doi:10.3969/j.issn.1672-5565.2016.02.08

基于 HP 模型的蛋白质折叠问题的研究

史 小 红
(西安工业大学理学院,西安 710032)

摘 要:基于蛋白质二维 HP 模型提出改进的遗传算法对真实蛋白质进行计算机折叠模拟。结果显示疏水能量函数最小值的蛋白质构象对应含疏水核心的稳定结构,疏水作用在蛋白质折叠中起主要作用。研究表明二维 HP 模型在蛋白质折叠研究中是可行的和有效的并为进一步揭示蛋白质折叠机理提供重要参考信息。

关键词:蛋白质折叠模拟;HP 模型;遗传算法;蛋白质构象
中图分类号:Q615 **文献标志码**:A **文章编号**:1672-5565(2016)02-112-05

Research on protein folding based on HP model

SHI Xiaohong
(*School of Science*, *Xi' an Technological University*, *Xi' an* 710032, *China*)

Abstract: An improved genetic algorithm for real-life protein folding simulation is proposed based on two dimensional hydrophobic polar (HP) model. The computing results show that the lowest energy conformation with a hydrophobic core and hydrophobic interaction is the main driving force for protein folding. Our studies indicate that the HP model for real-life protein folding problem is effective and reliable, and also provide the important reference information for understanding the overall folding mechanism.

Keywords: Protein folding simulation; Hydrophobic polar model; Genetic algorithm; Conformation of protein

"蛋白质折叠问题"被列为是 21 世纪生物物理学的重要课题之一,蛋白质折叠问题的主要目的是根据蛋白质的氨基酸序列确定其折叠路径和最终的具有蛋白质功能的三维天然结构。Anfinsen 等人的牛胰核糖核酸酶复性实验研究已证明:蛋白质的天然结构是完全由它的一级结构——氨基酸序列决定的。随后,Anfinsen 提出了蛋白质天然构象对应自由能最小结构的著名热力学假说。各种理论预测蛋白质折叠结构的方法都基于这样的理论与实验基础,预测蛋白质折叠结构对在分子水平理解蛋白质折叠的机理具有重要意义。这个问题自 20 世纪中期就被广泛研究,但到目前尚无满意的解决方法。Levinthal 认为,蛋白质折叠问题是极其复杂的,如果通过枚举法搜索自由能最小构象,蛋白质折叠需要耗费接近无穷大的时间长度。因此,无论是对于计算机模拟还是实验研究而言,蛋白质折叠研究仍然是极为困难的事情。近年,根据球状蛋白质折叠结构通常是由一个疏水核心紧密堆积而形成特定的空间结构的特性,提出了疏水作用力是蛋白质折叠的主要因素的亲疏水模型,组成蛋白质的氨基酸残基被简单化分为疏水和亲水两类,也称 HP (Hydrophobic-Polar)模型,HP 模型在二维格子中进行蛋白质折叠结构的计算机模拟成为了研究热点[1]。二维简化 HP 模型的蛋白质折叠研究能够为解决理论预测中面临的如何准确表达势能函数,如何有效搜索构象空间提供参考数据,将会增加我们对蛋白质结构形成过程以及蛋白质结构与功能之间关系的理解,也将会对功能蛋白质的设计和基因工程药物的筛选、研制有重要意义。目前,蒙特卡洛(Monte Carlo, MC)算法[2]、遗传算法[3-4](Genetic Algorithms, GA)和蚁群算法[5-6](Ant Colony Optimization Algorithm, ACOA)等应用在二维简化 HP 模型的蛋白质折叠研究中,这些算法在对一些短肽链的模拟中取得了一定的进展,在短肽链计算

收稿日期:2016-01-26;**修回日期**:2016-03-22.
作者简介::史小红,女,博士,副教授,硕士生导师,研究方向:蛋白质结构预测及蛋白质折叠动力学等;E-mail:ishxh@163.com.

