ELSEVIER

Biophysical Chemistry 105 (2003) 11–21

Biophysical Chemistry

www.elsevier.com/locate/bpc

# Predicting protein structure from long-range contacts

Jin Chen[a], Linxi Zhang[a,*], Li Jing[a], Youxing Wang[a], Zhouting Jiang[a], Delu Zhao[b]

[a]*Department of Physics, Zhejiang University, Hangzhou 310028, PR China*
[b]*Polymer Physics Laboratory, Center of Molecular Science, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100080, PR China*

Received 16 September 2002; received in revised form 6 January 2003; accepted 22 January 2003

## Abstract

Short-range and long-range contacts are important in forming protein structure. The proteins can be grouped into four different structural classes according to the content and topology of α-helices and β-strands, and there are all-α, all-β, α/β and α+β proteins. However, there is much difference in statistical property for those classes of proteins. In this paper, we will discuss protein structure in the view of the relative number of long-range (short-range) contacts for each residue. We find the percentage of residues having a large number of long-range contacts in protein is small in all-α class of proteins, and large in all-β class of proteins. However, the percentage of residues is almost the same in α/β and α+β classes of proteins. We calculate the percentage of residues having the number of long-range contacts greater than or equal to (≥) $N_L=5$, and 7 for 428 proteins. The average percentage is 13.3%, 54.8%, 41.4% and 37.0% for all-α, all-β, α/β and α+β classes of proteins with $N_L=5$, respectively. With $N_L$ increasing, the percentage decreases, especially for all-α class of proteins. In the meantime, the percentage of residues having the number of short-range contacts greater than or equal to $N_S$ ($\geq N_S$) in protein samples is large for all-α class of proteins, and small for all-β class of proteins, especially for large $N_S$. We also investigate the ability of amino residues in forming a large number of long-range and short-range contacts. Cys, Val, Ile, Tyr, Trp and Phe can form a large number of long-range contacts easily, and Glu, Lys, Asp, Gln, Arg and Asn can form a large number of long-range contacts, but with difficulty. We also discuss the relative ability in forming short-range contacts for 20 amino residues. Comparison with Fauchere–Pliska hydrophobicity scale and the percentage of residues having large number of long-range contacts is also made. This investigation can provide some insights into the protein structure.

*Keywords:* Protein structure; Short-range and long-range contact; Protein structural class

## 1. Introduction

In protein molecules, short-range and long-range contacts are very important, because there is an effective attraction among the contacts. The folding of a polypeptide chain into a compact, unique three-dimensional structure is directed and stabilized by molecular interactions between the constituent amino acid residues along the chain. The knowledge of inter-residue interactions can help us understand the mechanism of protein folding and stability. Tanaka et al. first categorized the inter-residue interactions into short-, medium and long-range and proposed a hypothesis for protein folding by a three-step mechanism based on these

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

0301-4622/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0301-4622(03)00033-4

interactions [1]. Later, many biologists, chemists and physicists have done a great deal of work on all kinds of fields [2–4]. Recently, Gromiha et al. analyzed the influence of medium- and long-range interactions in different structural classes of proteins, and investigated the importance of long-range interaction in protein folding [5–8].

On the other hand, globular proteins are grouped into four structural classes based on the content and topology of α-helices and β-strands in the three-dimensional structures [9]. α-Helices and β-strands may have a different exhibition in structure. We all know that the proteins can be categorized into four different structural classes, namely: all-α, all-β, α/β- and α+β-proteins. Some differences exist in protein inter-structure for four different structural classes. Gromiha investigated the effects of long-range contacts on protein folding according to the percentage of long-range contacts for different intervals with a step of 10 (4–10, 11–20, 21–30, 31–40, 41–50 and >50). In this paper, we will investigate the protein structure of four structural classes through the percentage of residue having a large number of long-range (short-range) contacts in all the residues of proteins, and we can predict the protein structure from the number of long-range contacts. In the meantime, we can also find the globular structure of proteins and the ability of amino acid residues to forming globular structure in more detail from the percentage of amino acid residues having a large number of long-range contacts.

## 2. Method of calculation

### 2.1. Database

In this paper, we study 428 globular protein structures. A database of these proteins is derived from the information about their three-dimensional structures available in the literature. The data of these globular protein structures are taken from the Protein Data Bank (PDB) [10]. The PDB codes for all the proteins used in the present study are listed in Table 1. The selected proteins are from four different structural classes: all-α, all-β, α/β and α+β proteins. We obtain the information about the structural class from SCOP [11] and CATH [12]. These are all-α class of proteins from no. 1 to no. 119, all-β class of proteins from no. 120 to 250, α/β class of proteins from no. 251 to 348, and α+β class of proteins from no. 349 to 428 in Table 1.

### 2.2. Computation of long-range contacts

Each residue in a protein molecule is represented by its α-carbon atom ($C^\alpha$). The center is fixed at the α-carbon atom of the first (N-terminal) residue and the distances between this atom and the rest of the α-carbon atoms in the protein molecule are computed. Residues whose distance between their center $C^\alpha$ atoms is shorter than $R_c$ are defined as a contact. This method has been shown in many articles [13–18], it is easy and effective to get the number of residue–residue contacts in protein. In our paper, we choose the value $R_c$ from 7.0 Å to 8.0 Å. In general, $R_c$ is chosen from 6.5 Å to 8.0 Å [13–16,18].

For a given residue, the composition of surrounding residues is discussed in terms of the location at the sequence level and the contributions from $> \pm 4$ residues treated as long-range contacts [1,2]. We also treated the contributions from $\leq \pm 4$ as short-range contacts. In this paper, our short-range contacts in fact include short-range contacts ($< \pm 3$) and medium-range contacts ($\pm 3$ and $\pm 4$).

