改进的势能曲面变平法在二维非格点模型中的应用*)

刘景发[1,3] 陈端兵[2] 刘朝霞[1]

(南京信息工程大学计算机与软件学院 南京 210044)[1]

(电子科技大学计算机科学与工程学院 成都 610054)[2] (衡阳师范学院数学系 衡阳 421008)[3]

摘 要 蛋白质结构预测问题是生物信息学中的一个重要问题。缺少一种有效的全局寻优方法是阻碍这一问题解决的关键。势能曲面变平(ELP)法是一种启发式的全局优化方法,是一种推广的 Monte Carlo 方法,已成功地应用于许多优化问题。在 ELP 法的基础上,提出了改进的势能曲面变平(ELP+)算法。将 ELP+算法应用于二维非格点的蛋白质 AB 模型,预测和发现四条链长分别为 13,21,34 和 55 的氨基酸序列的蛋白质结构。数值实验表明,ELP+算法是一种预测蛋白质结构的有效算法。

关键词 蛋白质结构预测,非格点模型,Monte Carlo 方法,ELP 方法

Improved Energy Landscape Paving Method and its Application in 2D Off-lattice Model

LIU Jing-fa[1,3] CHEN Duan-bing[2] LIU Zhao-xia[1]

(Computer and Software Institute, Nanjing University of Information Science and Technology, Nanjing 210044, China)[1]

(School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)[2]

(Department of Mathematics, Hengyang Normal University, Hengyang 421008, China)[3]

Abstract Protein structure prediction problem is one of the most prominent problems of bioinformatics. Lacking powerful optimization method is the key obstacle to solve this problem. The energy landscape paving (ELP) method is a class of heuristic global optimization algorithm and a generation of Monte Carlo method, and has been successfully applied to solving many optimization problems. Based on the ELP method, an improved energy landscape paving (ELP+) algorithm is put forward. The ELP+ algorithm is applied to the 2D off-lattice protein AB model to predict protein structures of four amino acids chains with lengths n=13, 21, 34, and 55, respectively. Experimental results show that the ELP+ algorithm is quite effective in the protein structure prediction problem.

Keywords Protein structure prediction, Off-lattice model, Monte Carlo method, ELP method

1 引言

蛋白质结构预测问题是生物信息学研究的核心内容之一。三联密码的破译使人们掌握了遗传信息从 DNA 到氨基酸序列的复制规律。然而,仅有氨基酸序列是不够的,氨基酸序列还必须形成一定的空间结构,才能真正完成蛋白质的合成,也才能行使其特定的生物学功能。X 射线晶体衍射分析法和多维核磁共振技术,是测定蛋白质空间结构的两种主要实验方法,然而实验方法不仅耗资耗时,还受到实验条件的限制,而且用实验方法测定结构的速度和人类的测序速度之间还存在很大的差距。20 世纪中叶,Anfinsen, Dill 等人根据变性的核糖核酸在一定条件下可以自发地再折叠形成天然酶分子的实验,提出蛋白质的氨基酸序列决定其空间结构的著名论断[1],这一发现揭示了氨基酸序列到蛋白质结构的折叠是一个热力学过程。一般认为蛋白质的天然构形是热力学最稳定的构形,即自由能最小的构形。如果能够建立一个表征蛋白质结构与能量关系的函数,利用最优化方法在蛋白质空间结构中找到能量函数的全局极小点,蛋白质结构预测问题便可以得到解决。然而,由于蛋白质是一个强韧性的大分子体系,其势能表面存在着极多的局部极小点。缺少一种有效的全局优化方法,一直困扰着蛋白质结构研究的进展。为解决这一问题,人们正从两方面进行努力:一是在保持精度的条件下,简化物理模型;二是寻找适合于蛋白质结构预测的全局优化方法。

