求解蛋白质结构预测问题的三维连续模型及其相应的拟物拟人算法*

武汉华中科技大学计算机科学与技术学院(430074) 黄文奇 杨朝阳

摘 要：提出了蛋白质结构预测问题的三维欧氏空间连续模型，为由此形成的数学问题找到了相应的物理模型，进而找到了相应的拟物拟人算法。

关键词：蛋白质结构预测 拟物拟人算法 引力势能

由于生物功能的主要体现者是蛋白质，而蛋白质所具有的生物学功能在很大程度上又与其空间结构相关，因此蛋白质结构的确定将是研究其生物功能的基础。目前虽然有一些技术手段可以测定蛋白质的三维结构，但这些技术都受到相当大的限制，如测定的时间相当长、只限于较小的蛋白质的结构测定等。由于蛋白质分子三维结构测定的速度仍远远落后于其氨基酸序列测定的速度，因此迫切需要一种不依赖晶体培养、迅速简便易行的确定蛋白质结构的方法。自上世纪50年代以来，C.B.Anfinsen，K.A.Dill及K.F.Lau等人就发现蛋白质的空间折叠结构取决于构成该蛋白质的氨基酸的序列。也就是说蛋白质的空间折叠结构的全部信息都隐藏在蛋白质的线型结构中，即氨基酸序列中。因此，根据蛋白质的氨基酸序列提供的信息从理论上预测其相应的高级结构就成为生物信息学中一个重要的课题。

蛋白质的氨基酸序列是由疏水氨基酸和亲水氨基酸组成的。蛋白质的空间结构有如下特点：疏水氨基酸紧密地聚合在一起。有学者相继提出一些模型，这些模型的共同特征是它们都是离散的整数模型。整数模型简洁优美，但蛋白质结构的预测问题是NP-Hard问题，当链长n较大时，计算时间随着问题规模的增加呈指数型增长，求解变得很困难。受离散整数模型的启发，结合作者在NP-Hard问题求解工作中的经验，为蛋白质结构的预测问题提出了三维连续数学模型。本文详细介绍该模型及相应的拟物拟人算法，最后给出了实算结果和评论。

1 用于蛋白质结构预测的三维连续数学模型

对于任意给定的正整数n和任意地被确定了黑白颜色的标号分别为$(1,2,\cdots\cdots,n)$的n个直径为1的球，如何调整这n个球的球心在三维欧氏空间上的位置才能使这n个球两两互不嵌入，标号相邻的二个球皆相切，并且使全部黑球达到尽可能聚合得紧的状态。

此问题更为形式化的描述是寻求$3n$个实数$x_1,y_1,z_1,\cdots\cdots x_n,y_n,z_n$，使在如下约束关系(1)、(2)得到满足的条件下，式(3)函数达到最小。

$$\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}\geqslant 1 \qquad (1)$$

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

$$U=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}U_{ij} \qquad (3)$$

式(3)中的U_{ij}定义如下：

$$U_{ij}=\begin{cases}-\dfrac{1}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}}, & \text{若标号为 } i,j \text{ 的球同为黑球时；}\\ 0, & \text{否则}\end{cases} \qquad (4)$$

(3)式中的U为略去正常数系数后的诸黑球间的万有引力势能，此势能最小意味着黑球之间聚合得最紧。

在此数学模型中，黑球代表疏水氨基酸，白球代表亲水氨基酸；长度为n的黑白球序列代表蛋白质氨基酸序列。数学模型(1)~(4)的解$(x_1,y_1,z_1,\cdots\cdots x_n,y_n,z_n)$即为所求蛋白质的在三维空间中的折叠结构。

2 求解蛋白质结构预测问题的拟物拟人算法

2.1 简 述

将(1)~(4)式中涉及到的几个直径为1的球想象为光滑的弹性实体，想象第i个球与第$i+1$个球的球心之间有一根原始长度为1的弹簧相连，$i=1,2,\cdots\cdots,n-1$。设想这n个球现在被随机地撒在空间中。于是在不同编号的二球之间会产生如下三种不同类型的作用力：

