基于格子模型的蛋白质设计方法*

刘 赟[1] 王存新[1]** 王宝翰[2] 陈慰祖[1]

（[1]北京工业大学生命科学与生物工程学院，北京 100022；[2]中国科学院生物物理研究所，北京 100101）

摘要 提出一个简单有效的蛋白质设计方法，这一方法完全基于物理学原理. 与同类工作相比，该方法在很大程度上可节省对序列空间进行的搜索，是对同类工作的简化与发展. 对三个平面格子模型进行的检验表明该方法是成功的. 该方法可进一步用于真实蛋白质的三维非格子模型.

关键词 蛋白质设计，蛋白质逆折叠，格子模型，非格子模型

学科分类号 Q615

蛋白质折叠与蛋白质逆折叠问题是分子生物学中的重要问题. 研究折叠问题的目的是从蛋白质序列出发来预测结构，这一问题已获得很大进展[1~3]. 蛋白质逆折叠问题又称为蛋白质设计，相当于给定结构和温度，在序列空间中搜索一个（或几个）能折叠到此结构并保持结构稳定的序列. 逆折叠问题广泛应用于分子生物学的基础研究以及药物设计领域. 从理论上来说，可以通过穷举法为特定结构的蛋白质找到适合的序列[4]，可以先给出一个需要设计序列的目标结构，再根据蛋白质链长列出所有序列，然后在特定条件下（例如给定温度），通过检验每一个序列是否能快速折叠成稳定的目标结构来挑选最适合的序列.

Shakhnovich 和 Gutin[5]最早用蒙特卡罗方法在目标结构中寻找能量最低的序列（SG 方法），在这个工作中，需要对残基的比例进行人为的限制. Kurosky 和 Deutsch[6,7]指出优化的目标函数应为自由能差，并求出了自由能累积（cumulant）展开的一阶近似（KD 方法），用格子模型进行了测试. 此方法与 SG 方法比起来，不需要对残基比例作出限制. 在格子模型上用模拟退火对此方法进行的测试表明，KD 方法优于 SG 方法. Seno 等[8]使用双重蒙特卡罗方法对构象空间和序列空间同时搜索，在特定格子模型上的测试获得了成功，此方法在原则上较前述方法更精确，但在双重空间上的搜索需要大量计算机 CPU 时间，很难应用于较大体系，也很难应用于非格子模型的真实蛋白质结构.

在这个领域，我们曾经提出了一个基于相对熵的理论方法，用于处理蛋白质折叠与逆折叠问题[9~11]. 其基本思想是从相对熵概念出发，对相对熵进行优化，以寻找一个给定氨基酸序列的最优构象（蛋白质折叠），或者用最陡下降法进行真实蛋白质逆折叠问题（蛋白质设计）的研究. 与同类工作相比，得到了较好的结果.

由于格子模型比真实蛋白质的非格子模型更为简单，易于进行理论研究，也便于推广到真实蛋白质的非格子模型，上述的 SG 方法、KD 方法、Seno 等的双重蒙特卡罗方法[5~8]以及其他一些理论性的研究[4,12,13]都广泛使用了格子模型. 本工作采用了格子模型，改进和简化了基于相对熵的蛋白质设计方法，不需要使用蒙特卡罗方法或最陡下降法，直接确定残基的种类，为真实蛋白质的研究提供了理论基础. 本研究仅采用亲水-疏水模型，还未涉及真实的氨基酸序列，所以离真实的蛋白质逆折叠，还有相当距离，需要进一步研究.

1 理论与方法

假设 $H(r, s)$ 是蛋白质分子体系的哈密顿量，可写为

$$H_r(s) = \frac{1}{2}\sum_{i,j\neq i}^{N} U(s_i, s_j)A(r_i - r_j) \qquad (1)$$

其中 N 为残基总数，$S = (s_1, s_2, \cdots, s_n)$ 表示蛋白质的残基序列，我们采用简单的疏水-亲水模型（HP 模型）来检验我们的算法，把所有氨基酸残基分为两类，即：非极性疏水氨基酸和极性亲水氨基酸. 定义 $s_i = 1$ 时表示疏水残基，$s_i = -1$ 表示亲水残基. $A(r_i - r_j)$ 表示残基之间相互作用强度

* 国家自然科学基金（10174005，30170230）和北京市自然科学基金（5032002）资助项目.