### 2.3. Percentages of short-range contacts and long-range contacts in four structural classes of proteins

There are many short-range and long-range contacts in globular proteins. As there exists a great deal of differences for α-helices and β-strands in the three-dimensional structures, there may be a different three-dimensional structure for all-α, all-β, α/β and α+β proteins. If we only consider the total number of short-range and long-range contacts of proteins, we cannot know the globular structure in more detail. In fact, if a protein has a globular structure, there are some residues in the inside of the globular and the other residues are on the outside of the globular. Therefore, we should know those residues that are in the outside

Table 1
The PDB code of proteins used in this paper

| | | | | | | |
|---|---|---|---|---|---|---|
| *All-α proteins* | | | | | | |
| 1. 1BBL | 2. 1CIF | 3. 1FIA-A | 4. 1FIA-B | 5. 1GCN | 6. 1LMB-3 | 7. 1LMB-4 |
| 8. 1PPT | 9. 1UTG | 10. 2MLT-A | 11. 2MLT-B | 12. 2PDE | 13. 3INS-A | 14. 3INS-B |
| 15. 3INS-C | 16. 3INS-D | 17. 451C | 18. 4ICB | 19. 1BP2 | 20. 1CCR | 21. 1CPC-A |
| 22. 1CPC-B | 23. 1CPC-L | 24. 1CPC-K | 25. 1ECD | 26. 1ECO | 27. 1FCS | 28. 1FHA |
| 29. 1HBG | 30. 1I55-A | 31. 1I55-B | 32. 1IFA | 33. 1LE4 | 34. 1LH1 | 35. 1MBC |
| 36. 1MBD | 37. 1MBS | 38. 1PP2-R | 39. 1PP2-L | 40. 1YCC | 41. 256B-A | 42. 256B-B |
| 43. 2C2C | 44. 2CCY-A | 45. 2CCY-B | 46. 2CDV | 47. 2CY3 | 48. 2GMF-A | 49. 2GMF-B |
| 50. 2HCO-A | 51. 2HCO-B | 52. 2LHB | 53. 2MHB-A | 54. 2MHB-B | 55. 2MHR | 56. 2WRP-R |
| 57. 4CPV | 58. 4MBN | 59. 5CPV | 60. 5CYT | 61. 1COL-A | 62. 1COL-B | 63. 1PRC-L |
| 64. 1AVH-A | 65. 1AVH-B | 66. 4CPP | 67. 2CTS | 68. 1ALA | 69. 1BAB-A | 70. 1BAB-B |
| 71. 1BAB-C | 72. 1BAB-D | 73. 1E12-A | 74. 1EA8-A | 75. 1F63-A | 76. 1F88-A | 77. 1F88-B |
| 78. 1FIP-A | 79. 1FIP-B | 80. 1H96-A | 81. 1KZU-A | 82. 1KZU-B | 83. 1LGH-A | 84. 1LGH-B |
| 85. 1LH2 | 86. 1LPE | 87. 1MBA | 88. 1MSL-A | 89. 1MSL-B | 90. 1MSL-C | 91. 1MSL-D |
| 92. 1MSL-E | 93. 1PPA | 94. 1PRC-C | 95. 1PRC-H | 96. 1PRC-L | 97. 1PRC-M | 98. 1PSS-H |
| 99. 1PSS-L | 100. 1PSS-M | 101. 1RCB | 102. 1RRO | 103. 1TRO-A | 104. 1TRO-C | 105.1TRO-E |
| 106. 1TRO-G | 107. 1YEA | 108. 2BRD | 109. 2CYP | 110. 2END | 111. 2FAL | 112. 2HBG |
| 113. 1BAB-A | 114. 1BAB-C | 115. 3C2C | 116. 3CYT-I | 117. 3CYT-O | 118. 4BP2 | 119.5TNC |
| *All-β proteins* | | | | | | |
| 120. 1BOV-A | 121. 1BOV-B | 122. 1BOV-C | 123. 1BOV-D | 124. 1BOV-E | 125. 1CDT-A | 126. 1CDT-B |
| 127. 1HIV-A | 128. 1HIV-B | 129. 1HOE | 130. 1TEN | 131. 1TGS-I | 132. 1TPA-I | 133. 2PCY |
| 134. 1A45 | 135. 1ACX | 136. 1CD8 | 137. 1CID | 138. 1COB-A | 139. 1COB-B | 140. 1LTS-D |
| 141. 1LTS-E | 142. 1LTS-F | 143. 1LTS-G | 144. 1LTS-H | 145. 1LTS-A | 146. 1P12-E | 147. 1RBP |
| 148. 1REI-A | 149. 1REI-B | 150. 1TIE | 151. 1TLK | 152. 2ALP | 153. 2AVI-A | 154. 2AVI-B |
| 155. 2AZA-A | 156. 2AZA-B | 157. 2ILA | 158. 2LAL-A | 159. 2LAL-C | 160. 2LTN- | 161. 2LTN-B |
| 162. 2PAB-A | 163. 2PAB-B | 164. 2RHE | 165. 2RSP-A | 166. 2RSP-B | 167. 2SNS | 168. 2SNV |
| 169. 2SOD-O | 170. 2SOD-Y | 171. 2SOD-B | 172. 2SOD-G | 173. 2STV | 174. 3SGB-E | 175. 1CA2 |
| 176. 1CSE-E | 177. 1EST | 178. 1FCI-A | 179. 1FCI-B | 180. 1HIL-A | 181. 1HIL-B | 182. 1HIL-C |
| 183. 1HIL-D | 184. 1HSB-A | 185. 1MAM-L | 186. 1MAM-H | 187. 1PPF-E | 188. 1TGS-Z | 189. 1TPA-E |
| 190. 2AYH | 191. 2CAB | 192. 2CAN | 193. 2GCH | 194. 2PTC-E | 195. 3CNA | 196. 3EST |
| 197. 4CHA-A | 198. 4CHA-B | 199. 4FAB-L | 200. 4FAB-H | 201. 5PTP | 202. 1CD1-A | 203. 1CD1-C |
| 204. 1NSB-A | 205. 1NSB-B | 206. 2APR | 207. 2ER7-E | 208. 2PIA | 209. 2POR | 210. 3APP |
| 211. 2BPA-1 | 212. 1BXW-A | 213. 1CTX | 214. 1F3G | 215. 1F53-A | 216. 1GPR | 217. 1H6X-A |
| 218. 1HJC-A | 219. 1HJC-D | 220. 1KL9-A | 221. 1MPP | 222. 1NN2 | 223. 1PAZ | 224. 1PRN |
| 225. 1PYP | 226. 1QD5-A | 227. 1QJ8-A | 228. 1QNY-A | 229. 1REE-A | 230. 1REE-B | 231. 1SGT |
| 232. 1SHF-A | 233. 1SHF-B | 234. 1STP | 235. 1FHG | 236. 1TON | 237. 2BB2 | 238. 2CA2 |
| 239. 2FCP-A | 240. 2MCM | 241. 2MPR-A | 242. 2MPR-B | 243. 2MPR-C | 244. 2OMF | 245. 2REN |
| 246. 2SGA | 247. 3EBX | 248. 3ERT-A | 249. 4FGF | 250. 4PEP | | |
| *α/β proteins* | | | | | | |
| 251. 1ABA | 252. 1FX1 | 253. 1GKY | 254. 1OFV | 255. 1OVB | 256. 1Q21 | 257. 1RNH |
| 258. 2FCR | 259. 2FOX | 260. 2TRX-A | 261. 2TRX-B | 262. 3ADK | 263. 3CHY | 264. 3DFR |
| 265. 4DFR-A | 266. 4DFR-B | 267. 5P21 | 268. 8ATC-B | 269. 8ATC-D | 270. 1BKS | 271. 1DHR |
| 272. 1EAF | 273. 1PRC-H | 274. 1RHD | 275. 1RVE-A | 276. 1REV-B | 277. 1TIM-A | 278. 1TIM-B |
| 279. 1TRE-A | 280. 1TRE-B | 281. 1ULA | 282. 2DRI | 283. 2SBT | 284. 3PGM | 285. 4BLM-A |
| 286. 4BLM-B | 287. 4CLA | 288. 5TIM-A | 289. 5TIM-B | 290. 1ABE | 291. 1ALD | 292. 1ETU |
| 293. 1GOX | 294. 1IPD | 295. 1MNS | 296. 1PFK-A | 297. 1PFK-B | 298. 1SBP | 299. 2ACH-A |
| 300. 2GBP | 301. 2HAD | 302. 2LIV | 303. 3CPA | 304. 4PFK | 305. 5ABP | 306. 5ADH |
| 307. 5CPA | 308. 6XIA | 309. 8ADH | 310. 1GLA-G | 311. 2AAA | 312. 2PGD | 313. 2TAA-A |
| 314. 2TS1 | 315. 3PGK | 316. 4ENL | 317. 4ICD | 318. 8CAT-A | 319. 8CAT-B | 320. 1CIS |
| 321. 1CRN | 322. 1CSE-E | 323. 1E49-P | 324. 1E6K-A | 325. 1EXT-A | 326. 1EXT-B | 327. 1FCB-A |
| 328. 1FCB-B | 329. 1GBP | 330. 1GPD-G | 331. 1GPD-R | 332. 1PEK-A | 333. 1PEK-B | 334. 1CW4 |
| 335. 1TFD | 336. 1THM | 337. 2CAB | 338. 2FX2 | 339. 2PRK | 340. 3CLA | 341. 3COX |
| 342. 3LDH | 343. 4CPA | 344. 4CPA-I | 345. 1AI2 | 346. 8DFR | 347. 1GVB | 348. 1RH3 |