在蛋白质结构预测中,最简单的模型是疏水-亲水氨基酸格点模型和非格点模型。本文引入了一个由 Stillinger 等[2,3]提出的蛋白质 AB 模型就是一种非格点模型,该模型也仅仅考虑两种氨基酸——疏水氨基酸(用 A 表示)和亲水氨基酸(用 B 表示)。文献[2-5]对该模型的二维形式进行了研究,并给出了不同氨基酸序列的最低能量构形。使用于优化最低能量的方法包括:基于 Metropolis 抽样的 Monte Carlo 方法(MMC)[3]、记忆禁忌搜索方法(MTS)[4]、基于非格点的剪枝-繁殖 Rosenbluth 方法(PERM)[5]以及 PERM 结合共轭梯度算法(PERM+)[5]。这些方法在不同程度上均具有全局优化的能力,是蛋白质结构预测的有效方法。本文介绍另外一种全局优化方法——基于 Monte Carlo 的势能曲面变平(Energy Landscape Paving, ELP)法[6,7]。在此方法基础上,本文提出改进的 ELP 方法,并将之应用到二维非格点的蛋白质 AB 模型进行蛋白质结构预测。计算结果表明,改进的 ELP 方法对于解决非格点模型的蛋白质结构预测问题是有效的。

*)本课题得到国家自然科学基金项目(10476006)和湖南省教育厅杰出青年基金项目(07B009)的支持。**刘景发** 副教授,博士研究生,主要研究方向为生物信息学、NP 难度问题求解、人工智能;**陈端兵** 博士研究生,主要研究方向为 NP 难度问题求解、数据挖掘。

2 二维非格点的蛋白质 AB 模型

对于有 n 个氨基酸残基的蛋白质链，链上任意两个相邻的氨基酸残基之间的距离为 1，链上不相邻的氨基酸残基通过 Lennard-Jones 势能相互作用。相邻的两个键向量（Bond Vectors）之间存在键角（Bond Angles）θ_i（$-\pi\leqslant\theta_i<\pi, 1\leqslant i\leqslant n-2$）变形，$n$ 个残基的构形由 $n-2$ 个键角 $\theta_1,\cdots,\theta_{n-2}$ 或 $n-1$ 个单位键向量 $u_1,u_2,\cdots,u_{n-1}$ 确定。键向量 u_i 和键角 θ_i 的定义如图 1 所示。势能函数[2-5]由二项构成：弯曲势能和非键相互作用的 Lennard-Jones 势能，它可表示为：

$$E=\frac{1}{4}\sum_{i=1}^{n-2}(1+\cos\theta_i)+4\sum_{i=1}^{n-2}\sum_{j=i+2}^{n}[r_{ij}^{-12}-C(\zeta_i,\zeta_j)r_{ij}^{-6}]$$

这里 r_{ij} 是氨基酸 i 和 j（$i<j$）之间的距离，ζ_i 是 A 或 B，当 $\zeta_i\zeta_j$ 分别是 AA，BB，AB（或 BA）时，$C(\zeta_i,\zeta_j)=1,\frac{1}{2},-\frac{1}{2}$，表现为 AA 之间具有强吸引，BB 之间具有弱吸引，AB（或 BA）之间为弱排斥。

二维 AB 模型的蛋白质结构预测问题更为形式化的提法是：已知氨基酸序列 $s=\zeta_1\zeta_2\cdots\zeta_n$，求解带有约束条件的优化问题：$\min\limits_{X\in G(s)}E(X)$，这里 $G(s)$ 是 s 所有相邻的两个残基距离为 1 的合法构形的集合。

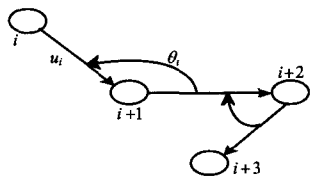

图 1 键向量 u_i 和键角 θ_i 的定义

3 改进的势能曲面变平法

势能曲面变平（ELP）法[6,7]是一种 Monte Carlo 方法，该方法融合了势能曲面变形（Energy Landscape Deformation）法[8]和禁忌搜索（Tabu Search）[9]的思想，它通过修改势能函数使搜索避开最近访问过的区域，也就是说，如果某一构形 X 被采样，则其相应的势能 $E(X)$ 将增加一个“惩罚”：用势能 $\tilde{E}=E+f(H(q,t))$ 代替 E。这里，$f(H(q,t))$ 是直方图 $H(q,t)$ 的函数，可以定义为 $\ln H(q,t)$ 或 $kH(q,t)$ 等，k 是常数；q 是“有序参数”，最简单的“有序参数”选择是势能本身，即 $q=E$。在我们的模拟中，使用势能 E 作为一个“有序参数”，变形的势能函数定义为 $\tilde{E}=E+H(E,t)$。直方图 $H(E,t)$ 在每次 Monte Carlo 迭代步中都要更新，是“有序参数”和迭代步数 t 的函数。一个构形 X 的抽样权重定义为：