** 通讯联系人.

Tel：010-67392724，E-mail：cxwang@bjut.edu.cn

收稿日期：2003-08-18，接受日期：2003-09-28

函数，如果第 i 个残基与第 j 个残基相邻且 i 不等于 $j-1$ 或 $j+1$，则 $A(r)=1$，否则 $A(r)=0$. r_i 表示第 i 个残基的坐标. $U(s_i, s_j)$ 为残基 i 与 j 之间的接触势，可定义为[9]：

$$U(s_i,s_j) = a + bs_i + cs_j + ds_is_j \qquad (2)$$

其中 a，b，c，d 为待定参数，由具体势函数模型决定. 有物理意义的 $U(s_i, s_j)$ 必须符合 Miyazawa 与 Jernigan[14] 或 Maiorov 和 Crippen[15] 提出的关于接触势函数满足的条件，也就是 $u(1,1)+u(-1,-1)<2u(1,-1)$.

相对熵 G 可表示为[9,11]

$$G(s) = \sum_r P_\alpha \ln(P_\alpha/P_0) \qquad (3)$$

其中下标 α 表示目标结构. P_0 表示在任一序列 $\{s\}$ 的条件下分子具有构象 $\{r\}$ 的几率，P_α 表示在任一序列 $\{s\}$ 的条件下，分子具有指定的固定构象 $\{r^\alpha\}$ 的几率. 相对熵 G 表示这两个几率函数相差的测度，我们希望 P_0 尽可能接近 P_α 从而使 P_0 和 P_α 之间的差异最小，因而必须优化相对熵 $G(s)$ 并使 $G(s)$ 极小化. 一般的能量优化方法是，对于给定目标构象 $\{r_i^\alpha\}$，从所有可能形成该构象的序列中寻找哈密顿量最小的序列作为最佳序列. 而基于相对熵的蛋白质设计方法是，对于给定的目标构象，从形成该构象的序列中，找到相对熵最小的序列，即对应于该构象的最佳序列. 利用最陡下降法优化相对熵 $G(s)$ 可以得到[11]

$$s_i^{k+1} - s_i^k = -\eta\beta\sum_{j\neq i}[A(r_i^\alpha - r_j^\alpha) - \langle A(r_i - r_j)\rangle_0](b + ds_j^k) \qquad (4)$$

此公式的详细推导参见文献[11]. 其中：η 是介于 0 和 1 之间的调节参数，用来控制迭代的收敛速度；$\beta=1/RT$，T 为绝对温度，R 为普适气体常数；$\langle\ldots\rangle$ 表示对于分布函数 P_0 的系综平均. 考虑到我们采用简单的疏水亲水模型（HP 模型），上式可进一步简化为以下的符号函数形式[11]

$$s_i^{k+1} = -\mathrm{sgn}(\eta\beta\sum_{j\neq i}[A(r_i^\alpha - r_j^\alpha) - \langle A(r_i - r_j)\rangle_0](b + ds_j^k)) \qquad (5)$$

其中 sgn 为符号函数，即

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{如果 } x \geqslant 0, \\ -1 & \text{其他情况} \end{cases} \qquad (6)$$

注意到在公式（2）中，只要 a、b、c、d 的选取能使 $U(s_i, s_j)$ 满足 $u(1,1)+u(-1,-1)<2u(1,-1)$ 这一基本条件，就能利用迭代公式（5）得出符合目标结构的最佳序列，而我们总是可以选取满足这一条件的适当参数，使 $d \ll b$，例如，可以选 $a=-0.995$，$b=c=-0.5$，$d=-0.005$，从而使 $U(1,1)=-2$，$U(1,-1)=-0.99$，$U(-1,-1)=0$，以满足上述条件. 这样，公式（5）中的 ds_j^k 项就可以忽略. 另外，公式（5）的迭代结果是收敛的[11]，因此当 $k\to\infty$ 时，$s_i^k\to$稳定值 s_i，这样就有

$$s_i = -\mathrm{sgn}(b\eta\beta\sum_{j\neq i}[A(r_i^\alpha - r_j^\alpha) - \langle A(r_i - r_j)\rangle_0]) \qquad (7)$$