机折叠模拟中找到自由能最小折叠结构对应一个具有疏水核心的稳定结构。但是,这些用于研究的短肽链大都来自一个理想的 HP 序列,没有涉及由 20 个氨基酸组成的一级序列的天然蛋白质结构。最近,Yan 将二维 HP 模型应用在了抗菌肽(Misgurin)折叠构象的研究[7],分析了天然抗菌肽所有 HP 二维构象;研究了甘氨酸(Gly)按照两种不同的疏水标度,其双重属性变化对抗菌肽折叠构象和自由能的影响,他们的研究为蛋白质折叠研究积累了有用的经验。

我们应用 GA 算法对二维 HP 模型的蛋白质折叠进行了计算机模拟,通过在能量函数计算中增加新的惩罚因子对劣质构象进行快速淘汰,提高了优化效率。对测试集的 HP 序列进行模拟研究,同时对天然蛋白质抗菌肽和牛胰岛素的 HP 模型也进行计算机的折叠模拟。研究结果表明:(1)改进的 GA 算法能够快速找到蛋白质折叠的自由能最小构象,自由能最小构象在二维格点图形中总是包含一个最大化的疏水核心,疏水核心结构对应蛋白质的稳定结构,疏水作用在蛋白质折叠中起主要作用。(2)将二维 HP 模型的疏水能量函数应用在真实蛋白质折叠研究,结果显示:自由能最小构象对应疏水核心结构。二维 HP 模型的蛋白质折叠模拟在真实蛋白质折叠研究中是可行和有效的。我们的研究将为蛋白质设计及基因工程等研究提供有效的参考数据,强化我们对蛋白质结构形成过程及蛋白质结构与功能之间关系的理解,促进蛋白质在生物制药等领域的广泛应用。

1　理论与方法

1.1　HP 模型

自然界有多种氨基酸,然而蛋白质中出现的只有 20 种,蛋白质的一级结构就指 20 种氨基酸残基由肽键连接起来的多肽链。其中疏水性表示有些残基侧链的疏水基团避开水的效应,其结果是形成了疏水残基埋藏在蛋白质分子内部,而有些残基侧链是极性的,很容易和水作用形成了极性亲水残基暴露在蛋白质分子与水接触的表面。本文采用适用最广泛的埃氏法(Eisenberg)的疏水标度进行两类 HP 模型的划分[8]。其中氨基酸 CFILMVWYPAG ∈ {H} 疏水集合,DEKNQRSTH ∈ {P} 亲水集合,则蛋白质序列 $\{S|s_1,s_2\cdots s_i\cdots s_n\} \in \{H,P\}^n$,$n$ 为残基数。S_i 代表第 i 个氨基酸残基。能量函数是基于疏水作用是蛋白质折叠的主要作用力的这一普遍共识,因此,我们将序列不相邻但是在结构相邻的两疏水残基之间能量设为-1,其它残基之间能量都为 0,这种能量函数 E 的建立,能够反映蛋白质折叠重要特征[9-10],又能方便计算机模拟去发现在蛋白质折叠过程中的重要规律。

$$E = \sum_{(i\neq j)} E_{S_iS_j}\Delta(\vec{r}_i - \vec{r}_j) \quad (1)$$

$$E_{HH} = -1 \quad (2)$$

$$E_{HP} = E_{PP} = 0 \quad (3)$$

其中 $\vec{r}_i$ 表示第 i 个氨基酸的位置,S_i 表示肽链上第 i 个氨基酸残基。如果 $\vec{r}_i$ 与 $\vec{r}_j$ 非序列直接相邻,则 $\Delta(\vec{r}_i-\vec{r}_j)=1$,否则 $\Delta(\vec{r}_i-\vec{r}_j)=0$。$HH$、$HP$、$PP$ 空间能量计算满足公式(2)和(3)。

在平面建立笛卡尔坐标系 x、y 坐标,其最小单位为整数 1,沿 x、y 坐标等距 1 画出网格连线,每个交叉节点将放置一个氨基酸残基,序列相邻两氨基酸残基在网格中也必须相邻。一个格点的氨基酸残基只有向前、向后、向上、向下四种连接方式,如图 1 所示,分别由随机数 00、11、01 和 10 表示。