$$w(\tilde{E}(X))=\exp(-\tilde{E}(X)/k_BT)$$

其中 k_BT 是在低温 T 处的热能，k_B 是 Boltzmann 常数。

在 ELP 迭代中，给定相等的罚因子 $f(H(q,t))$，模拟将更有利于搜索低能构形，而不利于高能构形的抽样。然而，计算将很快陷入局部极小点。随着计算呆在极小点“时间”t 的延长，惩罚项 $f(H(q,t))$ 将增加，以致于局部极小点的抽样权重 w 将减少而不再优先被采样；此时，计算将搜索更高能量的区域，而跳出该局部最优点。ELP 采用这种方法将局部势能曲面变平。然而，当计算跳出局部最优以后，它将很快又陷入一个新的局部最优点，再利用局部势能曲面变平的方法跳出该极小点。这种过程不断重复，直到原始的势能曲面被慢慢变平。

ELP 法有些类似于禁忌搜索。在禁忌搜索中，最近被访问过的区域不可能立即再被访问。而在 ELP 中，重新访问不是完全被禁止，只是相对于那些在能量上是够低的但很少被访问的区域有更低的抽样权重。

不难发现，在 ELP 法[6,7]中，为了跳出局部极小点而引入惩罚项 $f(H(q,t))$，这里直方图函数 $H(q,t)$ 将势能空间中的能量值分成有限的多个“区间”（Bin），这就存在一个技术上的缺陷：假若一个新的以前没有访问过的更低能量的极小点，其能量值正好落在包含早期被访问过的其它构形的能量所处的同一区间内。此时，由于“惩罚”项 $f(H(q,t))$ 的作用，接受这个新的极小点的可能性就会变小，在 ELP 模拟中很可能错过这个极小点。针对这一缺陷，对 ELP 法提出改进：一旦新产生的构形 $X^{(2)}$ 的能量 $E(X^{(2)})$ 低于原来构形 $X^{(1)}$ 的能量 $E(X^{(1)})$，就无条件地接受新的构形 $X^{(2)}$；否则，按照是否满足 $\mathrm{Ran}<\exp\{[\tilde{E}(X^{(1)})-\tilde{E}(X^{(2)})]/k_BT\}$ 而决定是否接受新的构形，这里 Ran 为（0,1）之间的随机数。

改进过的势能曲面变平（ELP+）法的算法描述如下：

Step 1 给定一个初始构形 $X^{(1)}$，令 $t=1$，初始化 $H(E,t)$，设置温度 $T=5K$。计算 $E(X^{(1)},t)$ 和 $\tilde{E}(X^{(1)},t)$；

Step 2 更新当前构形为新的构形 $X^{(2)}$（更新方法见第 4 节）；

Step 3 计算 $E(X^{(2)},t)$ 和 $H(E(X^{(2)}),t)$，令 $\tilde{E}(X^{(2)},t)\Leftarrow E(X^{(2)},t)+H(E(X^{(2)}),t)$；

Step 4 如果 $E(X^{(2)},t)<E(X^{(1)},t)$，则接受 $X^{(2)}$，使之成为当前构形，即令 $X^{(1)}\Leftarrow X^{(2)}$，$E(X^{(1)},t)\Leftarrow E(X^{(2)},t)$；否则，如果满足

$$\mathrm{Ran}<\exp\{[\tilde{E}(X^{(1)},t)-\tilde{E}(X^{(2)},t)]/k_BT\},$$

则仍然接受 $X^{(2)}$ 并使之成为当前构形，否则不接受 $X^{(2)}$，仍使 $X^{(1)}$ 为当前构形。这里 Ran 为（0,1）之间的随机数；