Table 1 *(Continued)*

| | | | | | | |
|---|---|---|---|---|---|---|
| $\alpha+\beta$ *proteins* | | | | | | |
| 349. 1CYO | 350. 1DUR | 351. 1FXD | 352. 1NRC-A | 353. 1NRC-B | 354. 2SAR-A | 355. 2SAR-B |
| 356. 1FKF | 357. 1LZ1 | 358. 1MSB-A | 359. 1MSB-B | 360. 2LYZ | 361. 2LZM | 362. 2MS2-A |
| 363. 2MS2-B | 364. 2MS2-C | 365. 3LYZ | 366. 3RN3 | 367. 3SSI | 368. 5FD1 | 369. 7RSA |
| 370. 9RNT | 371. 1PPN | 372. 2ACT | 373. 2TSC-A | 374. 2TSC-B | 375. 9PAP | 376. 1PAX |
| 377. 1PHH | 378. 1PRC-C | 379. 1PRC-M | 380. 4TLN | 381. 4TMS | 382. 6LDH | 383. 8TLN-E |
| 384. 102L | 385. 125L | 386. 190L | 387. 1AQP | 388. 1BKF | 389. 1CTF | 390. 1D9W-A |
| 391. 1E3V-A | 392. 1E3V-B | 393. 1EAF | 394. 1EZM | 395. 1FDD | 396. 1FKB | 397. 1FRH |
| 398. 1GWD-A | 399. 1HSB-A | 400. 1HSB-B | 401. 1I1Z-A | 402. 1I20-A | 403. 1IET | 404. 1IEU |
| 405. 1L3F-E | 406. 1LHH | 407. 1LHI | 408. 1LTS-D | 409. 1DZS-A | 410. 1DZS-B | 411. 1POP-A |
| 412. 1ROB | 413. 1SHA-A | 414. 2AAK | 415. 2ACH-A | 416. 2PAD | 417. 2PRF | 418. 3IL8 |
| 419. 3RUB-L | 420. 3RUB-S | 421. 3SIC-E | 422. 3SIC-I | 423. 4BLM-A | 424. 4BLM-B | 425. 4ENL |
| 426. 4LZM | 427. 8CAT-A | 428. 8CAT-B | | | | |

of the globular protein and those residues that are in the inside of the globular protein in order to know the structure of the protein. Here we introduce the percentages of short-range contacts and long-range contacts. If $N_{P_L}$ is the number of amino acid residues whose number of long-range contacts is greater than or equal to $N_L(\geq N_L)$, we have the percentage $P_L$ of residue having the number of long-range contacts greater than or equal to $N_L(\geq N_L)$

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

Here, $N$ is the total number of residues in a protein molecule. If a protein has a compact structure, there is a large value of $P_L$. If a protein has a loose structure, there is a small value of $P_L$. Sometimes, $P_L$ may be zero. Of course, $P_L$ depends on the value of $N_L$. In the meantime, the maximum value of $P_L$ is unity, this means that all the residues in a protein molecule have a large number of long-range contacts.

Here, we also define the percentage $P_S$ of residue having a number of short-range contacts greater than or equal to $N_S(\geq N_S)$ as

$$P_S=\frac{N_{P_S}}{N} \tag{2}$$

Here, $N_{P_S}$ is the number of amino acid residues whose number of long-range contacts is greater than or equal to $N_S(\geq N_S)$. Therefore, we can discuss the globular structure through calculation of percentages of short-range contacts and long-range contacts.

The average percentages of residue having the number of long-range contacts greater than or equal to $N_L(\geq N_L)$ per protein molecule in four structural classes are also considered in our work. We define it as

$$\overline{P_L}=\frac{\sum_{i=1}^{M}P_{L,i}}{M} \tag{3}$$

Here, $M$ represents the total number of proteins in four different structural classes, and $M=119$, 131, 98 and 80, respectively, for all-α, all-β, $\alpha/\beta$ and $\alpha+\beta$ proteins. $P_{L,i}$ is the percentage of residue having a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ for $i$th protein molecule from Table 1. For example, $P_{L,5}$ is the percentage of 1GCN.

