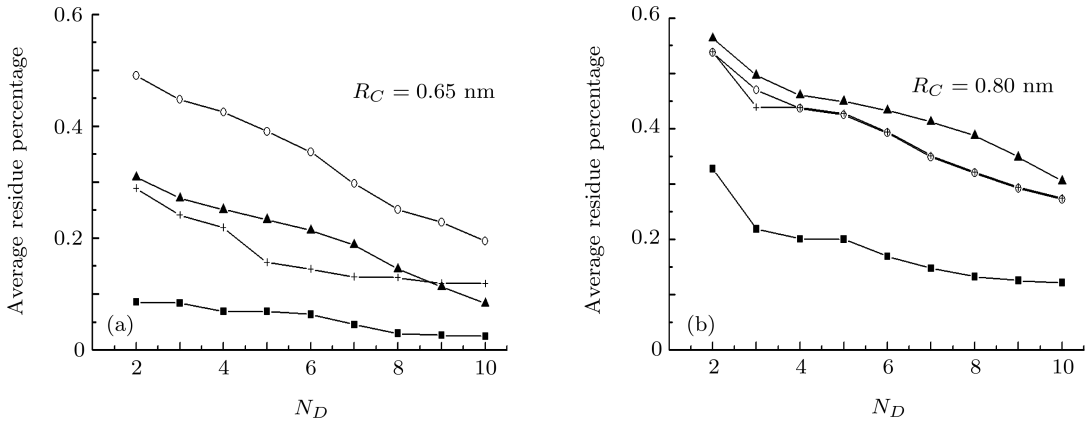


Fig.5. The average residue percentage $\bar{P}_i$ versus N_D for different R_C. Here $N_C = 3$, and (■), (○) , (▲) , (+) represents all-α, all-β, α/β and $\alpha + \beta$ proteins, respectively.

We have calculated the average sequential residue length $\bar{L}$ over all 611 protein samples under different boundary conditions as well as the value of $\bar{P}$. In order to compare with Fig.5, $N_C = 3$ is also chosen as the boundary condition and the typical curves are plotted in Fig.6. From these curves, the average sequential residue length $\bar{L}$ increases with increasing N_D. It is understandable since a stem with its length not longer than N_D, all residues, which as the component of that stem, are excluded from the statistical result. The figures show the following features: (i) The minimum of the average sequential residue length $\bar{L}$ is related to all-α proteins. It is also demonstrated that all-α proteins, because of the helix secondary conformation, are lacking in ability to form sequential long-range contacts. (ii) In Fig.6(a), the maximum of $\bar{L}$ comes from all-β proteins for $N_D \leq 6$. The gaps between all-β proteins and other three types of proteins are not obvious, similar to as the curves in Fig.5(a). Also, under the condition of $N_D > 6$, $R_C = 0.65$ nm, the maximum of $\bar{L}$ occurs at $\alpha + \beta$ proteins. In Fig.6(b), the maximum of $\bar{L}$ is related to α/β proteins when $N_D < 8$ and the values of $\bar{L}$ of all-β, α/β and $\alpha + \beta$ proteins are almost the same when $N_D \geq 8$ with $R_C = 0.80$ nm. (iii) The curves of all-β and $\alpha + \beta$ proteins are almost overlapped at $R_C = 0.80$ nm. It is demonstrated also that the structural characters on residue percentage and sequential residue length are supplemented by each other and give us the new way to comprehend the structural information of a protein. (iv) Unlike the curves in Fig.5, the values are not increased much from Fig.6(a) to Fig.6(b) with the increase of R_C. The radius R_C is not the main effect on the statistical results of $\bar{L}$, especially in the case of all-α proteins. According to these features we can conclude that residue percentage and sequential residue length are two important aspects of structural characters of a protein. In this work, besides the limit radius of sphere R_C, the boundary condition is expressed by two parameters N_C and

N_D, which are helpful to discover the structural information of globular proteins. In addition, the structural information of different categories of proteins is well discussed according to the influence of N_C and N_D.

