第 16 卷 第 2 期 化 学 物 理 学 报 Vol. 16, No. 2
2003 年 4 月 CHINESE JOURNAL OF CHEMICAL PHYSICS Apr. 2003

1003-7713/2003/02-117-5

用于真实蛋白质结构预测的一种新的优化方法*

卢本卓[a]， 王存新[b**]， 王宝翰[c]

（a. 中国科学技术大学天文与应用物理系，合肥 230026；
b. 北京工业大学生命科学与生物工程学院，北京 100022；
c. 中国科学院生物物理研究所，北京 100101）

摘 要： 用‘相对熵’作为优化函数，提出了一个有效快速的折叠预测优化算法. 使用了非格点模型，预测只关心蛋白质主链的走向. 其中只用到了蛋白质主链上的两两连续的 C_α 原子间的距离信息以及 20 种氨基酸的接触势的一个扩展形式. 对几个真实蛋白质做了算法测试，预测的初始结构都为比较大的去折叠态，预测构象相对于它们天然结构的均方根偏差（RMSD）为 5 ~ 7 Å. 从原理上讲，该方法是对能量优化的改进.

关键词： 蛋白质折叠预测；相对熵；优化方法；非格点模型

中图分类号： Q615 **文献标识码：** A

A New Minimization Method for Real Protein Folding Prediction*

Lu Benzhuo[a], Wang Cunxin[b**], Wang Baohan[c]

(*a. Department of Astronomy and Applied Physics, University of Science and Technology of China, Hefei* 230026;
b. College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022;
c. Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101)

Abstract A new effective and fast minimization approach is proposed for the prediction of protein folding, in which the " relative entropy " is used as minimization function. Unlike the energy minimization method, the essence of this approach is to search the conformation with high occupation probability, which corresponds to the state with low free energy instead of low energy. The off-lattice model is used, and the prediction just focuses on the frame of the main chain of protein. In this approach, only the distances between the consecutive C_α atoms along the peptide chain and a generalized form of the contact potential for 20 types of amino acids are used. Tests of the prediction algorithm are performed on real proteins with the initial structure fully denatured. The root mean square deviations (RMSD) of the structures of four folded target proteins versus the native structures are from 5 to 7 Å. The advantage of this approach is its simple potential function and fast performance. Moreover, it can be considered as an improvement on the energy minimization method in principle.

Key words Protein folding prediction, Relative entropy, Minimization method, Off-lattice model

1 引 言

近来，蛋白质三级结构预测（或折叠预测）已成为分子生物学中的热点问题. 如基于已知结构信息的 Threading 方法，已得到一些好的预测结果（对于小的蛋白质，均方根偏差的值在 3 ~ 7.5 Å）[1,2]. 然

* 国家自然科学基金资助项目（10174005、30170230 和 29992590-2）. ** 通讯联系人，E-mail：cxwang@bjpu.edu.cn
收稿日期：2002-05-30；修回日期：2002-07-09.

而,已经知道,小的蛋白质的氨基酸序列自身包含的信息已经足以决定它的折叠结构,即自由能最小的构象[3]. 这表明蛋白质的天然结构由氨基酸序列残基间的物理相互作用决定. 不少工作用 Monte Carlo、分子动力学(MD) 或其它的被称为 *ab initio* 的方法来优化体系的能量,预测蛋白质的结构[4-8],但一般能量优化方法并不考虑熵的效应,所以预测的结构并不对应自由能最小的态. 我们提出了一个简单快速的折叠预测算法,采用"相对熵"作为优化函数,代替了传统的体系的能量. 在实质上更接近于从自由能的角度考虑体系的优化. 采用了非格点模型,其中只用到了肽主链上的两两连续的 C_α 原子间的距离信息以及 20 种氨基酸的接触势的一个扩展形式. 对几个真实蛋白质作了算法测试,预测的初始构象都是充分去折叠的态,四个蛋白质例子的预测构象相对于它们天然结构的方均根偏差(RMSD)在 5 ~ 7 Å. 这一方法简单,且收敛速度快.

