

球状蛋白质分子中氨基酸紧密接触对性质的研究*

陈 进 李 靖 章林溪 姜舟婷 夏阿根 叶高翔

(浙江大学物理系 杭州 310028)

赵得禄

(中国科学院化学研究所分子科学中心 北京 100080)

摘 要 采用 CSU 软件 (Contacts of structural units), 对 61 种球状蛋白质分子中氨基酸紧密接触对 (Residue-residue contact) 进行了研究. 重点研究了不同氨基酸在形成远程紧密接触对 (Long-range contact) 和近程紧密接触对 (Short-range contact) 时的不同能力. 发现氨基酸 Leu, Val, Ile, Met, Phe, Tyr, Cys, Trp (疏性氨基酸, H) 比较容易形成远程紧密接触对, 氨基酸 Glu, Gln, Asp, Asn, Lys, Ser, Arg, Pro (亲水氨基酸, P) 比较难形成远程紧密接触对, 而氨基酸 Ala, Gly, Thr, His (中性氨基酸, N) 在形成远程紧密接触对时能力一般. 它们平均每个氨基酸可形成 6.03, 3.64 和 4.43 个远程紧密接触对. 同时它们在形成近程紧密接触对时能力非常接近, 平均每个氨基酸可形成的近程紧密接触对数目在 2.34 ~ 2.85 变化, 差别非常小. 亲水氨基酸 (P), 中性氨基酸 (N) 和疏水性氨基酸 (H) 在蛋白质分子结构稳定性上起着不同的作用.

关键词 球状蛋白质, 紧密接触对, CSU 软件, 折叠

球状蛋白质是一类非常普遍的蛋白质, 它的特点是蛋白质分子非常紧密, 内部空间的绝大部分被原子所充满, 体系的能量来源于氨基酸与氨基酸的紧密接触对的作用 (Residue-residue contact) (见图 1)^[1], 因此研究氨基酸与氨基酸的紧密接触对的相互作用在研究球状蛋白质结构和稳定性时显得特别重要. Miyazawa 和 Jernigan 首先研究了球状蛋白质分子中氨基酸与氨基酸的紧密接触对的相互作用^[2,3]. 我们也研究了氨基酸与氨基酸的紧密接触对的相互作用, 建立了新的蛋白质格点模型-HNP 格点模型^[4]. 最近 Gromiha 和 Selvaraj 开始研究氨基酸紧密接触对在蛋白质折叠过程中所起的重要作用^[5-7], 我们也研究了不同的氨基酸在形成紧密接触对能力的差异^[8]. 但是这些研究的最大问题是在研究紧密接触对时仅仅考虑了二个氨基酸中的 $C^\alpha-C^\alpha$ 距离 (氨基酸中直接与 $-COOH$ 相连的碳原子称为 C^α), 即如果 $R_{C^\alpha-C^\alpha} < R_c$, 就形成一个紧密接触对, 其中 R_c 是定义紧密接触对时的某一个常数, 可以取某一个特定的值, 如 $R_c = 0.65$ nm, 或者 0.80 nm 等^[2,3,6-8]. 这些研究就没有考虑氨基酸的差异, 实际上氨基酸在侧基上差异是非常大的. 如 Gly 氨基酸的侧基为 H, 而 Arg 氨基酸的侧基为 $(CH_2)_3NHC(NH)NH_2$; 另外氨基酸之间如能形成氢键, 这样氨基酸的相互

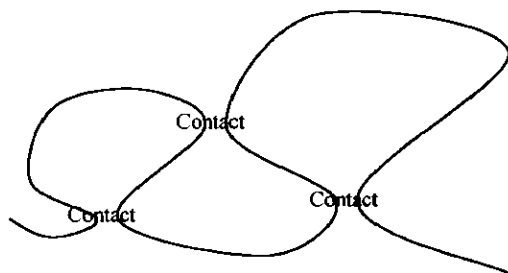


Fig. 1 A configuration of protein molecule. Many residues can form long-range and short-range contacts in globular protein.

