

二维蛋白质模型分子在折叠过程中的构象研究*

陈 进 董志君 章林溪 姜舟婷

赵得禄

(浙江大学物理系 杭州 310028)

(中国科学院化学所分子科学中心 北京 100080)

摘 要 蛋白质在折叠过程中其构象要发生明显变化.采用精确计数法,计算了在蛋白质折叠的不同阶段其尺寸大小及其分布情况.发现在折叠的初期,分子的尺寸比较大,其分布也比较宽.在折叠的后期,分子的尺寸比较小,其分布比较窄.不同的氨基酸序列,其分子尺寸的分布也不同.对于可折叠的氨基酸序列,在平均尺寸大小附近出现的几率特别大.同时还计算了比值 S_N/D_N .这里 S_N 为可设计序列数目(The number of designing sequences), D_N 为可设计构象数目(The number of designable conformations),并有关系

$$-1.684 + 0.3255N \leq S_N/D_N \leq -0.8664 + 0.3125N \quad (N \geq 13)$$

通过这些研究以提高对折叠过程的认识.

关键词 二维蛋白质模型分子, HP 格点模型, 精确计数法, 均方迴转半径, 分布函数, 折叠

蛋白质分子的折叠过程是指蛋白质分子从一般的状态变化到基态的复杂过程.它能使我们了解氨基酸序列是如何决定蛋白质分子结构,预测其结构及结构所表现出来的蛋白质分子的性质^[1-4].在这个过程中氨基酸与氨基酸紧密接触对(Residue-residue contact)的相互作用起着十分重要的作用^[5-7].人们用不同的方法对它进行了研究,如 Dill 等首先对紧密高分子链的构象进行了初步研究^[2,3],后来又用 Monte Carlo 方法研究了蛋白质在不同的氨基酸序列下的折叠过程^[4],有人还用图论的方法研究了生物大分子在折叠过程中的热力学性质的变化^[8].实际上在蛋白质的折叠过程中,其分子大小也要发生明显变化.同时不同序列的蛋白质分子,其分子大小及其分布情况也是不同的.这些问题非常有意义,它能够帮助我们更好地了解蛋白质的折叠问题.这里采用精确计数法^[9-11],研究二维蛋白质模型分子在折叠的不同阶段其尺寸大小以及分布情况,了解其构象变化情况.

1 计算方法

蛋白质分子在其折叠过程中,蛋白质的相互作用起着决定性的作用,其相互作用主要来源于氨基酸-氨基酸紧密接触对的相互作用,而这种相互作用有别于一般分子之间的相互作用.其体系

的哈密顿(Hamiltonian)能量为^[1-9]

$$E = - \sum_{i < j} E_{S_i - S_j} \delta(r_{ij}, a_0) \quad (1)$$

其中 $E_{S_i - S_j}$ 表示氨基酸 S_i 和 S_j 之间的相互作用能量, $\delta(r_{ij}, a_0)$ 是判断氨基酸 S_i 和 S_j 是否形成一个紧密接触对的条件,这里 a_0 是形成紧密接触对的最大距离,一般取 0.65 nm ^[5-7].对于真实的蛋白质分子,氨基酸 S_i 和 S_j 之间的相互作用能量 $E_{S_i - S_j}$ 存在 $\frac{1}{2} \times 20 \times 20$ 种可能,这样理论计算非常困难.后来人们采用简单的格点模型即 HP 格点模型.在这个模型中,只存在亲水基团(P)和疏水基团(H)二种基团,使得氨基酸 S_i 和 S_j 之间的相互作用能量 $E_{S_i - S_j}$ 从 200 种可能减少到 3 种(或者 2 种)可能,计算大大简化.在蛋白质分子的折叠过程中,体系的能量是慢慢减小的,因此研究在不同紧密接触对数目时的构象性质具有十分重要的作用. Dill 等已经研究了在不同紧密接触对数目时的构象数目,但还没有研究在折叠过程中其分子大小及其分布的变化情况.这里将用精确计数法来计算二维蛋白质模型分子的均方迴转半径及其分布情况.定义 $\langle S^2 \rangle_t$ 是紧密接触对数目为 t 时的均方迴转半径,即

* 2002-12-05 收稿, 2003-01-28 修稿; 国家自然科学基金(基金号 29874021, 20174036 和 20274040)和浙江省自然科学基金(基金号 101002)资助项目

$$\langle S^2 \rangle_i = \sum_{i=1}^{N+1} \sum_{j=1}^{N_i} \frac{[(S_{i,x} - S_{0,x})^2 + (S_{i,y} - S_{0,y})^2]}{N_i} \quad (2)$$

