

二维紧密蛋白质链折叠序列的特性研究*

章林溪**

(温州师范学院物理系 温州 325027)

摘要 采用完全计数法,研究了二维紧密蛋白质链在不同 HP 序列时的构象性质,特别是具有唯一基态能量的折叠序列的性质. 对于具有 N 个单体的紧密蛋白质链,发现有一定比例的序列为折叠序列. 在这些折叠序列中,疏水基团(H)的数目比亲水基团(P)多 20%,并同 200 种真实蛋白质分子的疏水基团和亲水基团的结果进行了比较. 对于不同的折叠序列,根据序列中其疏水基团的数目,把具有相同疏水基团数目的序列归在同一类,发现这样的序列在总的序列中的相对含量满足高斯分布. 同时还对序列中 H(或者 P)团簇大小及其分别进行了研究,发现折叠序列与无规随机序列不同. 还研究了不同折叠序列在不同链长时的比热情况,发现其相转变温度 T_c 主要与链长有关,与折叠序列无关.

关键词 二维紧密蛋白质, 完全计数法, 折叠序列, 比热, 相转变温度

球状蛋白质是一类非常重要的蛋白质,其特点是分子内部非常紧密,其相互作用主要来源于氨基酸之间的紧密接触对(residue-residue contact),同时构成蛋白质分子的氨基酸主要有 20 种,这样对于具有 N 个单体的蛋白质,其氨基酸序列数目有 20^N 种,因此理论上研究其性质极其复杂. 后来人们就采用比较简单的疏水(H)/亲水(P)紧密蛋白质链模型来研究球状蛋白质性质^[1-5]. 由于蛋白质分子要求具有基态的能量,同时还必须是简并度为 1,因此不是所有序列下的紧密蛋白质链都能代表球状蛋白质分子,把这些能够折叠成基态能量且具有简并度为 1 的序列称为折叠序列. 研究这些折叠序列的性质能够帮助我们了解为什么在自然界中蛋白质分子的序列数目远远小于 20^N ,同时还能帮助我们了解为什么自然界中存在着不同氨基酸序列的蛋白质而具有相似的空间结构. Lau 和 Dill 首先研究了链长为 9 (10 个单体)的紧密蛋白质链在所有序列下的构象性质^[2],但由于链长太短,其 HP 序列也非常少(2^{10}),因此得到的结果也是非常有限. 本文采用二维的 HP 格点模型,研究具有 N 个单体($N \leq 22$)的紧密蛋白质链的构象及其 HP 序列性质,通过对这些折叠序列中所包含的特性研究,能够帮助我们对蛋白质氨基酸序列性质的了解.

1 模型和计算方法

本文采用二维的正方形格点模型. 在这个模型中,紧密蛋白质链的键角分别可取, $\pm \pi/2, \pi$. 在计算时,考虑排斥体积效应,并排除两个完全相同的构象. 在考虑相互作用时采用紧密蛋白质链的 HP 格点模型,其体系的 Hamiltonian 能量为^[1-6]

$$E = -k_B T \sum_{i < j} S_i S_j \delta(r_{ij}, 1) \quad (1)$$

第 i 个基团如是亲水基团(P),则 $S_i = 0$; 第 i 个基团如是疏水基团(H),则 $S_i = 1$; r_{ij} 表示基团 i 和基团 j 的距离. 如果 $r_{ij} = 1$ (这里正方形的边长为 1),则 $\delta(r_{ij}, 1) = 1$,即形成一个紧密接触对;如果 $r_{ij} \neq 1$,则 $\delta(r_{ij}, 1) = 0$. 这里假定形成一个紧密接触对的两个疏水基团之间的相互作用能为 $-k_B T$. 为了计算方便,习惯上取 $k_B = 1$. 给定一个 HP 序列,计算其最小能量以及其对应的构象数目. 如果其最小能量对应的构象数目为 1,这表明该序列为折叠序列.