从原则上来说，只要能精确算出 $\langle A(r_i-r_j)\rangle_0$ 的值就能确定最适合目标结构的序列. 这里存在的困难是计算 $\langle A(r_i-r_j)\rangle_0$ 的值. $\langle A(r_i-r_j)\rangle_0$ 是指对于几率分布 P_0 的平均接触强度，需要对所有构象求和，所以求 $\langle A(r_i-r_j)\rangle_0$ 的值或仅仅给出较精确的估计值都是非常困难的，这里仅给出一个粗略的估计. 在以下的推导中，我们将逐步忽略 $\langle A(r_i-r_j)\rangle_0$ 对下标 i、j 的依赖性，以一个常数 $\bar{A}$ 作为 $\langle A(r_i-r_j)\rangle_0$ 的粗略估计值.

由公式（4）可知，当 $k\to\infty$ 时有 $\lim(s_i^{k+1}-s_i^k)=s_i-s_i=0$，从而有

$$\sum_{j\neq i}A(r_i^\alpha - r_j^\alpha)(b + ds_j) = \sum_{j\neq i}\langle A(r_i - r_j)\rangle_0(b + ds_j) = A_i\sum_{j\neq i}(b + ds_j) \qquad (8)$$

这里首先忽略了 $\langle A(r_i-r_j)\rangle_0$ 关于下标 j 的依赖性，粗略地认为 $\langle A(r_i-r_j)\rangle_0$ 只与下标 i 有关，以 A_i 近似表示 $\langle A(r_i-r_j)\rangle_0$. 注意到此时的 $\langle A(r_i-r_j)\rangle_0$ 表示序列 s 处于稳定值时关于 P_0 的系综平均，且 $\langle A(r_i-r_j)\rangle_0 = \dfrac{\sum_r e^{\beta H(\bar{s},r)}A(r_i-r_j)}{\sum_r e^{\beta H(\bar{s},r)}}$.

这里参照了求多粒子系统质心的方法. 因为对于多粒子系统有 $\sum_i m_i r_i = r_c\sum_i m_i$，其中 m_i，r_i 分别为第 i 个粒子的质量和坐标，r_c 为系统的质心坐标. （8）式中 $\langle A(r_i-r_j)\rangle_0$，$b+ds_j$ 和 A_i 分别对应于 m_i，r_i 和 r_c. 这样，（8）式就可以看作当序列处于稳定值时，作用于残基 s_i 上力场的条件. 由（8）式可得

$$A_i = \frac{\sum_{j\neq i}A(r_i^\alpha - r_j^\alpha)(b + ds_j)}{\sum_{j\neq i}(b + ds_j)} \qquad (9)$$

由于 $d \ll b$，上式中的 ds_j 项可以忽略，注意到 A_i 是和 i 相关的. 对于不同残基 i，上式中 A_i 的取值不

一样. 如果忽略 A_i 对 i 的依赖性，令 $\bar{A}=\frac{1}{N}\sum_i A_i$，可得

$$\bar{A}\approx\sum_{i,j\neq i}A(r_i^\alpha-r_j^\alpha)/N(N-1) \qquad (10)$$

结果与文献［11］相同[11]，这里给出了一个基于物理原理的解释，给出了 $<A(r_i-r_j)>_0$ 的一个粗略的估计值 $\bar{A}$. 用 $\bar{A}$ 代替 $<A(r_i-r_j)>_0$ 并注意到 $-b\eta\beta$ 是一个正数，公式（7）可以写成

$$s_i=\mathrm{sgn}(\sum_{j\neq i}[A(r_i^\alpha-r_j^\alpha)-\bar{A}]) \qquad (11)$$

作蛋白质设计的时候，目标结构中的每一个残基的坐标已给定，即 $A(r_i-r_j)$ 是已知的，从而可以利用（10）式估计 $\bar{A}$ 的值，最后用（11）式可以方便地确定每个残基的种类.