作用就非常强的, 也应该算是形成紧密接触对. 因此仅仅考虑 $C^\alpha-C^\alpha$ 之间的距离是不够的. Sobolev 等在研究紧密接触对时考虑了各种因素, 开发了 CSU (Contact of structural units) 软件^[9]. 这里我们采用 CSU 软件来研究不同氨基酸在形成紧密接触对能力的差异, 来揭示亲水氨基酸 (P), 中性氨基酸 (N) 和疏水性氨基酸 (H) 在形成远程紧密接触对和近程紧密接触对时的不同能力, 从而来研究蛋白质分子的稳定性.

1 计算方法

在蛋白质分子中, 有 20 种氨基酸. 在讨论紧密接触对时考虑: (1) 二个氨基酸中原子的最近距离; (2) 二个氨基酸的重叠面积; (3) 二个氨基酸之

* 2002-04-08 收稿, 2002-07-08 修稿; 国家自然科学基金 (基金号 29874021, 20174036 和 20274040)、国家重点基础研究专项经费 (批准号 95-预-12 和 G1999064800) 和浙江省自然科学基金 (基金号 101002) 资助项目

间形成的氢键;(4)二个氨基酸之间的疏水-疏水紧密接触对;(5)二个氨基酸之间的疏水-亲水紧密接触对.对于不同的球状蛋白质,从已知的 X-ray 结构数据出发(Protein data bank),采用 CSU 软件,来讨论他们之间形成的紧密接触对.

选择了 61 种不同的蛋白质,他们的代码为: 1ABE, 1AE4, 1ALD, 1AQP, 1BKS, 1C52, 1CA2, 1CCD, 1CD8, 1CI3, 1CTJ, 1DLW, 1DLY, 1DP2, 1DVO, 1DXD, 1EAF, 1EDY, 1EST, 1ETU, 1GC9, 1GRY, 1HCZ, 1ID2, 1IPD, 1LHL, 1MNS, 1NBC, 1NEU, 1OFV, 1PAX, 1POP, 1PPN, 1Q21, 1RHD, 1TPA, 1ULA, 2AAT, 2ACT, 2AYH, 2CAB, 2CCY, 2PIA, 2POR, 2SNS, 2STV, 3ADK, 3CHY, 3DFR, 3LYZ, 3PGM, 3RN3, 3SSI, 4CHA, 4CLA, 4CPV, 4MBN, 5CYT, 5PTP, 7GCH, 7RSA. 记录它们中形成紧密接触对的氨基酸,以及它们之间的相对“位置”.如果第 i 个氨基酸与第 j 个氨基酸形成一个紧密接触对且 $|j-i| \geq 5$,那么,就定义为一个远程紧密接触对;如果 $4 > |j-i| \geq 2$ 就定义为一个近程紧密接触对^[4-8].

2 结果与讨论

2.1 不同氨基酸之间形成紧密接触对的能力

我们记录不同氨基酸之间形成的紧密接触

对,根据它们之间的相对“位置”来决定近程紧密接触对和远程紧密接触对.由于不同氨基酸出现的几率是不同的,如 Leu, Ala, Gly 出现的几率比较大, Met, Cys, Trp 出现的几率比较小.因此单纯从 A 氨基酸与 B 氨基酸形成的紧密接触对数目来反映不同氨基酸之间形成的紧密接触对的能力不是很合理,因此这里引入一个 A 氨基酸和一个 B 氨基酸形成的平均紧密接触对数目 P_{A-B} , 定义为^[8]:

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

这里 N_{A-B} 是 A 氨基酸和 B 氨基酸形成的紧密接触对数目, N_A 和 N_B 分别为总的 A 氨基酸和 B 氨基酸的数目.在表 1 中给出了不同氨基酸之间的平均紧密接触对数目 P_{A-B} , 上三角的为远程紧密接触对的结果,下三角的为近程紧密接触对的结果.这里我们把 A-B 和 B-A 归纳在一起,发现 Leu-Leu, Val-Val, Leu-Val, Ile-Leu, Ile-Val, Ile-Ile, Phe-Leu, Phe-Phe, Phe-Ile, Ala-Leu 这 10 个氨基酸对是最容易形成远程紧密接触对,它们的 P_{A-B} 值都大于 0.5; 而 Glu-Cys, Asn-His, Asp-Cys, Lys-Cys, Arg-Met, Cys-Arg, Pro-Cys 这几个氨基酸对是最难形成远程紧密接触对,它们的 P_{A-B} 值都小于 0.1. 对于近程紧密接触对, P_{A-B} 的值都比较小,这主要它