对于给定的序列,其体系的比热为^[7,8]

$$C_V = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2} \quad (2)$$

这里 $\langle \rangle$ 表示对所有构象的平均,即

$$\langle E \rangle = \frac{\sum_i E_i \exp(-E_i/k_B T)}{\sum_i \exp(-E_i/k_B T)} \quad (3)$$

* 2004-03-30 收稿,2004-05-02 修稿;国家自然科学基金(基金号 20174036,20274040)和浙江省自然科学基金(基金号 R404047)资助项目;

** 通讯联系人, E-mail: lxzhang@hzenc.com

$\sum_i$ 表示对所有构象的求和。

2 结果与讨论

2.1 折叠序列内部的疏水基团(H)/亲水基团(P)分布

对于链长为 $N-1$ 的紧密蛋白质链, 有 2^N 个不同的疏水基团(H)/亲水基团(P)序列. 计算每个序列的最低能量以及其对应的构象数目. 为了减少工作量, 只需要计算能量比较低的构象数目. 比如对于 $N=16$, 只需要计算紧密接触对数目为 $t=8$ 和 $t=9$ 的构象, 对应的最低能量分别为 $-4, -5, -6, -7, -8$, 和 -9 (这里能量以 $k_B T$ 为单位). 根据其简并度确定是否为折叠序列, 即简并度为 1 的序列为折叠序列. 在这些序列中, 疏水基团(H)和亲水基团(P)的比例是不一样的, 正如每个蛋白质中氨基酸的含量是各不相同的. 但对大量蛋白质的氨基酸含量分析, 发现其 20 种氨基酸的含量是基本不变的. 根据这个特点, 计算了这些折叠序列中疏水基团与亲水基团的个数之比, 即 N_H/N_P , 结果见图 1, 这里 N_H 和 N_P 表示在所有的折叠序列中疏水基团与亲水基团的个数. 发现比值 N_H/N_P 对于较小的 N 值, 存在涨落. 当 $N > 16$ 时, 比较接近 $N_H/N_P = 1.197$. 实际上, 对于比较小的 N 值, 折叠序列的数目也是非常有限的, 如对于 $N=13$, 折叠序列的数目只有 173, 这样最后的结果不是特别可靠. 但是对于较大的 N 值, 折叠序列的数目就比较大, 如对于 $N=22$ 折叠序列的数目有 97478, 这样统计出来的结果就比较可靠的.

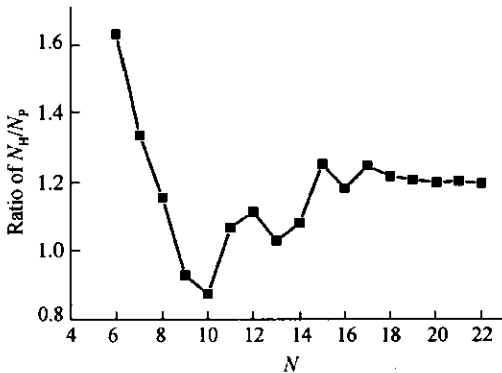


Fig. 1 Ratio of N_H/N_P versus number of monomers N
Here N_H (N_P) means total number of hydrophobic (H) (hydrophilic, P) in folding sequences

为了同真实的蛋白质分子比较, 还对 200 种蛋白质分子的 20 种氨基酸的相对含量进行了分

析计算, 这里的 200 种蛋白质分子的代码已在文献^[9]中给出. 同时根据 20 种氨基酸的 Fauchere-Pliska 疏水性(FPH)^[10] 的实验数值, 把 $FPH > 0$ 的氨基酸定义为疏水基团(H), $FPH < 0$ 的氨基酸定义为亲水基团(P). 发现其 N_H/N_P 值为 1.145. 虽然从二维紧密蛋白质链计算得到的 N_H/N_P 数值与真实的三维蛋白质分子的 N_H/N_P 数值这二者不能直接比较, 但是二者的结果还是能够说明二维紧密蛋白质链的折叠序列与真实蛋白质分子有着相似的性质.

对于这些折叠序列, 每个序列中疏水基团和亲水基团的含量各不相同, 我们已经对较短的二维紧密蛋白质链进行了研究 ($N=12, 14$). 发现疏水基团的含量在 40% ~ 60% 出现的序列几率最大^[6]. 这里对较长的二维紧密蛋白质链进行了研究, 结果见图 2. 发现其分布满足很好的高斯分布. 对于其他 $N=18, 19, 20$ 和 21 同样有着非常相似的高斯分布.

