求解 HP 模型蛋白质折叠问题的启发式算法*)

陈 矛 黄文奇

(华中科技大学计算机科学与技术学院 武汉 430074)

摘 要 构造了一个新的数学模型,把三维 HP 模型的蛋白质折叠问题由一个有约束的优化问题转化为无约束的优化问题,通过建立相对坐标和邻域结构,提出了一个局部搜索算法,并对文献中的链长不同的 7 个算例进行了测试。结果表明,该算法能在较短时间内找到其中 5 个算例的最优能量构形,对另外 2 个难例,则可以找到能量仅比最优构形高一个单位的次优构形。

关键词 蛋白质折叠,HP 模型,启发式算法

Heuristic Algorithm for Protein Folding Problem of HP Model

CHEN Mao HUANG Wen-Qi

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract A heuristic algorithm is proposed in this paper for the HP model protein folding problem. By constructing a new mathematical model, the three dimensional protein folding problem of HP model is converted from a nonlinear constraint-satisfied problem to an unconstrained optimization problem, which can be solved by the local search strategy based on the relative coordinate system and a novel neighbor structure. The computational results on seven benchmark sequences have shown that our algorithm is quite efficient, which can find optimal conformation for five sequences and near-optimal conformation for the other two harder sequences in very short time.

Keywords Protein folding problem, HP model, Heuristic algorithm

1 引言

蛋白质折叠问题,即蛋白质结构预测问题,是生物信息学领域的核心问题之一。蛋白质所具有的生物学功能在很大程度上取决于蛋白质的空间折叠结构,因此了解蛋白质的空间结构在生物学领域具有重要意义。通过实验手段,可以测出蛋白质链的构成,但要观测蛋白质的空间结构却非常困难[1]。Anfinsen 及 Dill 等人[2,3]的研究表明,根据蛋白质的氨基酸序列和能量模型,利用理论计算的方法来对蛋白质结构进行预测是一个可行的方法。目前,该方法已成为蛋白质工程中的一个重要工具。

由于真实蛋白质折叠问题的复杂性太高,理论界提出了一些简化模型。其中,研究最广泛的是 Dill 等提出的 HP 格点模型[4,5]。该模型把二十种氨基酸分为两类:疏水型氨基酸(H)和亲水型氨基酸(P)。每个氨基酸单体可以看作一个小球,H 记为黑球,P 为白球。这样,一条氨基酸序列可以看作是一个由黑白球组成的链,相邻两球的球心距离为 1。HP 格点模型把氨基酸链无重叠地摆放到二维或三维单位网格上,要求每个球必须放到一个格点上,且链上相邻的两球在放到格点平面或空间后位置仍相邻。HP 模型的能量函数为

$$E=\sum_{\substack{i,j=1\\ i<j-1}}^{n}\sigma_{ij} \tag{1}$$

其中,n 是链长。若 i,j 都是黑球且它们的球心距离为 1 时,$\sigma_{ij}=-1$;其它情况下,$\sigma_{ij}=0$。

虽然 HP 格点模型是最简单的简化模型,但对该模型蛋白质折叠问题的求解依然是困难的。该问题已被证明是 NP 完全问题,这意味着不存在既完整严格又不是太慢的求解算法[6]。因此,本文提出了一种启发式算法来求解三维 HP 模型的蛋白质折叠问题。对多个算例的测试结果表明,本文算法具有较高效率。

2 启发式算法

2.1 原始数学模型

在三维格点空间中,把氨基酸序列看作是一条由 n 个被确定了黑白颜色的、半径为 1/2 的球组成的链。对于第 i ($i=1,2,\cdots,n$)个球,它的球心坐标为(x_i, y_i, z_i)。由于球心位于格点上,因此球心坐标 x_i, y_i, z_i 都是整数。我们把这 n 个球的球心坐标的总体 x_1, y_1, z_1, $\cdots$, x_n, y_n, z_n 称为一个构形。蛋白质折叠问题就是问如何调整这 n 个球在格点空间中的位置,使得链上相邻两球球心之间的距离为 1,并使得能量函数 E 取得最低值。它的数学形式为:

$$\text{Min}\ (E) \tag{2}$$

并要求

$$\sqrt{(x_i-x_{i+1})^2+(y_i-y_{i+1})^2+(z_i-z_{i+1})^2}=1$$

x_i,y_i,z_i 都是整数,$i=1,2,n-1$ (3)

式(3)是保证链上相邻两球球心之间的距离为 1。我们把满足约束条件(3)的构形称为合法构形。可以看出,式(2)、(3)构成了一个非线性约束优化问题,它就是本文蛋白质折叠问题的数学模型。对该非线性约束优化问题,由于解空间不够连续、光滑,直接求解比较困难。