这里 N 是键长, N_i 为紧密接触对数目 i 时的总构象数目, $S_{i,x}$ 和 $S_{i,y}$ 分别表示第 i 个原子在 x, y 方向上的坐标分量, $S_{0,x}$ 和 $S_{0,y}$ 表示重心在 x, y 方向上的坐标分量. 在计算均方回转半径的分布函数 $P(S)$ 时, 如果 $P(S)$ 表示回转半径在 $S - S + \Delta S$ 的几率, N_s 为回转半径在 $S - S + \Delta S$ 的区间的构象数目, 即有^[9]

$$\int_S^{S+\Delta S} P(S) dS = \frac{N_s}{N_i} \quad (3)$$

如果 ΔS 足够小, 同时 $P(S)$ 为单调函数, 有

$$\int_S^{S+\Delta S} P(S) dS = P\left(S + \frac{\Delta S}{2}\right) \Delta S \quad (4)$$

故有

$$P\left(S + \frac{\Delta S}{2}\right) = \frac{N_s}{N_i \Delta S} \quad (5)$$

通过计算 N_s , 根据公式(5)可得分布函数 $P(S)$. 这里取 $\Delta S = 0.05 \langle S^2 \rangle^{1/2}$, $\langle S^2 \rangle^{1/2}$ 为均方根回转半径.

2 结果与讨论

2.1 具有相同紧密接触对数目的分子大小

用精确计数法计算了链长从 7 ~ 16 的二维蛋白质模型分子在不同紧密接触对数目(t)时的均方回转半径 $\langle S^2 \rangle_i$. 对于链长为 16 的二维蛋白质模型分子, 存在着 2155667 种构象, 同时二种构象不能完全相同, 因此计算链比较长的蛋白质模型分子时, 工作量非常大. 这里我们只计算到 $N = 16$ 的二维蛋白质模型分子的大小, 结果见图 1. 在图 1 中给出了紧密接触对数目从 $t = 0$ 到 $t = 4$ 时回转半径的特征比 $\langle S^2 \rangle_i / Nl^2$ 与链长 N 的关系, 发现 $\langle S^2 \rangle_i / Nl^2$ 与链长 N 有很好的线形关系

$$\langle S^2 \rangle_i / Nl^2 = a + bN \quad (6)$$

这里 l 为键长, a 与紧密接触对数目(t)有关, 而 b 与紧密接触对数目(t)无关. 对于具有相同数目的紧密接触对的二维蛋白质模型分子, $\langle S^2 \rangle_i / Nl^2$ 与链长 N 有相同的线形关系(斜率几乎相同). 这个与在 θ 状态时的高分子链大小相似, 但与考虑排斥体积时的高分子链不同. 可能的原因是在这里讨论的是具有相同数目的紧密接触对的蛋白质模型分子的均方回转半径, 它反映了蛋白质分子

在折叠过程中, 在不同的阶段所具有的构象性质, 与一般的高分子链应该有所不同. 对于相同链长的二维蛋白质模型分子, 没有紧密接触对时链的尺寸最大, 随着紧密接触对数目的增加, 链的尺寸也随之减小, 这说明在蛋白质的折叠过程中, 分子大小是慢慢变小的, 最后变成非常紧密的蛋白质基态分子. 对于图中有些点偏差比较大, 原因可能是由于链比较短, 构象数目比较少, 其统计性质存在一定涨落.

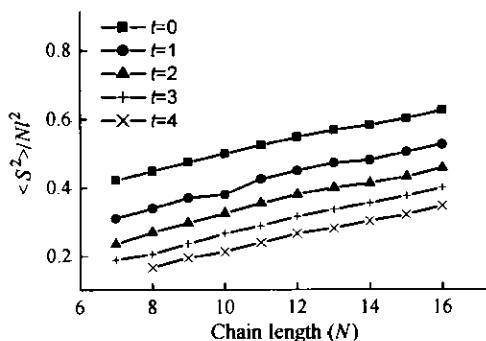


Fig. 1 Characteristic ratio $\langle S^2 \rangle_i / Nl^2$ as a function of chain length N with the number of contacts from $t = 0$ to $t = 4$

2.2 具有相同紧密接触对数目的分子大小分布函数 $P(S)$

蛋白质分子在折叠过程中, 其能量是随之减少, 同时构象的紧密接触对数目也随之增加. 从上面的分析可知, 在折叠过程中, 其尺寸也随之变小. 为了更清楚地了解其构象, 计算了均方回转半径的分布函数 $P(S)$, 结果见图 2. 这里仅给出了 $N = 13$ 的分布函数 $P(S)$, 但对于其他链长的二维蛋白质模型分子都具有相似的分布函数. 在图 2 中可知, 当 t 比较小的时候(折叠过程的初期), 其大小分布比较宽, 其最大的分布几率比较小. 随着折叠过程的进行, 回转半径的分布函数 $P(S)$ 随之变窄, 在 $S / \langle S^2 \rangle^{1/2} = 0.925$ 到 $S / \langle S^2 \rangle^{1/2} = 1.025$ 区间出现的几率变大. 当 $t = 7$ 时, 90% 的构象几乎都出现在 $S / \langle S^2 \rangle^{1/2} = 0.95$ 到 $S / \langle S^2 \rangle^{1/2} = 1.00$ 区间内. 由于这里不考虑氨基酸的序列, 因此还有一些构象在这个区间以外. 通过对分布函数 $P(S)$ 的研究, 我们可以初步了解在折叠过程的后期可能存在的构象.