We investigate the percentage of residue having a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ and the percentage of residue having a number of short-range contacts greater than or equal to $N_S(\geq N_S)$ for all 20 amino acid residues. Experiments and theoretical studies have shown the critical role played by the two types of residues, hydrophobic and polar [18–21]. 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 general, if the residue is hydrophobic, it is in the inside of the globular. The analysis of protein structure from the total

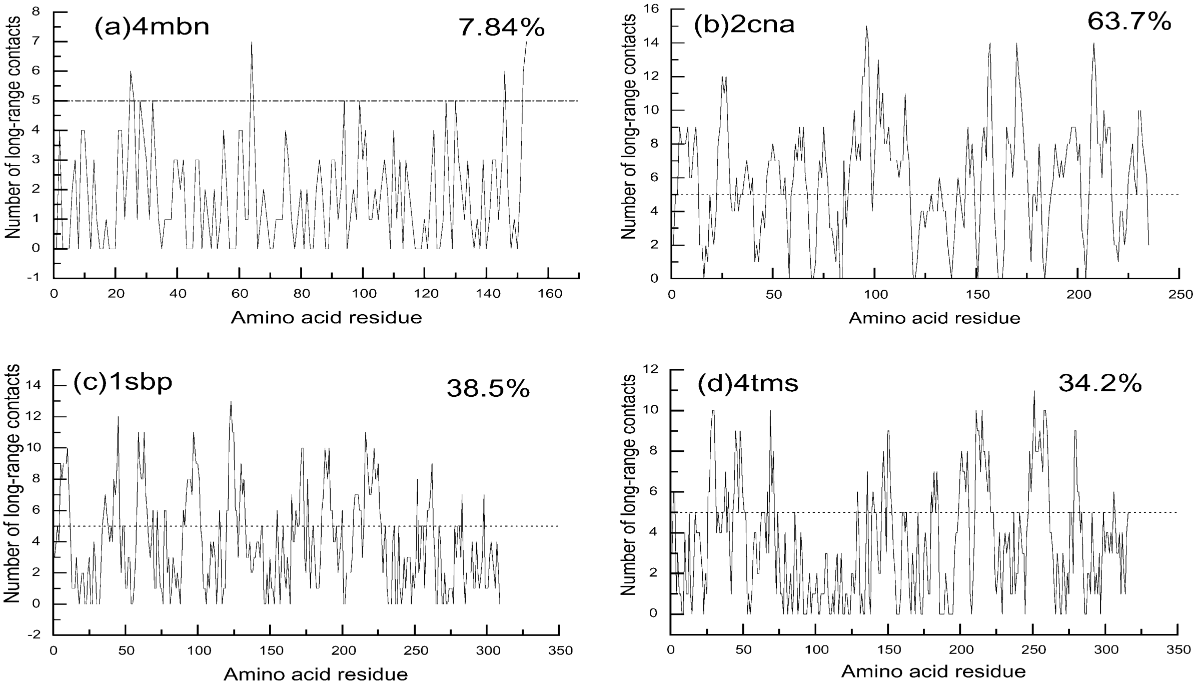

Fig. 1. Number of long-range contacts for four typical globular proteins in different structural classes: (a) 4mbn (all-α protein); (b) 2cna (all-β protein); (c) 1sbp (α/β protein); (d) 4tms (α+β protein).The symbol (...) represents five long-range contacts per residue, and 7.84%, 63.7%, 38.5% and 34.2% represent the percentage of residues having a number of long-range contacts greater than 4 (≥5) for 4mbn, 2can, 1sbp and 4tms, respectively.

number of long-range (or short-range) contacts may be unreasonable. Through calculation of the percentage of residues having a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ and the percentage of residue having the number of short-range contacts greater than or equal to $N_S(\geq N_S)$ for all the 20 amino acid residues, we can clearly determine protein structure.

## 3. Results and discussions

### *3.1. The difference in percentage of residues having large numbers of long-range contacts in four structural classes*

We first calculate the number of long-range contacts for every residue for 4mbn, 2can, 1sbp and 4tms, and the results are given in Fig. 1. The structural classes of these proteins are all-α, all-β, α/β and α+β proteins, respectively. From Fig. 1 we find that 4mbn protein (all-α class of protein) has a small percentage of residues with a large number of long-range contacts. Contrary to 4mbn (all-α class of protein), 2cna (all-β class of protein) has a large percentage of residues with a large number of long-range contacts. However, the situation of 1sbp and 4tms (α/β and α+β classes of proteins) are almost the same. For example, we assume $N_L=5$, the percentages of residue having a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ are 7.84%, 63.7%, 38.5% and 34.2%, respectively. In order to find the rule of different structural classes, we calculate the percentage of residues with a large number of long-range contacts in the 428 globular proteins, and the results are given in Fig. 2. Here, the number of proteins is given according to Table 1. For example, no. 1 in Fig. 2 represents the protein of 1BBL. In Fig. 2, $N_L$ is 5 and 7, respectively. We also calculate the percentages of residue having a

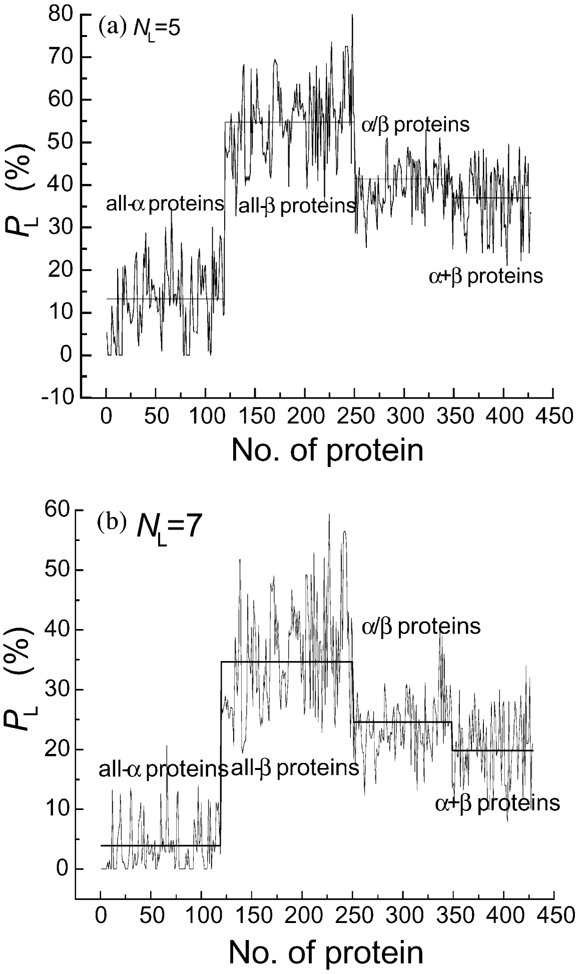

Fig. 2. Distribution of percentage of residues having a number of long-range contacts greater than or equal to $N_L$ ($\geq N_L$) in four structural classes of proteins. Here, no. of protein is given in Table 1, the total number of proteins is 428 and $R_c = 8.0$ Å.