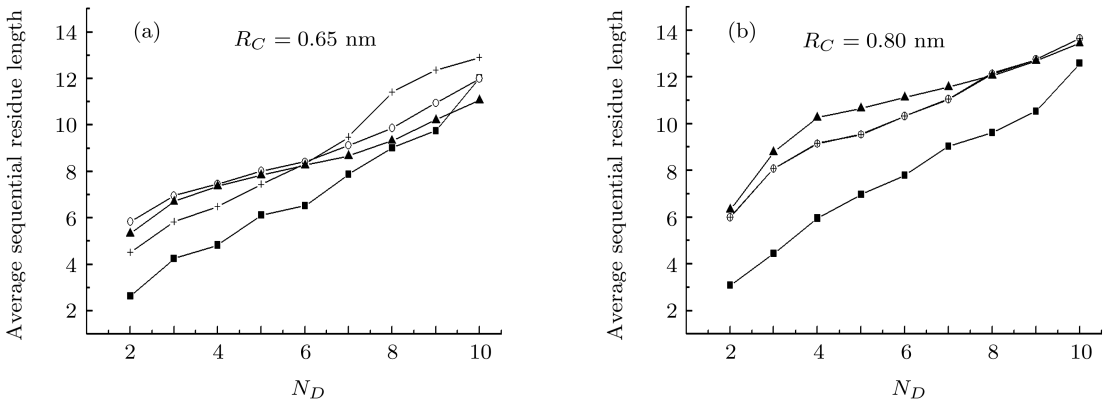


Fig.6. The average sequential residue length $\bar{L}_i$ versus N_D for different R_C. Here $N_C = 3$, and (■), (○), (▲) , (+) represents all-α, all-β, α/β and $\alpha + \beta$ proteins, respectively.

4. Conclusion

The interior characters of globular proteins in forming long-range contacts have been investigated by the statistical method. The statistical results demonstrate that hydrophobic amino acid residues have a strong attractive force in forming a long-range contact. Meanwhile, two parameters N_C and N_D are defined to confine the valid residues in detail. From the discussion in Section 3, we can conclude that the four categories of globular proteins have different abilities to construct long-range contacts, which are the important key in the protein folding process. All-β proteins, with the parallel and anti-parallel configurations of β sheets in secondary structure, are easy to form long-range contacts continuously along their primary sequence, while all-α proteins are hardly to form sequential long-range contacts, i.e., the long-range contacts along their primary sequence are always separated by their secondary configuration called helix. This property of globular proteins is strongly related to their secondary structure. In future work, we will continue to make effort for decreasing the big gap between the 3D structure and the available sequence data of proteins. We may focus on analysing the character in secondary structures of globular proteins in detail, and try to predict the structure of the protein and classify it.

References

[1] Anfinsen C, Haber E, Sela M and White F H 1961 *Proc. Natl. Acad. Sci.* U.S.A. **47** 1309
[2] Hubner I A, Deeds E J and Shakhnovich E I 2006 *Proc. Natl. Acad. Sci.* U.S.A. **103** 17747
[3] Cardenas A E and Elber R 2003 *Biophys. J.* **85** 2919
[4] Faccioli P, Sega M, Pederiva F and Orland H 2006 *Phys. Rev. Lett.* **97** 108101
[5] Rost B, Schneider R and Sander C 1997 *J. Mol. Biol.* **270** 471
[6] Kopp J and Schwede T 2004 *Nucleic Acids Res.* **32** D230
[7] Hermann J C, Arbona R M, Fedorov A A, Fedorov E, Almo S C, Shoichet B K and Raushel F M 2007 *Nature* **448** 775
[8] Kamtekar S, Schiffer J M, Xiong H, Babik J M and Hecht M H 1993 *Science* **262** 1680
[9] Dill K A, Bromberg S, Yue K, Fiebig K M, Yee K M, Thomas P D and Chan H S 1995 *Protein Sci.* **4** 561
[10] Yue K, Fiebig K M, Thomas P D, Chan H S, Shackhnovich E I and Dill K A 1995 *Proc. Natl. Acad. Sci.* U.S.A. **92** 325
[11] Regan L and DeGrado W F 1988 *Science* **241** 976
[12] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623
[13] Jiang Z, Zhang L, Chen J, Xia A and Zhao D 2002 *Polymer* **43** 6037
[14] Levitt M and Chothia C 1976 *Nature* **261** 552
[15] Hubbard T J P, Ailey B, Brenner S E, Murzin A G and Chothia C 1999 *Nucl. Acids Res.* **27** 254

[16] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534
[17] Zhang L and Sun T 2005 *Biophys. Chem.* **113** 9
[18] Sun T, Zhang L, Chen J and Jiang Z, 2004 *Chaos, Solitons and Fractals* **20** 1075
[19] Manavalan P and Ponnuswamy P K 1978 *Nature* **275** 673
[20] Ponnuswamy P K 1993 *Prog. Biophs. Mol. Biol.* **59** 57
[21] Gromiha M M and Selvaraj S 2000 *Recent Res. Dev. Biophys Chem.* **1** 1
[22] Debe D A and Goddard W A 1999 *J. Mol. Biol.* **294** 619
[23] Gromiha M M, Oobatake M, Kono H, Uedaira H and Sarai A 1999 *Prot. Eng.* **12** 549
[24] Gromiha M M 2001 *Biophys Chem.* **91** 71
[25] Jiang Z, Zhang L, Chen J, Xia A and Zhao D 2004 *Polymer* **45** 609
[26] DeVido D R, Dorsey J G., Chan H S and Dill K A 1998 *J. Phys. Chem.* B **102** 7272
[27] Fauchere J L and Pliska V 1983 *Eur. J. Med. Chem.* **18** 369
[28] Wimley W C and White S H 1998 *Nature Struct. Biol.* **3** 842