2.3 具有不同氨基酸序列的蛋白质模型分子大小分布函数 $P(S)$

这里采用 HP 格点模型, 研究具有不同氨基酸序列的蛋白质模型分子的均方回转半径分布函数

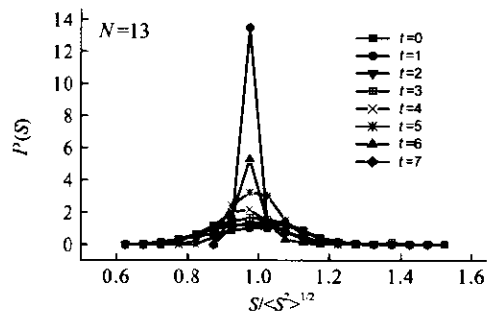


Fig. 2 Distribution function of $P(S)$ as a function of $S/\langle S^2 \rangle^{1/2}$ with the number of contacts from $t = 0$ to $t = 7$. Here chain length is $N = 13$.

$P(S)$. 结果见图 3. 图 3(a) 氨基酸序列为 HHPHHPPHPPHPPP, 图 3(b) 氨基酸序列为 HHPHPHHPPHPPHPPP. 虽然他们的亲水基团(P)和疏水基团(H)的数目相同,但是他们的氨基酸序列是不同的.最后他们的迴转半径分布函数 $P(S)$ 也有非常大的差别.从计算发现,序列 HHPHPHHPPHPPHPPP 的二维蛋白质分子是可以折叠到自然基态(Native state)的分子,是可折叠序列(Folding sequence);而序列 HHPHHPPHPPHPPHPPP 的二维蛋白质分子是不可以折叠到自然基态,或者它没有自然基态,是非折叠序列(Unfolding sequence).对于不能折叠到自然基态(Native state)的分子(见图 3(a),在迴转半径较小的区域,或者较大的区域,分子出现的可能性比较大,而在均方迴转半径附近,出现的几率相对要小一些.在 $E = -2$ 的曲线中存在双峰现象的原因可能是由于这里讨论的二维蛋白质模型分子比较短,构象数目相对比较少.但从图 3(b)可知,对于可以折叠到自然基态(Native state)的分子,在迴转半径较小的区域,或者较大的区域,分子出现的可能性比较小,而在均方迴转半径附近,出现的几率很大.随着能量的降低,这种趋势更加明显.

2.4 二维蛋白质模型分子在不同序列时的构象特点

采用精确计数法,计算了不同长度不同序列时的构象特点.这里计算了可设计序列数目 S_N (The number of designing sequences)和可设计构象数目 D_N (The number of designable conformations).它们分别表示可折叠的序列数目和构象数目.在图 4 中给出了比值 S_N/D_N 与链长 N 的关系.发现对于较小的 N 值,比值 S_N/D_N 几乎没有规律;但对

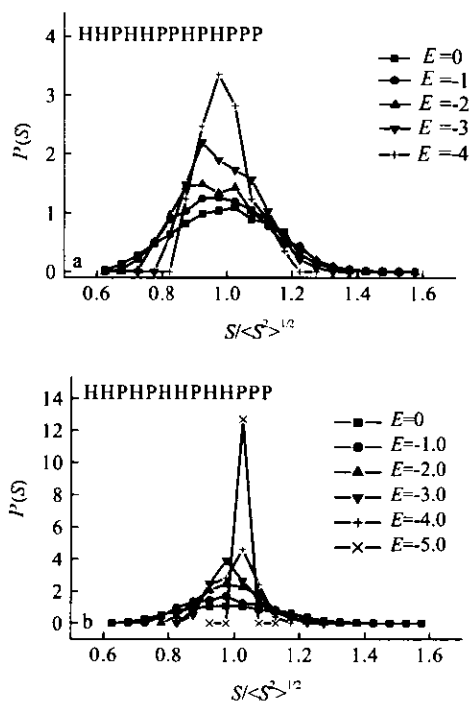


Fig. 3 Distribution function $P(S)$ as a function of $S/\langle S^2 \rangle^{1/2}$ with (a) unfolding sequence [HHPHHPPHPPHPPP] and (b) folding sequence [HHPHPHHPPHPPHPPP]

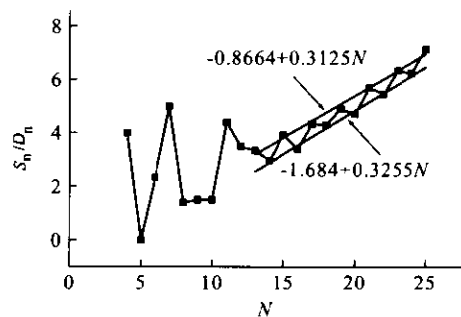


Fig. 4 The ratio S_N/D_N as a function of chain length N . Here S_N is the number of designing sequences and D_N is the number of designable conformations.