number of short-range contacts greater than or equal to $N_S(\geq N_S)$. Here, the value of $N_S$ is 3 and 5, respectively, and the results are also given in Fig. 3. We find that all-α class proteins have a small percentage of residue with the number of long-range contacts greater than or equal to $N_L(\geq N_L)$ (a small value of $P_L$), and a large percentage of residue with the number of short-range contacts greater than or equal to $N_S(\geq N_S)$ (a large value of $P_S$). However, all-β class of proteins have a large percentage of residue with a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ (a large value of $P_L$), and a small percentage of residue with the number of short-range contacts greater than or equal to $N_S(\geq N_S)$ (a small value of $P_S$). As the values of $N_L$ and $N_S$ increase, the percentage values of $P_L$ and $P_S$ decrease. For α/β and α+β classes of proteins, the percentage of residue with the number

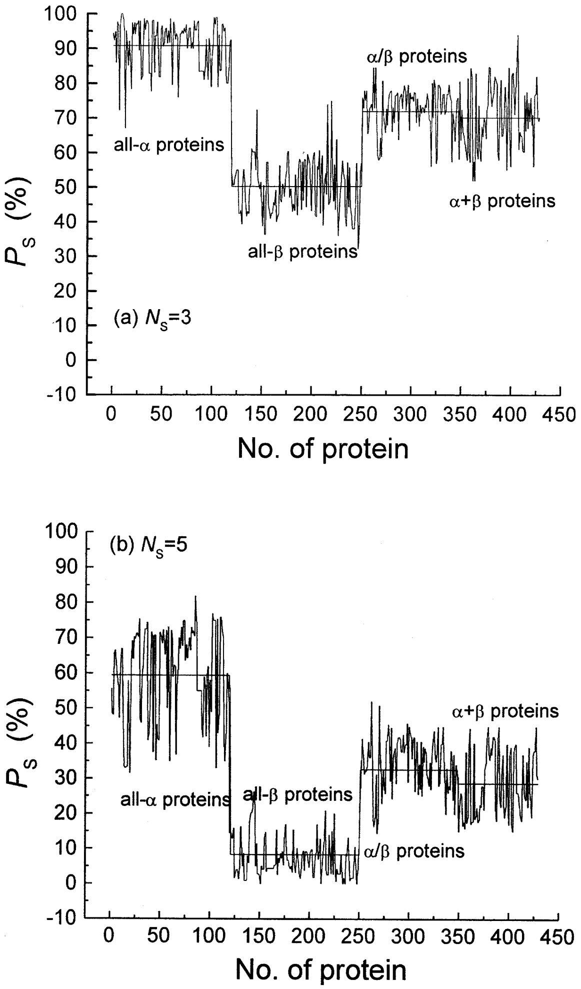

Fig. 3. Distribution of percentage of residues having a number of long-range contacts greater than or equal to $N_S$ ($\geq N_S$) in four structural classes of proteins. Here, no. of protein is given in Table 1, the total number of proteins is 428, and $R_c = 8.0$ Å.

Table 2
The average percentages of residues having the number of long-range contacts greater than or equal to $N_L$ ($\geq N_L$), and the number of short-range contacts greater than or equal to $N_S$ ($\geq N_S$) in four structural classes of proteins; here, $R_c = 8.0$ Å

| | $\bar{P}_L$[a] (%) | $\bar{P}_L$[b] (%) | $\bar{P}_S$[c] (%) | $\bar{P}_S$[d] (%) |
|---|---|---|---|---|
| All-α proteins | 13.3 | 3.91 | 90.8 | 59.4 |
| All-β proteins | 54.8 | 34.7 | 50.2 | 8.72 |
| α/β proteins | 41.4 | 24.6 | 71.9 | 32.6 |
| α+β proteins | 37.0 | 19.8 | 70.1 | 28.7 |

[a] $N_L = 5$.
[b] $N_L = 7$.
[c] $N_S = 3$.
[d] $N_S = 5$.

of long-range contacts greater than or equal to $N_L(\geq N_L)$ and the percentage of residue with the number of short-range contacts greater than or equal to $N_S(\geq N_S)$ are almost the same. The reason may be that for the α/β class of proteins there are approximately alternating α-helices and β-strands, and for α+β class of proteins, α-helices and β-strands that do not mix but tend to segregate into different domain. In Fig. 2, there are two cases of $N_L = 5$ and 7, and in Fig. 3 there are $N_S = 3$ and 5. In Fig. 2, we can find that the percentage $P_L$ is small in all-α class of proteins, and is large for all-β class of proteins. Although the percentage $P_L$ decreases with increasing the value of $N_L$, the relative relationship is almost the same.

We also calculate the percentage value of $P_L$ per protein in four structural classes, and the results are given in Table 2. The average percentage of the all-α class of proteins 13.3% is significantly smaller than that of the all-β class of proteins 54.8% for $N_L = 5$ and $R_C = 8.0$ Å. Over 50% of residues have a number of long-range contacts greater than or equal to 5 ($\geq 5$) in all-β class of proteins. However, there are only 13.2% in all-α class proteins. In the meantime, the average percentage of residues having a number of short-range contacts greater than or equal to 3 ($\geq 3$) in all-α class proteins is 90.8%, which is greater than that in all-β class proteins (50.2%) when $R_C = 8.0$ Å. The percentage decreases with increasing value of $N_S$, especially in all-β class of proteins. $P_S$ decreases from 50.2% to 8.31% in all-β class of proteins, while $P_S$ decrease from only 90.8% to 59.4% in all-α class of proteins when $N_S$ increases from 3 to 5 with $R_C = 8.0$ Å. This may be that α-helices have a short-range compact structure, whilst β-strands have a long-range compact structure.