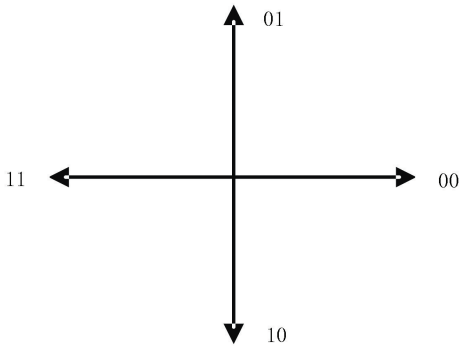

图 1　随机数表示的方向

Fig. 1　The direction of random number

在蛋白质折叠构象搜索中,天然的蛋白质结构必需满足两个或两个以上氨基酸残基不能占据同一个节点,即蛋白质空间结构中不能出现重叠、交叉和回路现象,氨基酸残基之间按顺序连接形成一个完整的肽链。

1.2　GA 算法设计

二维 HP 模型与真实蛋白质比较是一个简化的粗模型,但是,搜索蛋白质可能的构象数目仍随蛋白质序列数呈指数增长,由氨基酸序列搜索自由能最小构象的计算复杂性被证明仍然是 NP 类问题[11]。目前,遗传算法以其高效、实用的特点在蛋白质结构预测中取得较大进展,因此,我们应用 GA 算法进行蛋白质折叠研究。遗传算法(GA)是模仿生物进化机制的一种算法,即在所有可能的问题解中通过适者生存、优胜劣汰的法则找出一个最优解。遗传算法提供了一种求解复杂系统优化问题的通用操作:1)设计染色体编码方式。2)随机生成初始群体。3)计算种群

中每个个体适应度。4)选择优秀的个体复制、交叉、变异操作。5)是否满足优化规则,满足输出最优解。否则,返回 4)。其中交叉操作是最主要的遗传操作,对选中用于繁殖后代的个体,随机选择交叉位置 p,交换两个基因串 p 位置之后的基因串,产生两个新的个体,这两个新个体融合了其父代特征。GA 算法通过执行简单的交叉突变操作可以不断改善数据结构,每一次迭代中保留目标函数最优解,淘汰较差解,GA 可以逾越能量势垒,跳出局部最优搜索到全局最优解。因此,遗传算法已快速应用在蛋白质结构预测和设计中并取得可喜成果[12]。

MATLAB 软件以其强大的图形处理功能和使用简便直观特点,已发展成为适合多学科、跨平台的大型实用科学计算软件,其强大的内置函数,可避免在解决问题中进行繁琐的计算和设计。我们使用 MATLAB7.10 版本实现 GA 算法的步骤如下:

①输入数据。输入蛋白质一级序列 S,内置函数 strrep 将序列 S 转换为{H,P}^{n}序列,n = length(S)为序列长度,strfind 函数找到 H 的序列位置,length 函数记录 H 的个数。

②产生初始种群。使用 randint 函数生成由 0、1 随机数构成的 $2(n-1)\times N$ 的矩阵,每列数据是由 0 和 1 随机数列表示的一种折叠结构,N 表示产生的种群数。

③计算构象势能。记录残基位置坐标及格点数 m,根据格点坐标和能量函数公式计算每种构象的能量值。通过增加惩罚因子 p,增加惩罚能量 $E_0 = 10*p$,用以淘汰产生重叠及回路的结构,当 $m=n$,说明构象中没有重叠或回路发生,取 $p=0$,则 $E_0=0$;当 $m\neq n$,则构象中出现了重叠或回路结构,惩罚因子 $p=(n-m)$,则重叠格点越多的结构,给予惩罚的能量值越大,这些构象被淘汰的概率也越大。E 矩阵记录 N 个构象的能量值。

$$E = \sum_{i\neq j} E_{S_iS_j}\Delta(\vec{r}_i - \vec{r}_j) + E_0 \qquad (4)$$

④计算每个构象的适应度。Max 函数找到 E 中最大值 E_{max},$E' = E_{max} - E$,适应度函数为 f。则能量最小的构象适应性函数 f 对应最大。记录父代最优能量值 E_{best}。使用 figure 函数,记录最优能量值对应构象。