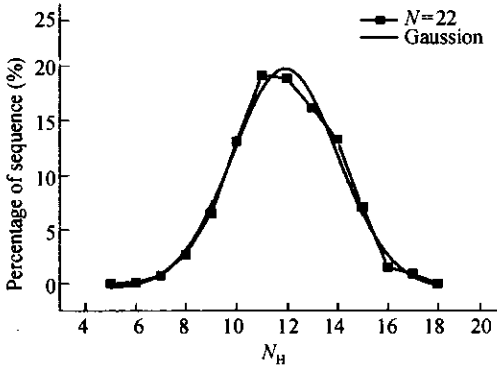


Fig. 2 Distribution of folding sequences with hydrophobic amino acid residues N_H in all folding sequences

2.2 折叠序列中疏水(H)/亲水(P)团簇的特性

对于 HP 格点模型, 一个折叠序列中只有 H 和 P 二个基团. 如序列 HHPPPHPPHPPHPPHPPH. 如果相邻二个基团相同, 就把它作为一个整体. 这样上面的这个序列中, 团簇大小 $S=4$ 的个数为 1, 团簇大小 $S=3$ 的个数为 1, 团簇大小 $S=2$ 的个数为 4, 团簇大小 $S=1$ 的个数为 5 (这里不区 H 和 P), 因而就可以知道其分布几率 $P(S)$, 这个方法与分析 DNA 序列的团簇非常相似^[11], 最后结果见图 3. 为了比较, 还给出了无规序列的分布几率. 发现二者有相似的分布, 即 $P(S)$ 都可以表示为

$$P(S) \sim \exp(-aS) \tag{4}$$

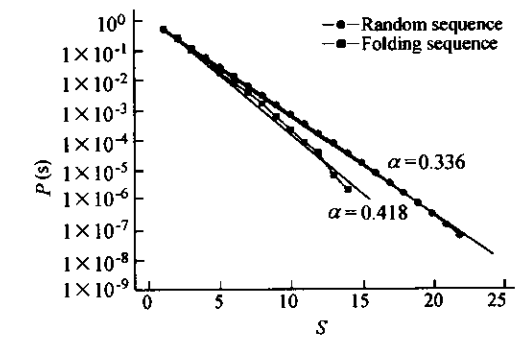


Fig. 3 Probability of cluster with size S in random sequence and folding sequence, here $N = 22$

但是二者的系数 α 是不一样的, 折叠序列 ($\alpha = 0.418$) 大于无规序列 ($\alpha = 0.336$). 另外最大的 S 值也不一样, 折叠序列 ($S_{\max} = 14$) 远小于无规序列 ($S_{\max} = 22$).

同时还计算了疏水(H)/亲水(P)团簇的平均大小, 结果见图 4. 发现无规序列的疏水(H)/亲水(P)团簇的平均大小为 $\bar{S} = 2.000$; 对于折叠序列,

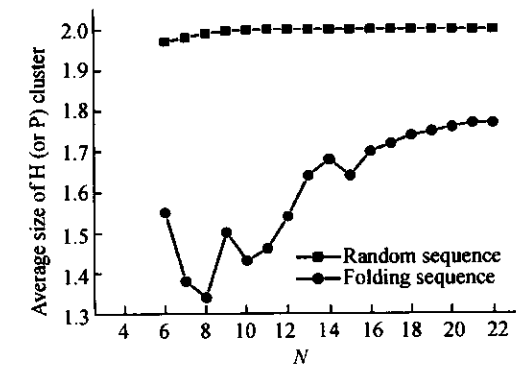


Fig. 4 Average size versus number of monomers N in random sequence and folding sequence