We also study the effects of $N_L$ on the average percentage of residues. Here, the value of $N_L$ ranges from 1 to 9, whilst $R_C$ ranges from 7.0 to 8.0 Å, and the results are given in Table 3. When $N_L = 1$, the average values of $P_L$ in all-β, α/β and α+β proteins are almost the same, especially in the case of $R_C = 8.0$ Å. With the increasing of $N_L$, the difference between all-β, α/β and α+β proteins increases, especially in the case of $R_C = 7.0$ Å. For all-α class proteins, the average values of $P_L$ decreases abruptly from 67.5% to 0.856% when $N_L$ increases from 1 to 9, and the average values

Table 3
The average percentage $\bar{P}_L$ of residues having the number of long-range contacts greater than or equal to $N_L$ ($\geq N_L$); here, $N_L$ ranges from 1 to 9, and $R_c = 7.0$ and 8.0 Å, respectively

| $N_L$ | $R_C = 8.0$ Å | | | | $R_C = 7.0$ Å | | | |
|---|---|---|---|---|---|---|---|---|
| | all-α proteins (%) | all-β proteins (%) | α/β Proteins (%) | α+β proteins (%) | all-α proteins (%) | all-β proteins (%) | α/β proteins (%) | α+β proteins (%) |
| 1 | 67.5 | 89.9 | 84.4 | 84.3 | 50.2 | 83.5 | 72.7 | 73.0 |
| 3 | 36.1 | 76.4 | 63.7 | 61.9 | 16.9 | 64.1 | 45.5 | 44.5 |
| 5 | 13.3 | 54.9 | 41.4 | 36.9 | 3.35 | 38.1 | 23.2 | 19.8 |
| 7 | 3.92 | 36.7 | 24.9 | 19.8 | 0.70 | 16.4 | 10.0 | 8.18 |
| 9 | 0.856 | 14.2 | 11.8 | 8.01 | 0.0801 | 2.10 | 1.72 | 0.926 |

Table 4
The average percentage $\bar{P}_S$ of residues having the number of short-range contacts greater than or equal to $N_S$ ($\geq N_S$); here, $N_S$ ranges from 2 to 6, and $R_c = 7.0$ Å and 8.0 Å, respectively

| $N_S$ | $R_C = 8.0$ Å | | | | $R_C = 7.0$ Å | | | |
|---|---|---|---|---|---|---|---|---|
| | All-α Proteins (%) | All-β proteins (%) | α/β proteins (%) | α+β proteins (%) | All-α proteins (%) | All-β proteins (%) | α/β proteins (%) | α+β proteins (%) |
| 2 | 98.5 | 97.5 | 98.9 | 97.8 | 95.9 | 77.1 | 88.0 | 88.1 |
| 3 | 90.1 | 50.4 | 72.1 | 70.4 | 84.1 | 33.5 | 59.0 | 57.1 |
| 4 | 80.1 | 28.4 | 55.6 | 53.2 | 72.4 | 16.9 | 45.4 | 41.6 |
| 5 | 59.1 | 9.00 | 32.6 | 28.7 | 52.0 | 5.75 | 25.6 | 22.6 |
| 6 | 45.3 | 4.96 | 22.5 | 18.5 | 39.5 | 3.58 | 18.3 | 14.9 |

of $\bar{P}_L$ decrease from 89.9% to 14.2% for all-β proteins in the case of $R_C = 8.0$ Å.

We investigate the average percentage of residues with a number of short-range contacts greater than or equal to $N_S(\geq N_S)$, where $N_S$ ranges from 2 to 6, and the results are shown in Table 4. The average percentages $\bar{P}_S$ are almost the same in the different structural classes for $N_S = 2$, especially for a large $R_C = 8.0$ Å. When $N_S$ increases, the average percentage $\bar{P}_S$ decreases, especially for all-β class proteins. For example, $\bar{P}_S$ decreases from 98.5% to 45.3% for all-α class proteins, however, $\bar{P}_S$ decreases abruptly from 97.5% to 4.96% for all-β class of proteins. This means that all-β class of proteins have a very small percentage of residues with a large number of short-range contacts. The situation of α/β and α+β proteins are almost the same. The average percentage $\bar{P}_S$ of α/β proteins is a little bit more than that of α+β proteins. We can assume that the features of α/β proteins should be similar to α+β proteins.

Above all, by studying all of the 428 global proteins, we conclude that the average percentage of long-range contacts for all-α class of proteins is greatly lower than that for all-β class of proteins, and the situation of α/β and α+β proteins is almost identical. These results can help us determine the globular protein structure. That is, that an α-helix can easily form short-range contacts, and a β-strand can easily form long-range contacts.

### *3.2. The percentage of long-range contacts for different amino acids residues in four structural classes*

In general, the amino acids are subdivided into two kinds of residues: hydrophobic (H) and polar

Table 5
The percentage of amino acid residues having a number of long-range contacts greater than or equal to $N_L$ ($\geq N_L$) the for 20 amino acid residues under different conditions

| | $R_C = 8.0$ Å | | $R_C = 7.0$ Å | |
|---|---|---|---|---|
| | $N_L = 5$ (%) | $N_L = 7$ (%) | $N_L = 5$ (%) | $N_L = 7$ (%) |
| Leu | 42.6 | 25.5 | 23.3 | 10.5 |
| Val | 57.6 | 38.7 | 36.7 | 18.0 |
| Ile | 54.0 | 35.3 | 33.7 | 15.3 |
| Met | 43.4 | 25.6 | 23.4 | 10.1 |
| Phe | 45.4 | 27.5 | 26.7 | 11.2 |
| Tyr | 49.7 | 30.5 | 31.5 | 13.1 |
| Cys | 67.2 | 45.0 | 41.6 | 24.0 |
| Trp | 46.7 | 28.2 | 31.0 | 12.0 |
| Ala | 42.6 | 23.6 | 22.4 | 10.2 |
| Gly | 43.8 | 28.1 | 25.9 | 10.9 |
| Thr | 44.5 | 26.2 | 27.9 | 11.1 |
| His | 35.7 | 20.0 | 18.1 | 7.28 |
| Glu | 21.9 | 11.6 | 11.1 | 4.90 |
| Gln | 30.1 | 17.7 | 17.3 | 6.57 |
| Asp | 26.2 | 14.3 | 14.2 | 5.02 |
| Asn | 33.0 | 16.1 | 17.2 | 5.85 |
| Lys | 25.4 | 13.0 | 13.7 | 4.80 |
| Ser | 37.8 | 20.7 | 21.5 | 9.56 |
| Arg | 31.2 | 17.1 | 17.7 | 6.69 |
| Pro | 35.1 | 19.0 | 17.3 | 4.59 |