Table 1 Average number P_{A-B} of residue-residue contacts for different amino acids A and B

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
Leu	0.760	0.669	0.588	0.269	0.533	0.381	0.178	0.292	0.513	0.334	0.323	0.215	0.286	0.206	0.212	0.199	0.232	0.264	0.256	0.260	Leu
Val	0.257	0.706	0.570	0.268	0.489	0.459	0.191	0.260	0.451	0.313	0.319	0.189	0.247	0.209	0.215	0.173	0.296	0.266	0.196	0.244	Val
Val	0.206	0.190	0.525	0.183	0.504	0.387	0.145	0.251	0.408	0.311	0.280	0.131	0.204	0.195	0.177	0.177	0.209	0.218	0.178	0.243	Ile
Ile	0.188	0.144	0.163	0.176	0.302	0.187	0.116	0.166	0.216	0.183	0.132	0.150	0.144	0.122	0.095	0.123	0.104	0.135	0.083	0.179	Met
Met	0.109	0.099	0.070	0.039	0.537	0.388	0.161	0.257	0.319	0.252	0.229	0.202	0.187	0.182	0.201	0.173	0.233	0.226	0.147	0.241	Phe
Phe	0.138	0.128	0.134	0.056	0.103	0.412	0.175	0.275	0.301	0.303	0.247	0.169	0.253	0.230	0.234	0.194	0.246	0.266	0.238	0.248	Tyr
Tyr	0.105	0.123	0.103	0.053	0.095	0.147	0.363	0.148	0.174	0.198	0.184	0.100	0.067	0.115	0.094	0.167	0.077	0.156	0.074	0.063	Cys
Cys	0.121	0.074	0.062	0.044	0.071	0.056	0.113	0.201	0.261	0.264	0.180	0.155	0.125	0.151	0.112	0.135	0.128	0.171	0.105	0.172	Trp
Trp	0.121	0.093	0.062	0.047	0.070	0.042	0.041	0.067	0.442	0.331	0.312	0.209	0.260	0.253	0.218	0.220	0.244	0.294	0.186	0.244	Ala
Ala	0.226	0.208	0.153	0.115	0.130	0.139	0.098	0.063	0.407	0.344	0.324	0.142	0.229	0.216	0.228	0.203	0.262	0.248	0.231	0.245	Gly
Gly	0.192	0.161	0.147	0.067	0.134	0.154	0.124	0.118	0.226	0.172	0.266	0.150	0.204	0.169	0.195	0.236	0.209	0.238	0.196	0.160	Thr
Thr	0.143	0.163	0.099	0.078	0.099	0.127	0.079	0.077	0.186	0.173	0.116	0.143	0.156	0.119	0.153	0.072	0.112	0.163	0.119	0.126	His
His	0.135	0.138	0.073	0.065	0.089	0.078	0.111	0.099	0.122	0.063	0.072	0.080	0.201	0.150	0.184	0.113	0.332	0.176	0.243	0.134	Glu
Glu	0.187	0.141	0.126	0.125	0.131	0.147	0.041	0.081	0.208	0.172	0.126	0.119	0.178	0.146	0.133	0.170	0.147	0.185	0.148	0.181	Gln
Gln	0.168	0.118	0.119	0.092	0.084	0.104	0.059	0.067	0.167	0.126	0.134	0.045	0.121	0.128	0.171	0.156	0.264	0.184	0.206	0.120	Asp
Asp	0.145	0.126	0.116	0.058	0.108	0.110	0.099	0.066	0.186	0.180	0.125	0.079	0.102	0.123	0.101	0.164	0.145	0.199	0.138	0.129	Asn
Asn	0.151	0.070	0.116	0.092	0.094	0.134	0.106	0.095	0.137	0.118	0.121	0.042	0.137	0.121	0.104	0.167	0.202	0.182	0.116	0.141	Lys
Lys	0.183	0.154	0.145	0.081	0.107	0.069	0.053	0.072	0.233	0.167	0.157	0.099	0.266	0.156	0.194	0.128	0.142	0.176	0.138	0.144	Ser
Ser	0.153	0.129	0.113	0.085	0.119	0.115	0.084	0.091	0.178	0.175	0.140	0.075	0.126	0.118	0.161	0.138	0.148	0.127	0.155	0.124	Arg
Arg	0.181	0.125	0.119	0.092	0.100	0.099	0.056	0.053	0.164	0.156	0.134	0.088	0.181	0.134	0.150	0.123	0.103	0.100	0.137	0.153	Pro
Pro	0.134	0.114	0.086	0.055	0.124	0.100	0.070	0.075	0.141	0.117	0.126	0.090	0.088	0.078	0.115	0.100	0.134	0.122	0.093	0.088	
	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	