2 基于相对熵的蛋白质体系的优化算法

假设 $H(s,\mathbf{r})$ 为具有残基序列 $S=(s_1,s_2\cdots,s_n)$ 和构象 $\mathbf{r}=(\vec{r}_1,\vec{r}_2,\cdots,\vec{r}_n)$ 的蛋白质分子的哈密顿量,其中 $\vec{r}_i$ 为第 i 个残基的 C_α 原子的位置坐标,s_i 为第 i 个残基的类型."相对熵"G 的定义为:

$$G(\{\vec{r}_i\}) = \sum_{\{s_i\}} P_\alpha \ln \frac{P_\alpha}{P_0} \tag{1}$$

其中 P_α 是对于一个给定的构象 $\mathbf{r}$,分子占有序列 $S=\{s_i\}$ 的概率:

$$P_0 = \frac{1}{Z_0} e^{-\beta H(s,\mathbf{r})}, \quad Z_0 = \sum_{\{s_i\}} e^{-\beta H(s,\mathbf{r})} \tag{2}$$

P_0 是对于一个给定的构象 $\mathbf{r}$,分子具有一个特定序列(目标序列)$S_\alpha=\{s_i^\alpha\}$ 的几率是

$$P_\alpha = \frac{1}{Z_a} e^{-\beta H(s,\mathbf{r})} \prod_i \delta_{s_i s_i^\alpha}$$
$$Z_\alpha = \sum_{\{s_i\}} e^{-\beta H(s,\mathbf{r})} \prod_i \delta_{s_i s_i^\alpha} \tag{3}$$

可证明"相对熵" $G \geqslant 0$,且当 $P_0=P_\alpha$ 时 G 最小.

对于给定的序列 $\{s_i^\alpha\}$,通过最小化 G,搜索构象空间找到最优的构象 $\{\vec{r}_i\}$,达到折叠预测的结果. 用最陡下降法进行优化:

$$\frac{d\vec{r}_i}{dt} = -\eta \frac{\partial}{\partial \vec{r}_i} G \tag{4}$$

其中 η 是一个在 0 和 1 之间的可调参数,用来控制迭代的收敛速度;i 为第 i 个 C_α 原子.

蛋白质体系的哈密顿量用简单的接触势形式

$$H = \frac{1}{2} \sum_{i,j\neq i} U(s_i^\alpha, s_j^\alpha) A(\vec{r}_i - \vec{r}_j) \tag{5}$$

其中,$A(\vec{r}_i-\vec{r}_j)$ 为依赖于残基间距离 $\vec{r}_{ij}$ 的无量纲接触强度函数;$U(s_i,s_j)$ 为残基 s_i 和 s_j 之间的接触势. 对真实蛋白,可取 *MJ* 接触势矩阵元[9] 来代替. 将(5)式代入到(4)式中,最终可得到优化算法的数值迭代公式:

$$\vec{r}_i^{\,k+1} - \vec{r}_i^{\,k} = -\eta\beta \sum_{j\neq i} [U(s_i^\alpha, s_j^\alpha) - \langle U(s_i,s_j)\rangle_0] \frac{\partial}{\partial \vec{r}_i^{\,k}} A(\vec{r}_i^{\,k} - \vec{r}_j^{\,k}) \tag{6}$$

其中,k 为第 k 次迭代;$\beta=1/RT$,T 为绝对温度,R 为气体常数;$\langle U(s_i,s_j)\rangle_0$ 为相对几率分布 P_0 的平均接触势,它不依赖于残基序列. 为了在真实空间中搜索,我们取一连续形式的函数 $A(\vec{r}_i-\vec{r}_j)$:

$$A(r_{ij}) = \frac{1}{\sqrt{2n\pi}} e^{-(r_{ij}^2-d^2)/2n} + \varepsilon\left(-\frac{\sigma^2}{r_{ij}^2} + \frac{\sigma^4}{r_{ij}^4}\right) \tag{7}$$