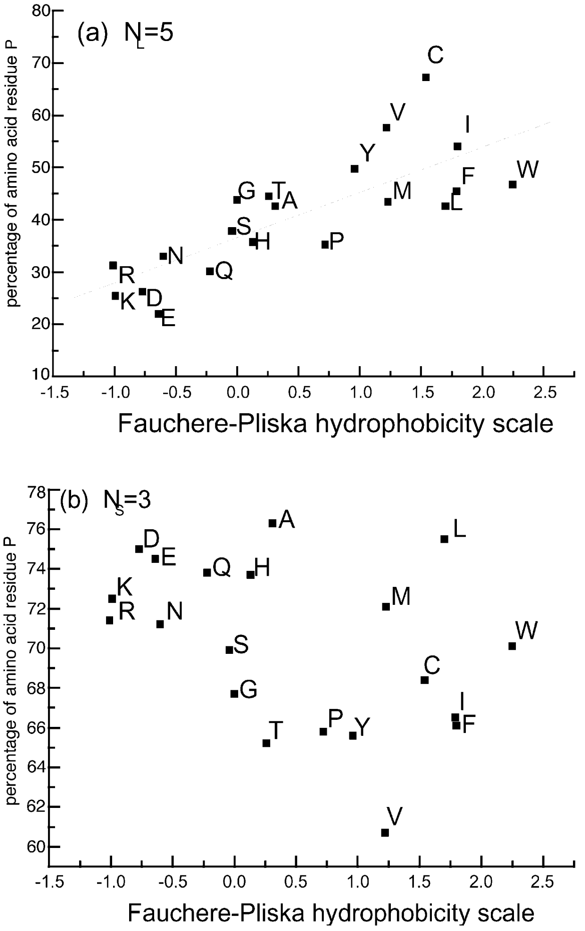

Fig. 4. Percentage of residues having a large number of long-range contacts $(P_L)$ and short-range contacts $(P_S)$ vs. Fauchere–Pliska hydrophobicity scale of amino acid residues. (a) $P_L$, $N_L=5$, and $R_C=8.0$ Å; and (b) $P_S$, $N_S=3$, and $R_C=8.0$ Å.

(P), and there are different amino acid residue–residue interactions. In principle, these interactions could be studied at a more fundamental level by using the potentials for each atom. However, for many applications, an amino acid-based approach is still preferred, because a calculation involving the pairwise interactions between thousands of atoms in a given protein is often not traceable with currently available computational power. These energies have been obtained by statistical methods from databases of protein native structures. Tanaka and Scheraga first advanced the idea in 1975 [1]. A comprehensive analysis was given by Miyazawa and Jernigan [15,16]. On the other hand, all of the 20 amino acid residues have a different ability to form a globular structure [18]. If the residue is hydrophobic, one may place itself quite easily inside the protein, and the percentage of residues having a large number of long-range contacts may be large. In general, the average number of long-range contacts per residue may be also large for the hydrophobic residue.

We calculate the percentage of amino acid residues having a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ for all of the 20 amino acids under different conditions. Here, $N_L=$ 5 and 7, and $R_c=8.0$ and 7.0 Å, and the results are given in Table 5. We find that the percentages per residue having a large number of long-range contact are large for Leu, Val, Ile, Met, Phe, Tyr, Cys, Trp, Ala, Gly and Thr. Those residues may be hydrophobic residues and easily placed in the

Table 6
The percentage of amino acid residues having a number of long-range contacts greater than or equal to $N_S$ $(\geq N_S)$ for the 20 amino acid residues under different conditions

| | $R_C=8.0$ Å | | $R_C=7.0$ Å | |
|---|---|---|---|---|
| | $N_S=3$ (%) | $N_S=5$ (%) | $N_S=3$ (%) | $N_S=5$ (%) |
| Leu | 75.5 | 37.7 | 65.7 | 34.0 |
| Val | 60.7 | 37.1 | 50.3 | 27.5 |
| Ile | 66.1 | 36.9 | 53.4 | 30.3 |
| Met | 72.1 | 35.6 | 64.1 | 33.8 |
| Phe | 66.5 | 35.0 | 56.2 | 32.2 |
| Tyr | 65.6 | 29.1 | 52.8 | 23.1 |
| Cys | 68.4 | 39.3 | 55.0 | 21.4 |
| Trp | 70.1 | 33.1 | 60.5 | 31.8 |
| Ala | 76.3 | 33.7 | 67.0 | 31.1 |
| Gly | 67.7 | 24.2 | 53.1 | 15.5 |
| Thr | 65.2 | 26.5 | 51.5 | 19.5 |
| His | 73.7 | 30.2 | 61.7 | 29.0 |
| Glu | 74.5 | 25.9 | 63.9 | 22.6 |
| Gln | 73.8 | 24.9 | 62.4 | 23.4 |
| Asp | 75.0 | 23.9 | 60.1 | 20.1 |
| Asn | 71.2 | 22.9 | 57.5 | 19.9 |
| Lys | 72.5 | 25.2 | 60.8 | 24.0 |
| Ser | 69.9 | 23.7 | 54.9 | 18.1 |
| Arg | 71.4 | 27.8 | 59.7 | 24.8 |
| Pro | 65.8 | 18.7 | 47.5 | 10.5 |

Table 7
The percentage of amino acid residues having a number of long-range contacts greater than or equal to $N_L$ ($\geq N_L$) for the 20 amino acid residues in four structural classes of proteins; here, $R_C = 8.0$ Å, and $N_L = 4$

| | All-α Proteins (%) | All-β proteins (%) | α/β proteins (%) | α+β proteins (%) |
|---|---|---|---|---|
| Leu | 28.1 | 77.0 | 56.9 | 59.7 |
| Val | 30.7 | 84.0 | 75.5 | 71.5 |
| Ile | 32.8 | 80.1 | 67.1 | 71.8 |
| Met | 29.5 | 75.7 | 56.3 | 58.0 |
| Phe | 26.8 | 76.9 | 63.2 | 63.6 |
| Tyr | 36.3 | 83.8 | 54.5 | 67.7 |
| Cys | 53.8 | 87.2 | 87.7 | 77.2 |
| Trp | 31.9 | 83.3 | 61.9 | 54.0 |
| Ala | 28.6 | 74.7 | 65.7 | 60.1 |
| Gly | 32.6 | 68.7 | 61.7 | 47.3 |
| Thr | 30.6 | 74.9 | 61.6 | 51.8 |
| His | 26.3 | 60.9 | 56.1 | 52.5 |
| Glu | 17.0 | 50.1 | 30.4 | 30.6 |
| Gln | 13.6 | 69.5 | 41.5 | 38.9 |
| Asp | 16.7 | 53.6 | 37.1 | 35.7 |
| Asn | 22.1 | 63.8 | 47.7 | 43.8 |
| Lys | 18.6 | 60.7 | 36.7 | 33.5 |
| Ser | 29.9 | 65.4 | 52.9 | 51.6 |
| Arg | 18.6 | 70.5 | 48.7 | 34.6 |
| Pro | 31.7 | 55.8 | 51.2 | 46.5 |

inside of globular proteins, especially for Cys, Val, Ile, Tyr, Trp and Phe. Our results agree with the free energies of transfer of the amino acids from water to non-polar environments by Fauchere and Plska [22]. In Fig. 4a, we plot the percentage of residues having a large number of long-range contacts vs. the Fauchere–Pliska hydrophobicity scale (FPH), and find that the value of $\bar{P}_{L,\alpha}$ increases with increasing FPH value, where $P_{L,\alpha}$ is the average of α amino acid residue. Here, we only give the relationship in the case of $N_L = 5$ and $R_C = 8.0$ Å. In fact, a similar relationship can also be given in the other cases. The relationship between $P_{L,\alpha}$ and the Fauchere–Pliska hydrophobicity scale (FPH) is expressed approximately as (in %)

$$\bar{P}_{L,\alpha} = a + b \times \text{FPH} \quad (\alpha = \text{Leu, Val, \ldots, Pro}) \qquad (4)$$

In this case, $a = 36.4$, and $b = 8.61$.