$$f = E'/sum(E') \qquad (5)$$

⑤复制操作。选择适应度大的构象进行复制,使用 sort 函数将种群中个体按 f 值由大到小排序,选择前 k 个初选的优良种群进行复制操作。

⑥交叉操作。选择交叉操作迭代次数 D,用 unidrnd 函数随机产生交叉操作位置,将优良种群两两在交叉位置进行交叉操作,正好产生 k 个不同的子代种群。调用能量函数执行步骤④中程序,对子代的能量值进行由小到大的排序。将新产生的最小能量 E_{mint} 与父代最优能量值 E_{best} 进行比较;如果 $E_{best}<E_{mint}$,保留父代最好构象,返回交叉迭代;如果 $E_{best}<E_{mint}$,用 E_{mint} 替代 E_{best}。使用 figure 函数,记录最优能量值对应构象。

⑦突变操作。给出突变概率 P_t,用 rand 函数产生随机数,如果 rand$<P_t$,执行突变操作。选择突变操作次数 T,用 unidrnd 函数随机产生突变操作位置,用 find 函数找到突变位置数,如果是 1,突变为 0,反之,突变为 1。调用步骤④能量函数程序,计算突变后能量 E_t。如果 $E_{best}<E_t$,保留最优构象,返回突变。如果 $E_t<E_{best}$,用 E_t 替代 E_{best}。使用 figure 函数,记录最优能量值对应构象。

⑧结果判断。判断种群进化是否达到最大迭代数。如果判断结果为否,则返回步骤④。判断结果为是,则此时 E_{best} 为目标函数值的最小值(适应度最大值),为全局最优解。使用 figure 函数,画出全局最优解对应的折叠构象,调用 figure 函数给出遗传算法进化过程图。

2 实验结果

2.1 测试集折叠实验

基于 HP 模型的蛋白质折叠研究是由给定的蛋白质一级结构即 HP 序列出发,折叠为最低能量的稳定构象的过程。测试数据集(见表 1)来自广泛使用的测试 HP 序列[1-2],应用 GA 算法对测试集中的序列作蛋白质折叠的计算机模拟,模拟结果如表 1 所示。图 2 是 HP 序列长度 14,对应最小能量-7 的蛋白质构象和 GA 算法迭代收敛图。图 3 是序列长度 20,最小能量为-9 的蛋白质构象图。显然,能量最小构象中形成了最大化的疏水核心,即疏水核心在稳定蛋白质结构中起重要作用。我们的研究在二维 HP 模型中再次验证了:自由能最小构象对应含疏水核心的稳定结构,疏水作用在蛋白质折叠中起主要作用的理论假设。

2.2 蛋白质折叠实验

蛋白质在合适的条件下能够快速折叠到自由能最低的天然构象,起到稳定结构的重要作用。目前,二维 HP 模型的研究较少涉及由 20 个氨基酸组成的一级序列的天然蛋白质结构,然而,要揭示蛋白质折叠机理就不能回避这个问题,需要对 20 个氨基酸残基在蛋白质折叠过程中所起的作用做深入细致的研究,因此,我们探索性的将 GA 算法的二维 HP 模型对抗菌肽(Misgurin)和牛胰岛素(Bovine Insulin)

B 链进行计算机折叠模拟,分别输入抗菌肽和牛胰岛素的氨基酸序列,进行基于 HP 模型的二维格点的折叠研究。抗菌肽和牛胰岛素 B 链的氨基酸序列及 HP 序列的详细情况见表 2。