其中,n、ε 和 σ 为可调参数;d 是一个残基接触距离附近的值,残基接触距离一般取 5.0 ~ 7.5 Å,这里取 $d=5.5$ Å. 在实际模拟中,两个连续的残基之间的距离用 SHAKE 算法[10] 来约束,因而,任何两个连续的残基间的相互作用不计算. 式(7)中的第一项指数项可以看作是 δ 函数的连续形式. 这一项和因子 $U(s_i,s_j)$ 一起考虑了蛋白质折叠的主要驱动力:疏水和亲水相互作用,其中 d 值(5.5 Å)正好对应了接触距离. 为防止一些残基(如疏水残基)紧密靠近,类似于式(7)中第二项的附加的作用力是有必要的. 本工作采用了一个类似于 van der Waals 的势,它比通常的 van der Waals 函数具有更光滑的距离依赖性,它在残基靠近的地方有一个势垒.

上面的优化算法从物理上可作如下解释:蛋白质折叠研究中,希望折叠的基态与更高能量的状态之间有较大的能隙,这样分子就能稳定的处于基态上,式(6)中的第二项$\langle U(s_i,s_j)\rangle_0$ 可以看作是与平均能量有关的项,因为天然态的能量比平均能量低,所以(6)式对能量差的优化结果倾向于导致大的能隙. 另外,由上面的定义,本方法实质上是寻找具有高占据率的状态,因而预测的结构应处在接近最低自由能的状态. 若适当修改本方法运用于格点模型的情形,可以发现本优化函数很接近于文献[11]中自由能优化函数,该工作处理的是蛋白质逆折叠的问题. 上面的方法与神经网络理论有许多共同之处.

一个棘手的问题是估计$\langle U(s_i,s_j)\rangle_0$的值，我们用简单的方法来估计$\langle U(s_i,s_j)\rangle_0$. 具有序列$s^\alpha=\{s_i^\alpha\}$的蛋白质能折叠到天然态构象的必要条件是只有当它当前状态的能量$E^\alpha=\langle H\rangle_\alpha$小于或等于某个平均能量[12,13]，比如刚好变性时的能量. 这个能量不是序列专一的. 把$\langle H\rangle_0$(即$H(s,\mathbf{r})$对序列的系综平均)考虑为变性态的能量，它可以写为：

$$\langle H\rangle_0=\frac{1}{2}\sum_{i,j\neq i}\langle U(s_i,s_j)\rangle_0 A(\vec{r}_i,\vec{r}_j) \tag{8}$$

而对于给定序列$s^\alpha=\{s_i^\alpha\}$的天然态的能量是

$$\begin{aligned}H&=\frac{1}{2}\sum_{i,j\neq i}U(s_i^\alpha,s_j^\alpha)A(\vec{r}_i-\vec{r}_j)\\&=\sum_i^{N-1}U(s_i^\alpha,s_{i+1}^\alpha)A(\vec{r}_i-\vec{r}_{i+1})+\\&\quad\frac{1}{2}\sum_{i,j\neq i\pm1}U(s_i^\alpha,s_j^\alpha)A(\vec{r}_i-\vec{r}_j)\end{aligned} \tag{9}$$

上式的第二项对应于长程相互作用，在去折叠(变性)态时假设这项等于零. 根据上面的分析就有：

$$\begin{aligned}\langle H\rangle_0&=\frac{1}{2}\sum_i\sum_{j\neq i}\langle U(s_i,s_j)\rangle_0 A(\vec{r}_i,\vec{r}_j)\\&=\sum_i^{N-1}U(s_i^\alpha,s_{i+1}^\alpha)A(\vec{r}_i-\vec{r}_{i+1})\\&=\bar{A}(N-1)\bar{U}\end{aligned} \tag{10}$$

其中，$\bar{A}=\dfrac{\sum_i^{N-1}U(s_i^\alpha,s_{i+1}^\alpha)A(\vec{r}_i-\vec{r}_{i+1})}{\sum_i^{N-1}U(s_i^\alpha,s_{i+1}^\alpha)}$