(1)编号相邻的二球之间的弹簧力；

(2)相互嵌入的二球之间的弹性排斥力；

(3)二黑球之间的万有引力。

* 基金项目：国家973计划项目(编号：G1998030600)资助

在任一时刻，每一球体受到的总外力即除自身之外其他 $n-1$ 个球对它的作用力之和。于是初始时刻以后每一球体在所受外力驱动之下都会进行运动。可以想象，由 n 个弹性球构成的这个体系在不停地运动一段时间后，会逐渐趋于稳定。此时这 n 个球的位置 $(x_1, y_1, z_1, \cdots\cdots x_n, y_n, z_n)$ 即是由(1)~(4)式所描述的蛋白质结构预测问题的精确解或近似解。

在此过程中弹簧拉力与弹性斥力的作用是在运动中逐渐使约束条件(1)和(2)得以满足；而万有引力的作用则是将全体黑球尽量拉拢。

将以上物理描述用数学表达式严格地写完整后，事实上即得到了预期的拟物算法。

2.2 模型中各球的受力分析及其物理意义

(1)第 j 个球施于第 i 个球的弹簧拉力为：

$$F_{拉ji}=k_{弹}\times[(\overline{r_j^{\omega}}-\overline{r_i^{\omega}})-\frac{(r_j^{\omega}-r_i^{\omega})}{|\overline{r_j^{\omega}}-\overline{r_i^{\omega}}|}] \tag{5}$$

式(5)中 $\frac{(r_j^{\omega}-r_i^{\omega})}{|\overline{r_j^{\omega}}-\overline{r_i^{\omega}}|}$ 表示从第 i 球心指向第 j 球心的单位向量。

其中 $k_{弹}$ 为物性系数，意义为二球间弹簧的倔强系数。式(5)的物理意义为关于弹簧拉力的虎克定律。

(2)第 j 个球施于第 i 个球的弹性斥力为：

$$\overline{F_{斥ji}^{\omega}}=k_{斥}\times d_{ij}\times\frac{(r_j^{\omega}-r_i^{\omega})}{|\overline{r_j^{\omega}}-\overline{r_i^{\omega}}|} \tag{6}$$

其中 $k_{斥}$ 为物性系数，意义为弹性球的倔强系数；式(6)的物理意义为关于弹性斥力的虎克定律。d_{ij} 为二球间的嵌入深度：

$$d_{ij}=\begin{cases}1-|\overline{r_i^{\omega}}-\overline{r_j^{\omega}}|, & 若|\overline{r_i^{\omega}}-\overline{r_j^{\omega}}|<1;\\ 0, & 否则。\end{cases} \tag{7}$$

(3)第 j 个球施于第 i 个球的万有引力为：

$$F_{引ji}=\begin{cases}\frac{k_{引}}{|\overline{r_j^{\omega}}-\overline{r_i^{\omega}}|^2}\times\frac{(r_j^{\omega}-r_i^{\omega})}{|\overline{r_j^{\omega}}-\overline{r_i^{\omega}}|}, & 若标号为\ i,j\ 的球皆为黑球\\ & 且|\overline{r_j^{\omega}}-\overline{r_i^{\omega}}|>1;\\ 0, & 否则。\end{cases} \tag{8}$$

其中 $k_{引}$ 为物性系数，意义为牛顿万有引力系数。(7)式的物理意义为万有引力定律。

(4)各球所受合力

第 i 个球 $(i=1,2,\cdots\cdots,n)$ 在任一时刻受到的力 F_i 为除它自身以外的诸球施于它的各种类型的力的合力：

$$\overline{F_i^{\omega}}=\sum_{j=1,j\neq i}^{n}(\overline{F_{弹ji}^{\omega}}+\overline{F_{斥ji}^{\omega}}+\overline{F_{引ji}^{\omega}}) \tag{9}$$

2.3 模型中各球的运动方程

根据低速运动的牛顿第二定律，在忽略掉一些传统力学细节并考虑到计算方便后，可得到如下的体系的运动方程：

$$\overline{r_i^{\omega(t+1)}}=\overline{r_i^{\omega(t)}}+\varepsilon\times F_i^{\omega(t)},\ i=(1,2,\wedge,n) \tag{10}$$

其中 ε 为一个小的正实数常数。

2.4 跳坑的拟人策略

由于计算的初始格局，即初始化时 n 个球心的位置 $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \cdots\cdots x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$ 是随机给出的，因此在计算过程中往往会遇到这样的情况：各球均以其平衡位置为中心微幅振荡，趋于静止，但并不满足式(1)和(2)这二个约束条件。可以把这种情形看作是计算落入了局部极小值。如果把一个氨基酸序列看作一根绳子或者是由 n 个球串在一起而形成的链子，则产生该情形的根源在于这根绳子打了一个或多个结。