于较大的 N 值,比值 S_N/D_N 出现交替变化,即一个比较大,下一个比较小.同时可以表示为

$$-1.684 + 0.3255N \leq S_N/D_N \leq -0.8664 + 0.3125N \quad (N \geq 13) \quad (7)$$

当 $N \geq 13$ 时,在图中出现了交替变化的规律,这是格子模型模拟中经常出现的“奇偶效应”,这里非常明显.这些结果能帮助我们了解蛋白质的构象及其折叠过程.有关三维蛋白质分子的折叠构象正在研究之中.

REFERENCES

- 1 Kamtekar S, Schiffer J M, Xiong H, Babik J M, Hecht M H. *Science*, 1993, 262: 1680 ~ 1685
- 2 Lau K F, Dill K A. *Macromolecules*, 1989, 22: 3986 ~ 3997
- 3 Dill K A, Bromberg S, Yue K, Fiebig K M, Yee K M, Thomas P D, Chan H S. *Protein Sci*, 1995, 4: 561 ~ 602
- 4 Yue K, Fiebig K M, Thomas P D, Chan H S, Shakhnovich E I, Dill K A. *Proc Natl Acad Sci USA*, 1995, 92: 325 ~ 329
- 5 Gromiha M M. *Biophys Chem*, 2001, 91: 71 ~ 77
- 6 Gromiha M M, Selvaraj S. *Biophys Chem*, 1999, 77: 49 ~ 68
- 7 Jiang Zhouting, Zhang Linxi, Chen Jin, Xia Agen, Zhao Delu. *Polymer*, 2002, 43: 6037 ~ 6047
- 8 Chen S J, Dill K A. *Proc Natl Acad Sci USA*, 2000, 97: 646 ~ 651
- 9 Jiang Zhouting, Zhang Linxi, Chen Jin, Zhao Delu. *Polymer*, 2002, 43: 1461 ~ 1466
- 10 Zhang Linxi(章林溪), Jin Jinsheng(金进生), Wang Xianhong(王向红). *Acta Polymerica Sinica(高分子学报)*, 2002, (4): 479 ~ 483
- 11 Chen Jin(陈进), Li Jing(李靖), Zhang Linxi(章林溪), Jiang Zhouting(姜舟婷), Xia Agen(夏阿根), Ye Gaoxiang(叶高翔), Zhao Delu(赵得禄). *Acta Polymerica Sinica(高分子学报)*, 2003, (2): 171 ~ 174

CONFORMATION STUDIES OF TWO-DIMENSIONAL MODEL MOLECULES OF PROTEINS IN THE PROCESS OF FOLDING

CHEN Jin¹, DONG Zhijun¹, ZHANG Linxi¹, JIANG Zhouting¹, ZHAO Delu²

(¹ Department of Physics, Zhejiang University, Hangzhou 310028)

(² Polymer Physics Laboratory, Center of Molecular Science, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100080)

Abstract There are many changes for protein molecules in conformations in the process of folding. The mean-square radius of gyration and its distribution function $P(S)$ were calculated by using the exact enumeration calculation method. It was found that the dimension was large at the beginning of folding, and the range of its distribution was also wide. However, the dimension became small and compact at the end of folding, and the range of its distribution also became small. The mean-square radius of gyration $\langle S^2 \rangle_t$ with the number of contacts t is shown as

$$\langle S^2 \rangle_t / Nl^2 = a + bN$$

where a depends on the number of contacts (t) and b is independent of the number of contacts (t), and N is the chain length of two-dimensional protein molecules. There are different distributions for various residue sequences. The value of $P(S)$ nearby the mean-square radius of gyration is more large for the folding sequence. A ratio of S_N/D_N was also calculated. S_N is the number of designing sequences, and D_N is the number of designable conformations. It was found that the ratio may be expressed as

$$-1.684 + 0.3255N \leq S_N/D_N \leq -0.8664 + 0.3125N \quad (N \geq 13)$$

These results may provide some insights into the folding of proteins molecules.

Key words Two-dimensional model protein, HP model, Exact enumeration calculation method, Mean-square radius of gyration, $P(S)$, Folding