通过进行拟物处理[7],引入新的能量并构造新的数学模型,本文把基于三维 HP 模型的蛋白质折叠问题由一个有约束的优化问题转化为无约束的优化问题,从而方便求解。

2.2 新模型及其能量函数

在具有最低能量的蛋白质结构中,疏水氨基酸单体彼此

*)国家自然科学基金资助项目(10471051)和国家 973 计划资助项目(2004CB318000)。陈 矛 博士生,主要研究方向为算法设计。黄文奇 教授,博士生导师,主要研究领域为 NP 难问题现实求解。

靠近，抱成一团，形成一个结构稳定的核，隐藏在蛋白质分子内部。疏水氨基酸之间的较强吸引力是形成核并使得蛋白质结构稳定的主要原因[8]。为了更确切地描述疏水氨基酸单体之间的相互吸引力，引入 $u_{引ij}$ 来描述任意两黑球 i 和 j 之间的引力势能[9]：

$$u_{引ij}=\begin{cases}-\dfrac{k_1}{r_{ij}^{21}}, 若\ r_{ij}\geqslant 1\\ M, 若\ r_{ij}<1\end{cases} \tag{4}$$

其中 r_{ij} 是 i 和 j 球心之间的距离，k_1 是引力系数，M 是正实数。

在调整球心位置时，构形中可能会有两个球彼此嵌入的情形，此时构形不满足式(3)，称为不合法构形。我们把球体想象为弹性实体[7]，根据弹性力学，受到挤压的物体都有要恢复自己形状大小的趋势，我们引入 $u_{斥ij}$ 来描述构形中任意两球 i 和 j 之间的斥力势能：

$$u_{斥ij}=\begin{cases}0, 若\ r_{ij}\geqslant 1\\ M, 若\ r_{ij}<1\end{cases} \tag{5}$$

这样，构形的体系能量 U 就包括引力势能和斥力势能两部分：

$$U=U_{引}+U_{斥}=\sum_{\substack{i=1\\ i是黑球}}^{n}\sum_{\substack{j>1\\ j是黑球}}^{n}u_{引ij}+\sum_{i=1}^{n}\sum_{j>i}^{n}u_{斥ij} \tag{6}$$

从式(4)～(6)可以看出，体系能量 U 是球心坐标 x_1，y_1，z_1，…，x_n，y_n，z_n 的已知函数：

$$U=U(x_1,y_1,z_1,\cdots,x_n,y_n,z_n) \tag{7}$$

基于新的能量函数，蛋白质折叠问题就转化为对总势能的优化问题，优化目标为找到一个具有最小能量的构形（x_1^*，y_1^*，z_1^*，…，x_n^*，y_n^*，z_n^*）。显然，该问题是一个无约束优化问题。对该问题的求解，将会比求解原始问题容易。

2.3 启发式算法

定义（相对坐标） 在任一构形中，球 i 的球心坐标为（x_i，y_i，z_i），球 $i+1$ 的球心坐标为（x_{i+1}，y_{i+1}，z_{i+1}）。若 $x_{i+1}=x_{i+1}$，则定义球 $i+1$ 的相对坐标为右；若 $y_{i+1}=y_{i+1}$，则定义球 $i+1$ 的相对坐标为上；若 $z_{i+1}=z_{i+1}$，则定义球 $i+1$ 的相对坐标为前。依此类推，可知球 $i+1$ 的相对坐标为左、下以及后的情形。

我们把构形第一个球放到坐标原点(0,0,0)处，则从第二个球开始，所有球的坐标都可以用相对坐标表示。利用相对坐标，构形 X 就可以由一串字符表示：$X=$（SUDRF…BLRD）。其中，S是表示第一个球，U,D,R,L,F,B则是其它各球的相对坐标，它们分别表示上，下，右，左，前和后。

在构形 X 中随机选取球 i（$2\leqslant i\leqslant n$）并改变 i 的相对坐标，将会导致球 i～球 n 各球位置的变化，从而得到一个新的构形。由于 i 是随机产生的，并且球 i 的相对坐标有5个变化位置可以选择，这样，X 就有 $5(n-1)$ 个可供选择的邻域构形。通过改变构形中一个或两个球的相对坐标来产生新的构形，并利用局部搜索策略，可以达到搜寻最低能量构形的目的。具体算法如下：

(1) 令 $k_1=10^6$，$M=10$，计数器 $T=0$，并指定计数终止值 $T_{end}=100$。随机产生一个初始构形 X_0，$X\leftarrow X_0$；

(2) 若 $T<T_{end}$，随机选一球 $i(i\neq 1)$，试探性地改变球 i 的相对坐标得到构形 X'，转(4)；

(3) $T\leftarrow 0$，转(5)；

(4) 若 $U(X')<U(X)$，$X\leftarrow X'$；否则，$T\leftarrow T+1$，$k_1\leftarrow k_1-2000$，$M\leftarrow M+0.02$，转(2)。