为解脱此种困境，纯粹的拟物算法是重新随机地选取初始值 $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \cdots\cdots x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$，然后再进行一轮新的拟物计算。这种计算虽然从原则上讲具有最终收敛的性质，但是实算的经验说明其效率是低的。此时，有前途的办法是提出好的“跳出陷阱”的策略，将计算点从局部极小值的陷阱中取出，而置入到具有更好的前景的位置上，然后接着进行新的拟物计算。这种“跳出陷阱”的策略可通过观察与体会人类的生活经验而得出，因而被称为拟人策略。

在日常生活中人们有这样的经验：如果一根绳子乱作一团，用手抓住绳子的一端用力抖几下，绳子有可能被抖开。借助这个思想，这里用称之为“抖结”的策略来跳出局部极小值：将物性系数 $k_{拉}$ 和 $k_{斥}$ 设置得较大，计算一定次数。因为 $k_{拉}$ 和 $k_{斥}$ 设置得较大，所以球的运动被放大，产生振荡，从而有可能把结解开。

2.5 拟物拟人算法

拟物拟人算法的步骤：

(1)随机地给出 n 个球心的位置 $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \cdots\cdots x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$。

(2) 设置物性系数 $k_{引}=5$，$k_{弹}=k_{斥}=50$，$\varepsilon=0.005$ 以式(5)~(9)为基础，利用式(10)开始计算，在此计算过程中，每计算 200 次物性系数 $k_{引}$ 的值减半。

(3)计算到第 5 000 次时检查是否落入了局部极小值。如果没有落入局部极小值则跳至步骤(5)继续计算；如果是，再检查这是从步骤(1)开始计算以来第几次落入了局部极小值，如果在 5 次以内(含第 5 次)则转步骤(4)进行跳坑，否则跳转到步骤(1)重新开始计算。

(4)设置物性系数 $k_{引}=5$，$k_{弹}=k_{斥}=150$，$\varepsilon=0.005$，计算 50 次，跳转到步骤(2)继续计算。

(5)设置物性系数 $k_{引}=0$，$k_{弹}=k_{斥}=50$，$\varepsilon=0.002$ 继续计算。当算到第 t 次，各球最大的位移小于 10^{-6} 时，统计此时该构形的能量值，如低于目标能量值，则停机，输出此时的 $(x_1^{(t)}, y_1^{(t)}, \cdots\cdots, x_n^{(t)}, y_n^{(t)})$ 即为所求的问题之解；否则跳至步骤(1)继续计算。

3 结果及评论

本文计算了三个测试序列，计算结果如表 1 所示。

表 1　三个氨基酸序列

链长	氨基酸序列	最小势能值
20	BWBWWBBWBWWBWBBWWBWB	−22
50	BBWBWBWBWBBBBWBWWWBWWWBWWW WBWWWBWWWBWBBBBWBWBWBWBB	−49
64	BBBBBBBBBBBBWBWBWWBBWWBBWWBW WBBWWBBWWBWWBBWWBBWWBWBW BWBBBBBBBBBBBB	−81

在二维整数模型中有一种正六边形格点模型，将其推广到三维即为正四面体模型。本文的计算结果中黑球的构形基本上与正四面体模型是一致的。计算结果表明三维连续模型以及拟物拟人策略对求解该问题是有效的。

研究的目的是利用三维欧氏空间的连续模型为蛋白质结构的预测工作提供有力的新算法。连续物理模型比离散整数模型更真实，更接近于自然界的本来面貌。再者，在连续模型之下的拟物拟人算法，与近年国际学术界广泛流行的模拟退火算法、遗传算法等有很好的相容性与互补性。在此拟物拟人法的基础上，恰当地结合退火和遗传的思想，预计有可能为蛋白质结构预测问题设计出更高性能的新的求解算法。

参考文献

1　黄文奇，詹叔浩.求解 packing 问题的拟物方法.应用数学学报，1979；2(2.2)

2　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)

3　靳利霞，唐焕文.蛋白质结构预测方法简述.自然杂志，2003；23(4)

4　Wenqi H，yan K.A heuristic quasi-physical strategy for solving disks packing problem.Simulation Modeling Practice and Theory，2002；(10)

5　倪红春，卫翼飞.基于遗传算法的蛋白质折叠模拟系统.上海大学学报(自然科学版)，2001；7(4)

(收稿日期：2003-11-07)