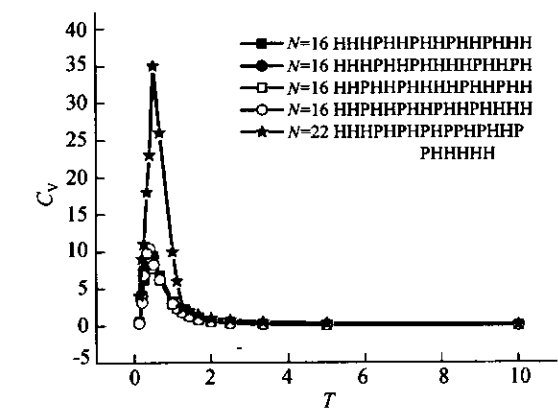


Fig. 5 Heat capacity C_V versus temperature T for different sequences with $N = 16$, and $N = 22$, respectively

疏水(H)/亲水(P)团簇的平均大小为 $\bar{S} = 1.771$. 由此可见, 折叠序列与无规序列确实不一样, 折叠序列有其特殊的性质.

2.3 具有折叠序列的紧密蛋白质链的比热

紧密蛋白质链的比热一方面与其分子的构象与相互作用有关, 另一方面其与相转变有密切关系, 同时它又可以通过实验的方法测得, 因此对它的研究非常重要. 这里计算了在不同温度下相同长度不同序列和不同长度的情况下的比热, 结果见图 5. 在图 5 中发现, 对于不同的序列和长度, 都有一个最大值. 这个比热最大值所对应的温度就是相转变温度 T_c , 这里发现对于 $N = 16$ 的 4 个序列, T_c 非常接近, 即 $T_c = 0.4$; 对于 $N = 22$ 的序列, $T_c = 0.5$, 同时这里的折叠序列是随机选取的, 由此可见, 相转变温度主要与链长有关, 与折叠序列无关. 这些研究能够帮助对蛋白质序列以及蛋白质结构多样性和复杂性的了解.

REFERENCES

1 Chan H S, Dill K A. *Macromolecules*, 1989, 22: 4559 ~ 5567

2 Lau K F, Dill K A. *Macromolecules*, 1989, 22: 3986 ~ 3997

3 Yue K, Fiebig K M, Thomas P D, Chan H S, Shakhovich E I, Dill K A. *Proc Natl Acad Sci U S A*, 1995, 92: 325 ~ 329

4 Sun T T, Zhang L X, Chen J, Shen Y. *J Chem Phys*, 2004, 120: 5469 ~ 5475

5 Chen Jin(陈进), Zhang Linxi(章林溪), Jiang Zhouting(姜舟婷), Zhao Delu(赵得禄). *Acta Polymerica Sinica(高分子学报)*, 2003, (4): 577 ~ 580

6 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): 515 ~ 519

7 Lam P M. *Phys Rev B*, 1987, 36: 6988 ~ 6992

8 Zhang L X, Sun T T. *Polymer*, 2004, 45: 3547 ~ 3554

9 Jiang Z, Zhang L, Chen J, Xia A, Zhao D. *Polymer*, 2004, 45: 609 ~ 622

10 Fauchere J L, Pliska V. *Eur J Med Chem*, 1983, 18: 369 ~ 75

11 Sun T T, Zhang L X, Chen J, Jiang Z T. *Chaos, Solitons and Fractals*, 2004, 20: 1075 ~ 1088

A STUDY OF FOLDING SEQUENCE OF TWO-DIMENSIONAL COMPACT PROTEINS

ZHANG Linxi

(*Department of Physics, Wenzhou Normal College, Wenzhou 325027*)

Abstract Statistical properties of folding sequences of two-dimensional compact protein were investigated by using the enumeration calculation method. The number of folding sequences, S_N , may be expressed as $S_N \propto 2^N$. Here 2^N means the total number of sequences for N -monomer compact chains. For folding sequences, $H_H/N_P \approx 1.197$, here N_H and N_P represent the total number of hydrophobic (H) and hydrophilic (P) groups in all folding sequences. Comparisons with real proteins were also made. Distributions of folding sequences in all sequences according to the number of hydrophobic amino acid residues were investigated, and they had the form of the Gaussian distribution. The cluster size and its distributions were also studied, and there existed a difference between folding sequences and random sequences. The coil-to-globule transition temperature T_c , estimated from the location of the peak on the heat capacity, depended on the chain length and was almost sequence independent.

Key words Two-dimensional compact protein, Enumeration calculation method, Folding sequence, Heat capacity, Transition temperature T_c