2 结果与讨论

一般来说，真实蛋白质在转变温度下能快速折叠到自由能最低的构象，这样的构象是稳定的. 人们用格子模型研究了蛋白质的折叠问题，从给定的序列出发，寻找自由能最低的构象.

为简单起见，这里采用二维格子模型来检验我们的方法. 用格子模型对较短链长的模型进行研究表明，任意一个序列都拥有多个不同的构象，通常只存在一个特殊构象，其中的疏水残基最大程度地聚集在一起，这相当于真实球状蛋白质内部含有一个疏水核心. 在文献［12］给出了一个残基数为 14 的短链例子，首先确定一个序列，然后在构象空间中搜寻自由能最低的构象，其结果见图 1. 给定序列在这个结构中具有最低自由能，可认为给定序列的蛋白质链能够折叠成此结构并保持稳定.

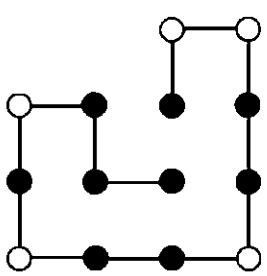

Fig. 1 The unique conformation of lowest free energy[12]

The dark and light circles represent polar and nopolar residues, respectively.

我们反过来为这个结构设计序列，以检验我们的蛋白质设计方法是否有效. 首先由（10）式计算出 $\bar{A}\approx0.076$，再按照（11）式确定这条链的每一个残基的种类，最终得到的序列与图 1 中的序列完全符合. 这就表明我们设计的序列使相应构象具有最低的自由能.

一般而言，在转变温度以下，自然界存在的蛋白质应该能从任意展开构象快速折叠成紧凑的稳定构象，通过对此类结构设计序列可以进一步检验我们的理论是否成功. 在文献［13］中给出了能够快速折叠的序列及相对应的结构（图 2），所示的序列能使这个有 20 个残基的蛋白质链快速折叠成图 2 中的结构，并以此结构作为其唯一的稳定结构，任何其他结构都会使构象的能量增加.

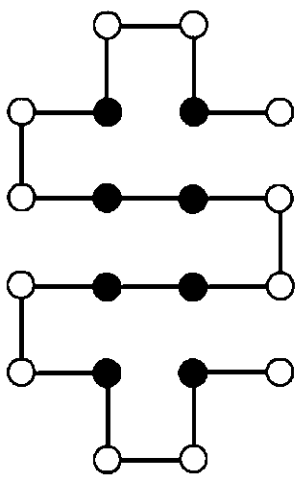

Fig. 2 A sequence is shown in its unique native structure[13]

The dark and light circles represent polar and nopolar residues, respectively.

与上一个例子类似，我们为图 2 所示的结构设计序列，在此例中 $\bar{A}\approx0.058$，由此得到的结果与图 2 中的序列完全符合. 这说明我们设计的序列能使此蛋白质链快速折叠成稳定构象，进一步证明我们的理论是有效的.

这两个模型的链长较短，对于更长的链效果如何还有待检验，为此，我们选取有 50 个残基的链[16]，这就需要在 $2^{50}\approx10^{15}$ 个序列组成的序列空间中相对于每一个结构进行搜索，运算量是巨大的. 文献［16］对含 50 个残基的目标结构设计序列，从 10^5 个随机序列中挑选出结果，并证明最终挑选出的序列是以图 3 中的目标结构为其唯一的基态.

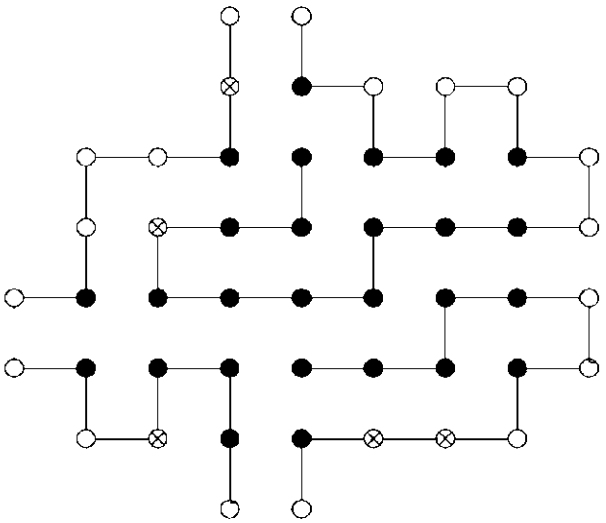

Fig. 3 Target structure with $N=50$ monomers[14]