Step 5 如果 $t\geqslant5\times10^6$，则停止迭代；否则，令 $t\Leftarrow t+1$，转 Step 2。

4 构形更新方法

随机产生正整数 $i(2\leqslant i\leqslant n)$，$n$ 为链长。设当前构形为 $X^{(1)}=(X_1^{(1)},X_2^{(1)},\cdots,X_n^{(1)})$，其中 $X_i^{(1)}$ 表示第 i 个氨基酸的坐标。由于当前构形 $X^{(1)}$ 中每一键向量的键长 $|u_i|$ 固定（$|u_i|=1, i=1,2,\cdots,n-1$），第 i 个氨基酸在以第 $i-1$ 个氨基酸为圆心的单位圆上。为了更新第 i 个氨基酸的位置，将第 i 个氨基酸沿其能量的负梯度方向暂时变为 $\tilde{i}$，即 $\tilde{X}_i^{(1)}=X_i^{(1)}-\varepsilon\nabla E_i$，其中 $X_i^{(1)}$ 和 $\tilde{X}_i^{(1)}$ 分别表示氨基酸 i 和 $\tilde{i}$ 的坐标，ε 为步长因子，$\nabla E_i=(\frac{\partial E}{\partial x_i},\frac{\partial E}{\partial y_i})$。连接 $i-1$ 和 $\tilde{i}$，与圆相交于点 i'。正式将氨基酸 i 移动到 i'，第 $i+1$ 至第 n 个氨基酸作为刚体保持整体平移，部分链的方向不变（如图 2 所示），得到新的构形 X'。显然，在构形 X' 中任何两个相邻氨基酸之间的距离仍然为 1。

在 X' 的基础上，再随机产生正整数 $j(2\leqslant j\leqslant n)$。采用上述类似的方法，将氨基酸 j 移动到 j'，第 $j+1$ 至第 n 个氨基酸作为刚体保持整体平移，得到新的构形，记为 $X^{(2)}$。这样，在构形 $X^{(1)}$ 的基础上，通过两次应用梯度法产生了新的构形 $X^{(2)}$。在梯度法中，取步长因子 $\varepsilon=0.2$。

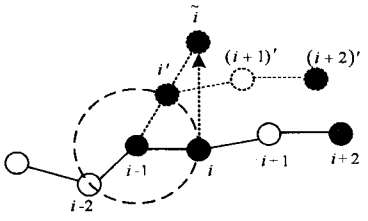

图 2　构形更新

5　计算结果与分析

我们在 Pentium IV 2.0GHz 的 PC 机上用 C++执行 ELP+算法。为了检测算法的效率，用文献[2-5]中的所有 4 条链长 $13 \leqslant n \leqslant 55$ 的 Fibonacci 序列(见表 1)作为测试实例。Fibonacci 序列递归地定义如下：假设 $S_0 = A, S_1 = B$，则 $S_{i+1} = S_{i-1} * S_i$，这里"*"是连结运算。例如，开始的几个序列为：$S_2 = AB, S_3 = BAB, S_4 = ABBAB, \cdots$。这些序列的长度按 $n_{i+1} = n_{i-1} + n_i$，也就是按 Fibonacci 数给出。