Upper triangle is the long-range contacts and lower triangle is the short-range contacts in protein samples.

仅仅与蛋白质中氨基酸的序列有关^[8]。

2.2 不同氨基酸形成紧密接触对的能力

上面分析了一个 A 氨基酸和一个 B 氨基酸形成紧密接触对的能力,如果说 A 氨基酸形成紧密接触对的能力比较弱,或者 B 氨基酸形成紧密接触对的能力比较弱,那么 P_{A-B} 就比较小.反之,如果知道了 P_{A-B} ,也不能判断 A 或 B 氨基酸形成紧密接触对的能力.因此这里就考虑一个氨基酸形成紧密接触对的能力.定义一个氨基酸能形成紧密接触对的数目 C_a :

$$C_{a,A} = \frac{\sum_{\beta = \text{Leu, Val, \dots, Pro}} [N_{A-\beta} + N_{\beta-A}]}{N_A} \quad (a = S \text{ or } L)$$

上式中考虑了 A-β 和 β-A 的紧密接触对都是与 A 氨基酸有关的紧密接触对,S 和 L 分别表示的近程或远程紧密接触对. N_A 为 A 氨基酸的总数目, $N_{A-\beta}$ 表示 A-β 紧密接触对的数目.计算结果见表 2.

Table 2 Average number of contacts per residue

20 Amino acids	Three types of amino acids	P (%)	C _S	$\bar{C}_S$	C _L	$\bar{C}_L$
Leu	Hydrophobic (H)	8.31	2.56	2.55	5.50	6.03
Val		7.68	2.27		5.58	
Ile		4.81	2.46		6.20	
Met		1.86	2.59		5.38	
Phe		3.71	2.50		6.69	
Tyr		3.59	2.53		6.62	
Cys		1.46	2.64		4.91	
Trp		1.64	2.85		7.33	
Ala	Neutral (N)	9.56	2.56	2.49	4.18	4.43
Gly		7.77	2.32		4.23	
Thr		5.69	2.38		4.30	
His		2.5	2.69		4.63	
Glu	Hydrophilic (P)	6.03	2.62	2.49	3.65	3.64
Gln		4.14	2.56		3.76	
Asp		5.44	2.37		3.44	
Asn		4.47	2.47		3.46	
Lys		6.58	2.56		3.45	
Ser		5.90	2.36		3.71	
Arg		4.13	2.63		3.74	
Pro		5.19	2.34		3.92	

$C_S(C_L)$ is the 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 $C_S(C_L)$.

在表 2 中,发现氨基酸 Leu, Val, Ile, Met, Phe, Tyr, Cys 和 Trp 的 C_L 值都比较大,其值大于 5.00,其平均值有 6.03.氨基酸 Ala, Gly, Thr 和 His 的 C_L 值在 4.18 ~ 4.63 之间,平均值为 4.43.但是氨基酸 Glu, Gln, Asp, Asn, Lys, Ser, Arg 和 Pro 的 C_L 值就比较小,只有 3.44 ~ 3.92,平均值只有 3.64.我们把它归类于疏性氨基酸(H),中性氨基酸(N)和亲水氨基酸(P),与 Jernigan 得到的结论是一致的^[2].这个结论非常重要,对于了解蛋白质结构的稳定性非常重要.