The dark and light circles represent polar and nopolar residues, respectively. Our results differ from that of Irback *et al.* in five sites (crosses). In our predicted sequence, the positions $i=10$, 11, 18, 28 ($i=1$ corresponds to the lower of the two end points) are polar and $i=47$ is the nopolar one.

用我们的方法按照（10）式可算出 $\bar{A}\approx 0.028$，再由（11）式得到最终结果. 与文献[16]的结果比较，我们设计的序列中有5个残基不同（在图3中用⊗表示）. 这表明我们的结果与文献[16]的结果有90%是相同的，即同源性达到90%. 之所以产生差异，是因为适合目标结构的序列可能有多个[4]，使用不同的方法所得结果有可能存在差异.

3 结 论

本文从基本物理学原理出发，在格子模型上研究了蛋白质设计问题. 估计了残基间接触函数对于Boltzmann分布的平均值. 与同类工作相比，本文给出了一个判断残基疏水性（亲水性）的简单方法，节省了在序列空间的搜寻时间，同时指出了 $<A(r_i-r_j)>_0$ 在设计过程中的重要性，是对同类工作的简化与发展. 本方法在格子模型的情况下是成功的，可以方便地推广到非格子模型的情况，这一工作也已经完成，对4种不同结构类型的真实蛋白质进行了检测，与同类工作相比，得到了较好的结果（待发表）.

需要指出，本文只局限于确定残基的亲水性与疏水性，残基间的相互作用势过于简化，自然界的蛋白质通常有20种残基，如何把我们的理论发展到20种残基的情形还有待进一步研究. 另外，本文是在格子模型的基础上研究蛋白质设计问题的，发展到真实蛋白质的非格子模型，如何选择残基间的相互作用接触强度函数还需要进一步讨论.

致谢 感谢李春华博士、薛颖同学以及王屹华同学对本工作提供的帮助.

参 考 文 献

1 Bryngelson J, Onuchic J N, Socci N D, *et al.* Funnels, pathways, and the energy landscape of protein folding. Proteins, 1995, **21** (3): 167 ~ 195

2 Fersht A. Nucleation mechanism of protein folding. Curr Opin Struct Biol, 1997, **7** (1): 10 ~ 14

3 Shakhnovich E I. Theoretical studies of protein-folding. Curr Opin Struct Biol, 1997, **7** (1): 29 ~ 40

4 Li H, Helling R, Tang C, *et al.* Emergence of preferred structures in a simple model of protein folding. Science, 1996, **273** (2): 666 ~ 669

5 Shakhnovich E I, Gutin A M. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA, 1993, **90** (15): 7195 ~ 7199

6 Kurosky T, Deutsch J M. Design of copolymeric materials. J Phys A: Math Gen, 1995, **27** (14): L387 ~ L393

7 Deutsch J M, Kurosky T. New algorithm for protein design. Phys Rev Lett, 1996, **76** (2): 323 ~ 326

8 Seno F, Vendruscolo M, Maritan A, *et al.* Optimal protein design procedure. Phys Rev Lett, 1996, **77** (9): 1901 ~ 1904

9 Wang B H, Yun Z X, Wang Z X, *et al.* A unified design approach for the inverse folding and direct folding of protein. J Bioscience, 1999, **24** (suppl 1): 61

10 卢本卓，王存新，王宝翰. 用于真实蛋白质结构预测的一种新的优化方法. 化学物理学报，2003，**16**（2）：117 ~ 121
Lu B Z, Wang C X, Wang B H. Chinese Journal of Chemical Physics, 2003, **16** (2): 117 ~ 121