(5) 若 $T<T_{end}$，随机选两球 i 和 $j(i,j\neq 1)$，试探性地改变 I 和 j 的相对坐标得到构形 X'，转(7)；

(6) 按式(1)计算构形的能量，停机退出；

(7) 若 $U(X')<U(X)$，$X\leftarrow X'$；否则，$T\leftarrow T+1$，$k_1\leftarrow k_1-2500$，$M\leftarrow M+0.03$，转(5)。

在算法开始时，我们把 k_1 取得很大，这样使得较强的吸引力引导构形向能量低的状态演化。在搜索过程中，逐渐减小 K_1，并慢慢增加 M，增加对不合法构形的惩罚，使得构形逐渐向合法构形演化。这样，算法结束时，构形就变成一个合法构形，且能量也很低。

3 算例与实验结果

我们在2.4GHz的P4 PC(内存512M)上用C语言实现了本文算法，并对文[8,10,11]中的7个经典算例(见表1)进行了计算。其中算例5～7是公认的难例。对每个算例，表中给出了目前文献中报道的最低能量 E_{min}。

从表1可以看出，对算例1～5，本文启发式算法可以找到算例的最低能量构形。对算例6和7，本文算法可以找到次优构形，能量仅比最低能量高一个单位。算例5虽然链长不长，但却是一个难例，其最低能量构形仅在文[10]中报道过。在269.4秒内，本文算法就可以找到算例5最低能量构形。从表中可以看出，对所有算例的计算均在270秒以内完成。

图1～4给出了利用本文算法得到的算例3,4以及算例5,7的最低能量构形，图中黑球表示H，白球表示P。

表1 算例及计算结果

序号	链长	HP链	E_{min}	本文稿得到的能量	时间(秒)
1	13	HPPHPPHPHPPHP	−5	−5	36.2
2	20	PHHHHHHPHHHHPHHPHHPP	−13	−13	45.7
3	21	PHPHPPHPHPPHPPHPHPPHP	−10	−10	31.8
4	34	HPPHPPHPHPPHPPHPHPPHPHPPHPPHPHPPHP	−19	−19	78.5
5	46	PHHHPHHHPPPHPHHPHHPPHPHHHHPHPPHHHHHPHPHHPPHHP	−34	−34	269.4
6	48	PHPPPPHPHHHPHPHHHHPHHPHHPPPHPHPPPHHHPPHHPPHHPPPPH	−32	−31	185.2
7	55	PHPHPPHPHPPHPPHPHPPHPHPPHPPHPHPPHPPHPHPPHP	−32	−31	234.7