而$\bar{U}$是可以算出来的：

$$\bar{U}=\frac{1}{N-1}\sum_i^{N-1}U(s_i^\alpha,s_{i+1}^\alpha)$$

更进一步，可以调节$A(\vec{r}_i-\vec{r}_j)$的值达到最接近于某个平均值$\langle A(\vec{r}_i-\vec{r}_j)\rangle=\tilde{A}$，使得满足

$$\begin{aligned}\langle H\rangle_0&=\frac{1}{2}\sum_{i,j\neq i}\langle U(s_i,s_j)\rangle_0 A(\vec{r}_i,\vec{r}_j)\\&=\frac{1}{2}\sum_{i,j\neq i}\langle U(s_i,s_j)\rangle_0\tilde{A}\\&=\langle U(s_i,s_j)\rangle_0\tilde{A}\frac{N(N-1)}{2}\end{aligned} \tag{11}$$

因此，

$$\langle U(s_i,s_j)\rangle_0=\frac{\bar{A}}{\tilde{A}}\frac{2}{N}\bar{U}=k_m\frac{2}{N}\bar{U} \tag{12}$$

其中$k_m=\dfrac{\bar{A}}{\tilde{A}}$作为一个可调参数出现，可能依赖于蛋白质的大小. 式(6)与能量优化方法不同之处在于增加了一项附加项$\langle U(s_i,s_j)\rangle_0$. 可以证明这一项的出现可与体系的自由能的一阶近似联系起来，也可以与文献[11]讨论的自由能优化做类似对比.

3　模拟计算及结果讨论

基于上述算法，我们编制了一个蛋白质折叠预测的程序包. 在所有的测试运行中，取$\eta=0.2$，$T=1$. 先在蛋白质 BPTI 上测试并选择可调参数，它们的取值为：$n=360$ Å^2，$\varepsilon=0.17$，$\sigma=1.35$ Å，$k_m=7.2$. 当每个节点(一个 bead 代表一个氨基酸)在连续两次迭代中的位置差都小于 0.001 Å 时，就认为迭代已收敛了. SHAKE 约束的精度也置为 0.001 Å. 应当指出的是，当初始结构的键长偏离要约束的长度太大时，就先采用一个谐振势进行键长约束，然后再使用 SHAKE 约束.

同通常的非格点模型一样，蛋白质简化表示为由一组节点组成，每一个节点代表一个氨基酸，它的坐标为相应残基的 C_α 原子的位置. 选择了四个小蛋白做测试的靶蛋白. 它们在蛋白质数据银行(PDB)中的代码分别是 1bpi、1ejg、1fcl 和 1ubq. 每个蛋白首先被充分去折叠成线团(coil)后作为折叠过程的初始结构，这些初始构象与它们的天然态的 RMSD 分别是：1bpi 为 16.4 Å，1fcl 为 10.0 Å，1ejg

表 1　折叠预测的结果(NC 是天然接触数) *

Table 1　Results of folding predictions (NC is the number of native contacts) *

PDB code	Residues number	NC of native conformation	NC of initial conformation	NC of final conformation	RMSD/Å
1bpi	58	180	0	92	6.8
1fcl	56	179	0	96	5.6
1ejg	46	144	0	76	5.2
1ubq	76	229	18	135	5.5

* The distance criterion for contact of two residues is 7.5 Å. The values of RMSD are obtained from the final folded structure versus the native structure of PDB.

为10.3 Å，1ubq 为15.6 Å. 这些蛋白的性质和预测结果列在表1中. 表中 RMSD 数据在最近报道的 *ab initio* 蛋白质折叠预测结果范围之内[14,15]. 位置 RMSD 的值在6 Å 左右，可认为是小蛋白的预测目标精度范围[15].

优化程序中，除了两两连续的 C_α 原子之间的距离(一般在3.8 Å 附近)外，不使用其它的目标蛋白的已知结构信息，比如天然二级结构、二硫键、回转半径等. 图1显示了1bpi 和1ejg 的天然结构及最后折叠的结构比较. 天然结构取自 X 射线衍射分析. 折叠前的初始结构都是没有任何二级结构的线团，然而，如图1(a)、(b)显示的肽1bpi，折叠构象明显地恢复了保持在天然结构中两端的螺旋及部分折叠片. 如图1(c)、(d)的1ejg，折叠结构中左边的两个片段是部分的螺旋，这也是天然结构中含有的. 因而，折叠计算能够得到和保持一些天然结构中含有的二级结构.