Except for Cys, the relative deviation between the Fauchere–Pliska hydrophobicity scale and our expression of $\bar{P}_{L,\alpha}$ is small. From Table 6, we also plot the percentage of residues having large number of short-range contact vs. the Fauchere–Pliska hydrophobicity scale (FPH) in Fig. 4b. In Fig. 4b, a similar relationship is not found.

We investigate the percentages of residues having a number of long-range contacts greater than or equal to 4 ($\geq 4$) for the 20 amino acids in four structural classes of proteins, and the results are given Table 7. In this paper, we choose $N_L$ at random as a last resort. Of course, if $N_L$ is too large, our percentages may become zero for all proteins, and if $N_L$ is too small, our percentages may become 100% for all proteins (see Tables 3 and 4). Therefore, we choose $N_S$ from 4 to 7, and $N_L$ from 2 to 6. In fact, here we discuss mainly the relative ability to forming long-range (short-range) contacts for residues (proteins). The effects of the cut-off of 4 (or 5/6) residues on our relative results in four structural classes of proteins is insignificant. In Table 7, the percentage of residues having a large number of long-range contacts in all-α class of proteins is smaller than that in the other type of proteins, especially in all-β class of proteins. In all-α class of proteins, the top-most three residue is Cys, Tyr and Ile, and Gln, Asp and Glu have a small percentage. In all-β class of proteins, the top-most three residue is Cys, Val and Tyr, and Glu has a minimum value of $P_L$. However, the value of $P_L$ in all-β class of proteins is 2–3 times value in all-α class of proteins. In the α+β class of proteins and α/β class of proteins, the top-most three residues are the same, i.e. Cyr, Val and Ile. Through our investigation of the percentage of residues having a large number of short-range contacts, we can determine the globular structure of proteins clearly. In the meantime, we can provide some insights into the structure difference in four classes, and the importance of α-helical and β-strand in stability of protein structure.

## Acknowledgments

This research was financially supported by NSFC (Nos. 29874012, 20174036 and 20274074), Natural Science Foundation of Zhejiang Province (No. 10102) and the Special Funds for Major State

Basic Research Projects (G1999064800). The authors thank the referees for their critical reading of the manuscript and their useful ideas.

## References

[1] S. Tanaka, H.A. Scheraga, Model of protein folding: inclusion of short-, medium-, and long-range interactions, Proc. Natl. Acad. Sci. 72 (1975) 3802–3805.

[2] I. Bahar, M. Kaplan, R.L. Jernigan, Short-range conformational energies, secondary structure propensities, and recognition of correct sequence–structure matches, Proteins 29 (1997) 292–308.

[3] C. Zhang, J.L. Cornette, C. Delisi, Consistency in structural energetics of protein folding and peptide recognition, Protein Sci. 6 (1997) 1057–1064.

[4] D.S. Gottfried, E. Haas, Nonlocal interactions stabilize compact folding intermediates in reduced unfolded bovine pancreatic trypsin inhibitor, Biochemistry 31 (1992) 12353–12362.

[5] M.M. Gromiha, S. Selvaraj, Influence of medium and long range contacts in TIM barrel proteins, J. Biol. Phys. 23 (1997) 209–217.

[6] M.M. Gromiha, S. Selvaraj, Importance of long range interactions in protein folding, Biophys. Chem. 77 (1999) 49–68.

[7] M.M. Gromiha, S. Selvaraj, Important amino acid properties for determining the transition state structures of two-state protein mutants, FEBS Lett. 526 (2002) 129–134.

[8] M.M. Gromiha, S. Selvaraj, Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long range order to folding rate prediction, J. Mol. Biol. 310 (2001) 27–32.

[9] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–554.

[10] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, et al., The Protein Data Bank: a computer-based archival file for macromolecular structures, J. Mol. Biol. 112 (1977) 535–542.

[11] T.J.P. Hubbard, B. Ailey, S.E. Brenner, A.G. Murzin, C. Chothia, SCOP: a structural classification of proteins database, Nucl. Acids Res. 27 (1999) 254–257.

[12] C.A. Orengo, F.M.G. Pearl, J.E. Bray, A.E. Todd, A.C. Martin, L. Loconte, et al., The CATH database provides insights into protein structure/function relationships, Nucl. Acids Res. 27 (1999) 275–279.

[13] M.M. Gromiha, Prediction of secondary structures in globular and membrane proteins, Recent Res. Dev. Protein Eng. 2 (2001) 161–178.

[14] M.M. Gromiha, S. Selvaraj, Inter-residue interactions in the structure, folding and stability of proteins, Recent Res. Dev. Biophys. Chem. 1 (2000) 1–14.

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

[16] S. Miyazawa, R.L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, Macromolecules 18 (1985) 534–552.

[17] G.M. Crippen, P.A. Kollman, Distance geometry methods for protein structure calculations, Biopolymers 18 (1979) 939–957.

[18] J. Zhouting, Z. Linxi, C. Jin, X. Agen, Z. Delu, Effect of amino acid on forming residue–residue contacts in proteins, Polymer 43 (2002) 6037–6047.

[19] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, M. Hecht, Protein design by binary patterning of polar and nonpolar amino acids, Science 262 (1993) 1680–1682.

[20] K.F. Lau, K.A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins, Macromolecules 22 (1989) 3986–3997.

[21] K.A. Dill, S. Bromberg, S. Yue, K. Fiebig, K.M. Yee, D.P. Thomas, et al., Principles of protein folding—a perspective from simple exact models, Protein Sci. 4 (1995) 561–602.

[22] J.L. Fauchere, V. Pliska, Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino-acid amides, Eur. J. Med. Chem. 18 (1983) 369–375.