同样还计算了氨基酸的近程紧密接触对,发现 20 种氨基酸的 C_S 值都比较接近.3 类氨基酸(H, N, P)的平均 C_S 值分别为 2.55, 2.49 和 2.49.可见不同氨基酸在形成近程紧密接触对能力是相同的,它仅仅与蛋白质的氨基酸序列有关.

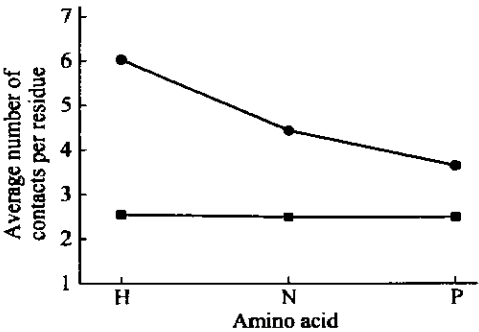


Fig. 2 Average number of contacts per residue for three types of amino acids Here H, N and P represent hydrophobic, neutral and hydrophilic residues, respectively ■(●) is the average number of short-range (long-range) contacts per residue.

图 2 给出了 3 类氨基酸(H, N, P)的平均 C_S 和 C_L 值,可见疏性氨基酸的 C_L 非常大.如果没有疏性氨基酸,没有远程紧密接触对,就没有球状蛋白质的结构.当然如果全部为疏性氨基酸,虽然能量是非常低的,但是简并度不为 1,结构也是不稳定的.通过对氨基酸近程或远程紧密接触对的研究,能够加深对蛋白质结构的了解,能帮助我们认识蛋白质的稳定和折叠问题.

REFERENCES

- 1 Chan H S, Dill K A. *Macromolecules*, 1989, 22: 3988 ~ 3997; 4559 ~ 4573
- 2 Miyazawa S, Jernigan R L. *Macromolecules*, 1985, 18: 543 ~ 552
- 3 Miyazawa S, Jernigan R L. *J Mol Biol*, 1996, 256: 623 ~ 644
- 4 Zhang Linxi(章林溪), Jin Jinsheng(金进生), Wang Xianghong(王向红), Ye Gaoxiang(叶高翔), Zhao Delu(赵得禄). *Acta Polymerica Sinica(高分子学报)*, 2002, (4): 497 ~ 501
- 5 Gromiha M M. *Biophys Chem*, 2001, 91: 71 ~ 77
- 6 Gromiha M M, Selvaraj S. *Biophys Chem*, 1999, 77: 49 ~ 68
- 7 Gromiha M M, Selvaraj S. *Inter J Biol Macromol*, 2001, 29: 25 ~ 34
- 8 Jiang Z, Zhang L, Xia A, Ye G, Zhao D. *Polymer*, 2002, 43: 6037 ~ 6047
- 9 Sobolev V, Sorokine A, Prilusky J, Abola E. *Bioinformatics*, 1999, 15: 327 ~ 330

A STUDY OF RESIDUE-RESIDUE CONTACTS IN GLOBULAR PROTEIN MOLECULES

CHEN Jin, LI Jing, ZHANG Linxi, JIANG Zhouting, XIA Agen, YE Gaoxiang

(*Department of Physics, Zhejiang University, Hangzhou 310028*)

ZHAO Delu

(*Polymer Physics Laboratory, Center of Molecular Science, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100080*)

Abstract The residue-residue contacts in 61 globular protein molecules were studied by using CSU software. The aim is to find the ability of forming the short-range and long-range residue-residue contacts for several residues. It is found that it is easy to form the long-range residue-residue contacts for Leu, Val, Ile, Met, Phe, Tyr, Cys, Trp residues, and it is difficult to form the long-range residue-residue contacts for Glu, Gln, Asp, Asn, Lys, Ser, Arg, Pro residues. The residue are called residues hydrophobic (H) and hydrophilic (P) residues, respectively, and the other residues is called neutral residue (N). The average number of long-range contacts per residue is 6.03, 4.43 and 3.64 for H, N and P residues respectively. However, the average number of short-range contacts per residue ranges from 2.34 to 2.85, and it is almost no difference among those residues in forming the short-range contacts. Our research may provide some insights into the stability and folding of protein molecules.

Key words Globular Protein, Contact, CSU software, Folding