在四个目标蛋白中，1bpi 和1ejg 都有三个二硫键，1ubq 和1fcl 没有二硫键. 优化表明，当计算中二硫键也做约束时，预测的精度会有提高.

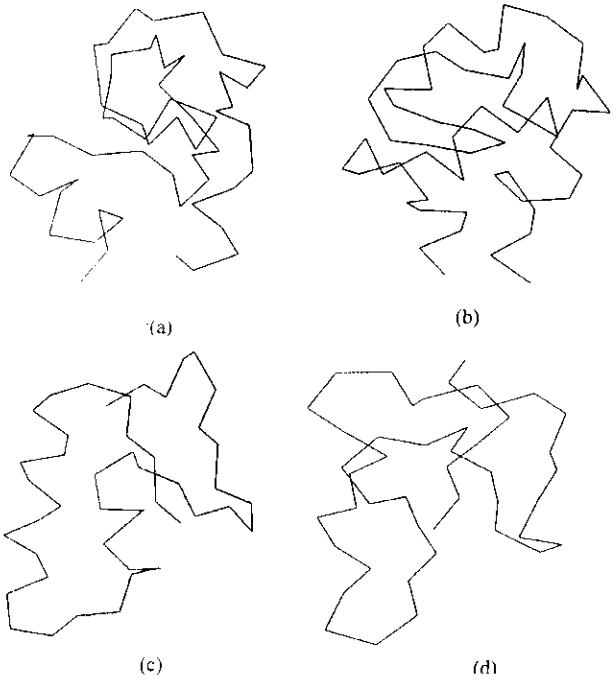

图1 (a)1bpi 的天然结构；(b)1bpi 的最后折叠结构；(c) 1ejg 的天然结构；(d) 1ejg 的最后折叠结构.

Fig. 1 (a) The native structure of 1bpi；(b) The final folded structure of 1bpi；(c) The native structure of 1ejg；(d) The final folded structure of 1ejg.

这一方法的特点是附加项：式(6)中的平均接触势$\langle U(s_i,s_j)\rangle_0$，这是一个小的值，对于本例中的蛋白，通常比 $U(s_i^\alpha,s_j^\alpha)$ 小若干倍. 但这一项对预测结构和算法的收敛速度确实有影响. 有和没有这一项的两种算法也在1ejg 上作了比较. 在我们的计算中，有这项的算法在2257次迭代后就收敛了，并产生5.2 Å 的 RMSD 和76个天然接触(天然态中总共有144个接触). 然而，略去了式(6)中的平均势的算法却需3673步才收敛，结果有5.9 Å 的 RMSD 和71个天然接触. 多数情况下含有平均势的算法导致天然接触数的增加因而也有更好的预测结果. 以前的能量优化的方法旨在找到一个具有最小能量的结构，而正是这一项的不同使得本方法的处理更接近于认为蛋白质的天然态并不正好处在能量最低的状态，而是对应于自由能最低的状态的观点.

4 结 论

综上所述，提出了一个新的有效的蛋白质折叠算法，其实质是按照 Boltzmann 分布搜索构象空间. 这一方法完全基于物理学原理，从根本上有别于其它的利用同源性建模、Threading 和基于对已知晶体结构作统计比较等的结构预测方法. 本方法仅采用了简单的扩展接触势而没有包含如文献[8]中的键角、二面角或其它形式的势. 另外，类似于 van der Waals 的约束势在本方法中也是有必要的，当然也可以探索其它形式的约束势函数. 这一项在短程范围内才有效，实际上在两个靠近的残基(约2 Å)起排斥力的作用. 势函数的选择是这一方法应当改进的. 由于 *MJ* 矩阵的统计性质，残基间相互作用的某些特性(如作用的方向依赖性、侧链堆积等)被平均掉了. 这些特殊的性质对建立蛋白质折叠的更真实的势是必要的. 另外，应该指出的是，和其它优化方法一样，本方法也必然会遇到局部最小的问题，因而预测的结果与初始构象有关. 我们只选择了一组较大的变性态(没有二级结构)作为初始构象，对任意初始态的预测仍然是个挑战. 这一方法从原理上讲可以作为蛋白质折叠和逆折叠的统一的处理框架[16]. 若对 P_0 和 P_α 的定义做相应修改，在 $\{s_i\}$ 空间的求和变为在 **r** 空间求和，则 P_0 和 P_α 以 $\{s_i\}$ 为变量，这样式(1)就可以导致适于蛋白质逆折叠问题的算法[16].