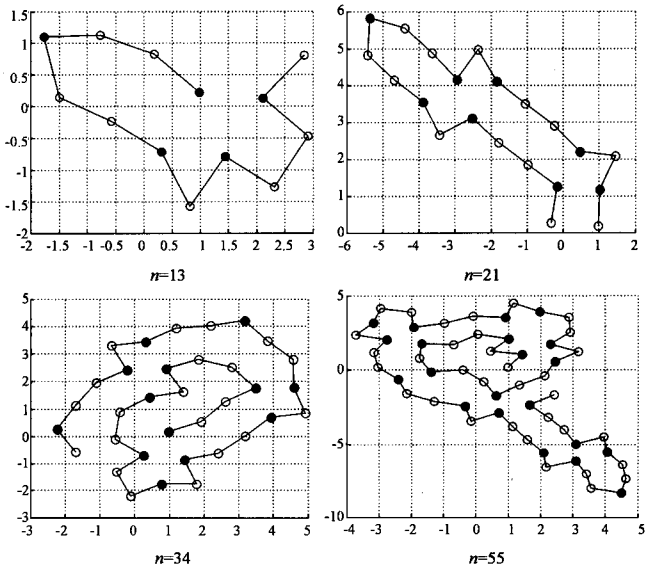

图 3　由 ELP+算法得到的表 1 中给出的 4 条氨基酸序列的最低能量构形

对于每一条序列，分别独立运行 ELP+算法 50 次，将最低能量值列于表 1。由表 1 可见，对于链长为 55 的氨基酸序列，用 ELP+算法得到的最低能量值分别低于基于 Metropolis 抽样的 Monte Carlo 方法(MMC)[3]、记忆禁忌搜索方法(MTS)[4] 和基于非格点的剪枝-繁殖 Rosenbluth 方法(PERM)[5]得到的最低能量。对于链长为 34 的序列，ELP+的计算结果也比 MMC 和 PERM 得到的最优结果要好，但比 MTS 得到的结果要稍差。对于其它两条长度分别为 13 和 21 的序列，ELP+的计算结果则略逊于 MMC，MTS 和 PERM 得到的结果。用 ELP+算法得到的各条序列的最低能量构形如图 3 所示，图中实心圆表示疏水氨基酸(A)，空心圆表示亲水氨基酸(B)。尽管用 ELP+得到的结果与 PERM 结合共扼梯度法(PERM+)[5]得到的最优值相比还有一定的差距，但由文献[5]中图 1 和本文图 3 可以看出，两图中各序列的最低能量构形的结构在本质上是一致的，即疏水氨基酸形成束，总是被亲水氨基酸包围。这表明，ELP+算法仍是一种预测蛋白质结构的有效算法。

表 1　改进的势能曲面变平(ELP+)方法与基于 Metropolis 抽样的 Monte Carlo 方法(MMC)[3]、记忆禁忌搜索方法(MTS)[4]、基于非格点的剪枝-繁殖 Rosenbluth 方法(PERM)[5]和 PERM 结合共轭梯度算法(PERM+)[5]的比较

n	序列	E_{min}^{MMC}	E_{min}^{MTS}	E_{min}^{PERM}	E_{min}^{PERM+}	E_{min}^{ELP+}
13	ABBABBABABBAB	−3.2235	−3.1989	−3.2167	−3.2939	−3.1589
21	BABABBABABBAB-BABABBAB	−5.2881	−6.1828	−5.7501	−6.1976	−5.2688
34	ABBABBABABBAB-BABABBABABBAB-BABABBAB	−8.9749	−9.7018	−9.2195	−10.7001	−9.4239
55	BABABBABABBAB-BABABBABABBAB-BABABBABBABAB-BABABBABBABAB-BAB	−14.4089	−14.5608	−14.9050	−18.5154	−14.9456

结束语　本文扼要地介绍了一种全局优化方法——势能曲面变平(ELP)法，在此方法基础上提出了改进的势能曲面变平(ELP+)法。将 ELP+应用于二维非格点的蛋白质 AB 模型去预测和发现蛋白质结构，数值实验表明 ELP+算法是一种预测蛋白质结构的有效方法。尽管测试的蛋白质 AB 模型有一些不真实的因素，这主要体现在模型中仅仅包含两种氨基酸且为 Fibonacci 序列，而且没有考虑扭曲势能，但这并不影响它作为蛋白质结构预测的一个较好的简化模型。当然，尽管 ELP+算法具有较强的全局寻优能力，但由于它是一种随机抽样方法，因此仍容易错过最优点而陷入一些局部极小点。今后，我们打算提出新的抽样法，进一步改进 ELP+法，并将之应用到具有真实能量的全原子模型中去，为蛋白质结构预测问题设计出具有更高性能的各种新的求解算法。

参 考 文 献

[1] Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, 181(96): 223-230

[2] Stillinger F H, Head-Gordon T, Hirshfeld C L. Toy model for protein folding. Phys. Rev. E, 1993, 48 (2): 1469-1477

[3] Stillinger F H, Head-Gordon T. Collective aspects of protein folding illustrated by a toy model. Phys. Rev. E, 1995, 52 (32): 2872-2877

[4] Yue X H, Tang H W, Guo C H. A tabu search and its application in 2D HP off-lattice model. Computer and Applied Chemistry, 2005, 22 (12): 1101-1105

[5] Hsu H P, Mehra V, Grassberger P. Structure optimization in an off-lattice protein model. Phys. Rev. E, 2003, 68 (3)

[6] Hansmann U H E, Wille L T. Global optimization by energy landscape paving. Phys. Rev. Lett, 2002, 88 (6)

[7] Schug A, Wenzel W, Hansmann U H E. Energy landscape paving simulations of the trp-cage protein. J. Chem. Phys, 2005, 122 (19)

[8] Besold G, Risbo J, Mouritsen O G. Efficient Monte Carlo sampling by direct flattening of free energy barriers. Comput. Mater. Sci, 1999, 15 (3): 311-340

[9] Cvijovic D, Klinowski J. Taboo search: An approach to the multiple minima problem. Science, 1995, 267 (3): 664-666