11 刘 赟，王宝翰，王存新，等. 基于相对熵的蛋白质设计新方法. 中国科学G辑，2003，**33**（4）：348 ~ 356
Liu Y, Wang B H, Wang C X, *et al.* Science in China (Series G), 2003, **33** (4): 348 ~ 356

12 Dill K A. Dominant forces in protein folding. Biochemistry, 1990, **29** (31): 7133 ~ 7154

13 Dill K A. Principles of protein folding-A perspective from simple exact models. Protein Science, 1995, **4** (4): 561 ~ 602

14 Miyazawa S, Jernigan R L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 1985, **18** (3): 534 ~ 552

15 Maiorov V N, Crippen G M. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol, 1992, **227** (13): 876 ~ 888

16 Irback A, Peterson C, Potthast F, *et al.* Design of sequences with good folding properties in coarse-grained protein models. Structure With Folding & Design, 1999, **7** (3): 347 ~ 360

A Protein Design Procedure Based on The Lattice Model*

LIU Yun[1)], WANG Cun-Xin[1)**], WANG Bao-Han[2)], CHEN Wei-Zu[1)]
([1)] *College of Life Science and Bioengineering, Beijing University of Technology, Beijing* 100022, China;
[2)] *Institute of Biophysics, The Academy Sciences of China, Beijing* 100101, China)

Abstract A new simple and effective approach completely based on the physical theory is proposed for protein design. Compared with the similar works, the algorithm saves a vast deal of trouble in exploring sequence space. The method is an improvement over previous works. The design procedure was tested on three lattice models in two dimensions and the successful results have been obtained. The method can be readily implemented for three-dimensional off-lattice models of real proteins.

Key words protein design, inverse protein folding, lattice model, off-lattice model

* This work was supported by grants from The National Natural Sciences Foundation of China (10174005, 30170230) and Beijing Natural Science Foundation (5032002).

** Corresponding author. Tel: 86-10-67392724, E-mail: cxwang@bjut.edu.cn

Received: August 18, 2003 Accepted: September 28, 2003

第三届亚洲视觉科学会议（ACV2004）第一轮通知

由中国科学院视觉信息加工重点实验室、上海神经科学研究所和第三军医大学西南眼科医院联合主办的第三届亚洲视觉科学会议（ACV）原定于2003年11月举行，因受SARS影响，改为2004年11月15～19日在重庆举行.

大会组织委员会：主 席：李朝义（中国）

副主席：Keiji Uchikawa（日本），Chan-Sup Chung（韩国），赫荣乔（中国）

大会学术委员会：主 席：王书荣（中国）

副主席：Makoto Ichikawa（日本），Choongkil Lee（韩国）

大会执行委员会：主 席：阴正勤（中国），副主席：李兵（中国），谢汉平（中国）

会议主题：

1）Visual Neuroscience; 2）Visual Perception; 3）Depth and Spatial Vision; 4）Color Vision 5）Visual Attention; 6）Eye Movements and Visuo-Motor Coordination; 7）Mathematical Models on Vision; 8）Retina; 9）Object Recognition; 10）Clinical Vision Studies; 11）Neural Imaging of Visual System; 12）Vision and Other Modalities.

重要时间：1）会议回执：2004年4月30日

2）论文摘要截止日期：2004年8月30日

有关会议地点、注册费、论文摘要格式等信息请详见第二轮通知. 第二轮通知将于2004年6月登出，届时将通过E-mail或邮寄发出，并请见会议网址：www.ibp.ac.cn/acv2004 敬请在网上下载会议回执.

第三届亚洲视觉科学会议（ACV2004）回执

（2004年11月15日～19日，重庆）

姓 名：_____ 性别：_____ 职务：_____

工作单位：_____

通讯地址：_____

邮政编码：_____

电 话：_____ 传真：_____

E-mail _____

□ 我拟参加第三届亚洲视觉科学会议，请寄第二轮通知

□ 我拟提交会议论文摘要

会议回执寄至：北京100101 朝阳区大屯路15号 中国科学院生物物理研究所 魏舜仪

电话：010-64889894，传真：010-64853625，E-mail：acv2004@sun5.ibp.ac.cn