注：E_{min} 是文献中已报道的最低能量

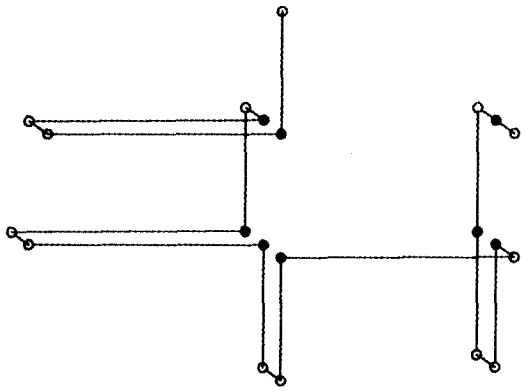

图 1　算例 $n=21$ 的最低能量构形，$E=-10$

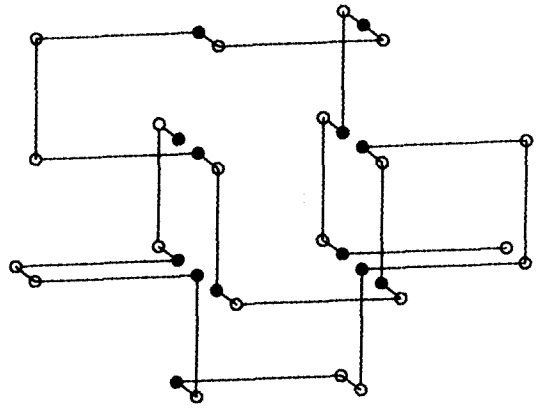

图 2　算例 $n=34$ 的最低能量构形，$E=-19$

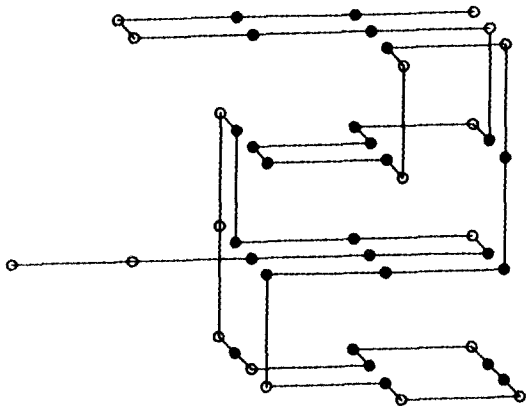

图 3　算例 $n=46$ 的最低能量构形，$E=-34$

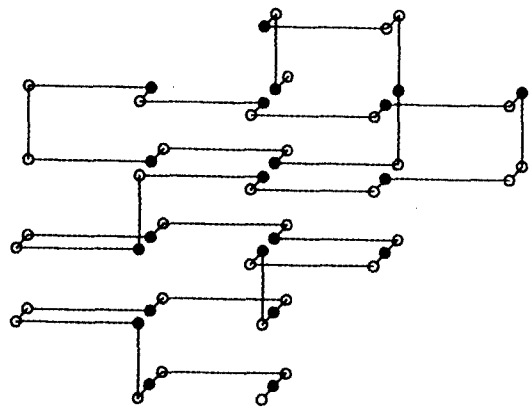

图 4　算例 $n=55$ 的最低能量构形，$E=-31$

结束语　受物理世界中物体间相互作用的规律的启发，本文对 HP 模型的蛋白质折叠问题提出了一个新的数学模型，并给出了一个启发式算法。实算结果表明该算法效率较高。沿着启发式的思路，今后我们将进一步提高算法效率，并把它应用到其它模型的蛋白质折叠问题之中。

参 考 文 献

1　Huang Wenqi, Lu Zhipeng. Personification algorithm for protein folding problem: Improvements in PERM. Chinese Science Bulletin, 2004, 49(19): 2092～2096

2　Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, 181: 233～240

3　Dill K A. Theory for the folding and stability of globular proteins. Biochemistry, 1985, 24: 1501～1509

4　Dill K A, Bromberg S, Yue K, et al. Principles of Protein Folding: A Perspective from Simple Exact Models. Protein Sci., 1995, 4: 561～602

5　Dill K A, Fiebig K M, Chan H S. Cooperativity in Protein Folding Kinetics. In: Proc. Natl Acad Sci USA. 1993, 90: 1942～1946

6　Crescenzi P, et al. On the complexity of protein folding. Journal of Computational Biology, 1998, 5(3): 409～422

7　黄文奇，许如初. 支持求解圆形 packing 问题的两个拟人策略. 中国科学(E辑), 1999, 29(4): 347～353

8　Michael B, Handan A, Wolfhard J. Multicanonical study of course-grained off-latticed models for folding heteropolymers. Physical Review E 71, 2005, 031906

9　黄文奇，黄勤波，石赫. 求解蛋白质结构预测问题的二维连续模型及其相应的拟物算法. 计算机研究与发展, 2004, 41(11): 959～965

10　Lucio T, Salvatore T. Contact interactions method: A new algorithm for protein folding simulations. Protein Science, 1996, 5: 147～153

11　Yue K, Fiebig K M, Thomas P D, Chan H S, Shakhnovich E I, Dill K A. A test of lattice protein folding algorithms. Proc Nut/Acud Sei, 1995, 92: 325～329

（上接第 44 页）

5　Li Lei, Abe S. A Micro-mobility Scheme based on Explicit Multicast. WM12-3. In: Asia-Pacific Conference on communication and International Symposium on Multi-Dimensional Mobile Communication (APCC2004/MDMC2004), August 2004

6　卢汉成，李津生，洪佩琳. 基于重叠网络的移动 IPv6 快速切换. 电路与系统学报, 2004, 9(5)

7　Min T. A seamless handoff approach of Mobile IP based on dual link. In: First International Conference on Wireless Internet (WICON 2005), July 2005

8　Makela J, Ylianttila M, Pahlavan K. Handoff decision in multiservice networks. In: 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications(PIMRC 2000), 2000, 1: 655～659

9　McNair J, Zhu Fang. Vertical handoffs in fourth-generation multinetwork environments. IEEE Wireless Communications, 2004, 11(3): 8～15

10　Goff T, Moronski J, Phatak D S, et al. Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments. IEEE INFOCOM'00, Tel-Aviv, Israel: IEEE Computer and Communications Societies, March 2000

11　Thomson S, Narten T. IPv6 Stateless Address Autoconfiguration. Internet RFC 2462, December 1998

12　Nakajima N, Dutta A, Das S, Henning Schulzrinne. Hndoff Delay Analysis and Measurement for SIP based mobility in IPv6. IEEE International Conference on Communications(ICC 2003), 2003, 2: 1085～1089