参 考 文 献

[1] Moult J, Hubbard T, Fidelis K, Pedersen J T. *Proteins Suppl.*, 1999, **3**: 2

[2] Venclovas C, Zemla A, Fedelis K, Moult J. *Proteins Suppl.*, 1999, **3**: 231

[3] Anfinsen C B. *Science*, 1973, **181**: 223

[4] Shakhnovich E I. *Phys. Rev. Lett.*, 1994, **72**: 3907

[5] Hinds D A, Levitt M. *J. Mol. Biol.*, 1994, **243**: 668

[6] Zhou Y Q, Karplus M. *J. Mol. Biol.*, 1999, **293**: 917

[7] Huang E S, Samudrala R, Ponder J W. *J. Mol. Biol.*, 1999, **290**: 267

[8] Lee J, Liwo A, Scheraga H A. *Proc. Natl. Acad. Sci. USA*, 1999, **96**: 2025

[9] (a) Miyazawa S, Jernigan R L. *Macromolecules*, 1985, **18**: 534
(B) Miyazawa S, Jernigan R L. *J. Mol. Biol.*, 1996, **256**: 623

[10] Ryckaert J P, Ciccotti G, Berendsen H J C. *J. Comp. Phys.*, 1977, **23**: 327

[11] Deutsch J M, Kurosky T. *Phys. Rev. Lett.*, 1996, **76**: 323

[12] Shakhnovich E I. *Fold & Design*, 1998, **3**: R45

[13] Shakhnovich E I, Gutin A M. *Proc. Natl. Acad. Sci. USA*, 1993, **90**: 7195

[14] Bonneau R, Strauss C E M, Baker M. *Proteins*, 2001, **43**: 1

[15] Reva B A, Finkelstein A V, Skolnic J. *Fold. Design*, 1998, **3**: 141

[16] Wang B H, *et al. J. Bioscience*, 1999, **24** (suppl **1**): 61

中国科协 2003 年学术年会通知

中国科协 2003 年学术年会定于 2003 年 9 月 13～16 日在辽宁省沈阳市召开。本届学术年会由中国科协和辽宁省人民政府联合主办，沈阳市人民政府承办。

2003 年学术年会的宗旨：以邓小平理论和“三个代表”重要思想为指导，贯彻党的十六大精神，围绕全面建设小康社会的宏伟目标，组织专家学者研讨科技工作者的历史责任。通过学术交流，展示科技进展和学术观点，为努力发展先进生产力和先进文化，促进社会主义物质文明和精神文明建设贡献力量。

学术年会以各全国性学会、协会、研究会，各省、自治区、直辖市科协，各计划单列市、副省级城市科协，新疆生产建设兵团科协，部分中心城市科协，部分高校科协和企业科协为组团单位，组织科技人员和专家学者报名。报名参加学术年会的代表，需填写个人报名表报组团单位。

中国科协接受团体报名的截止时间为 2003 年 5 月 20 日。各组团单位接受报名的截止时间要相应提前。

会前正式出版论文摘要文集。论文摘要集将收录报名参加年会主题会场和分会场交流的学术论文摘要，每位与会代表限报 1 篇，文集只刊载 2002 年 1 月 1 日至 2003 年 4 月 30 日之间发表的学术论文、科研报告、综述文章的摘要（如在学术期刊上发表过，需注明刊名、期号）。希望全国广大的科技工作者能将自己的最新科研成果展示于此。同时，本文集不保留知识产权，作者可继续向其他刊物投稿。