表 1　测试的 HP 序列集

Table 1　The set of HP sequence for analysis

序号	HP 序列	序列长度(n)	最优能量值(E)
1	HHHPHPHHPHHPPH	14	-7
2	HPHPPHHPHPPHPHHPPHPH	20	-9

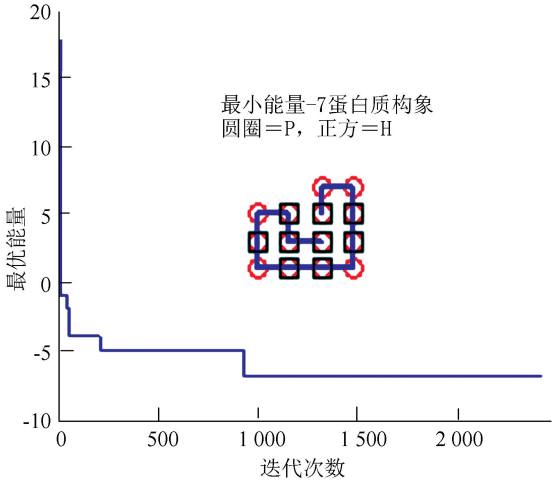

图 2　序列 1 模拟结果

Fig. 2　Simulation results of No.1 sequence

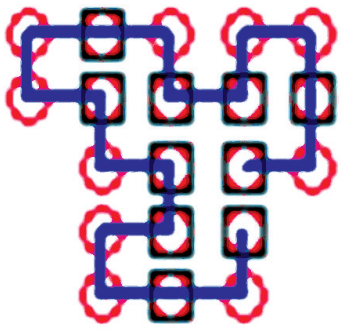

图 3　序列 2 模拟结果

Fig. 3　Simulation result of No.2 sequence

表 2　蛋白质氨基酸序列

Table 2　Amino acid sequence of real-life protein

蛋白质	氨基酸序列	长度(n)	最优能量值
抗菌肽	RQRVEELSKFSKKGAAARRRK PPPHPPHPPHPPPHHHHPPPP	21	-5
牛胰岛素 B 链	FVNQHLCGSHLVEALYLVCGERGFF HHPPPHHHPPHHPHHHHHHHPPHHH	25	-12

抗菌肽(Misgurin)序列由 21 个氨基酸残基组成,是具有很强抗菌活力的短肽,氨基酸序列为:RQRVEELSKFSKKGAAARRRK,使用 GA 算法进行的抗菌肽折叠模拟结果见图 4,结果显示抗菌肽最优能量值为-5,最小能量构象对应一个疏水核心结构,分别由 7L-10F、7L-4V、10F-15A、15A-4V、14G-17A 形成疏水核心,这些疏水氨基酸残基在稳定抗菌肽最小能量构象中起主要作用。

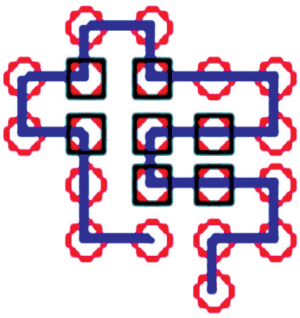

图 4　抗菌肽模拟结果

Fig. 4　Simulation result of Misgurin

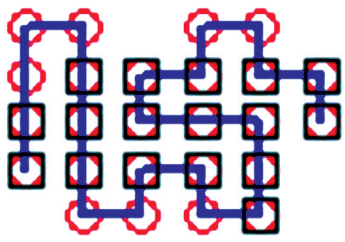

图 5　牛胰岛素 B 链模拟结果

Fig. 5　Simulation result of Bovine Insulin B

牛胰岛素包含 4 条肽链,其中 B 链由 25 个氨基酸残基组成,其氨基酸序列为:FVNQHLCGSHLVEALYLVCGERGFF,经过 GA 算法的蛋白质折叠模拟结果见图 5。牛胰岛素 B 链的疏水核心最大化的聚集在一起对应最小能量-12 的构象。疏水残基 1F-8G、2V-7C、6L-19C、7C-18V、6G-11L、11L-18V、12V-17L、12V-15L、16Y-25F、16Y-23G、23G-20G、17L-20G 在折叠模拟中起到减小构象自由能的主导作用。肽链中疏水核心结构的形成对稳定蛋白质构象有重要意义。

3 总结与展望

应用 GA 算法对二维格点 HP 模型的蛋白质折叠进行了计算机模拟,通过在能量函数计算中增加新的惩罚因子对劣质构象进行快速淘汰,提高了优化效率。通过实验数据测试,改进的 GA 算法能够快速找到蛋白质折叠的自由能最小构象,自由能最小构象在二维格点图形中聚集了一个最大化的疏水核心,疏水核心结构对应蛋白质的稳定结构,疏水作用在蛋白质折叠中起主要作用。同时,对天然蛋白质抗菌肽和牛胰岛素 B 链的二维 HP 模型进行折叠模拟,改进的 GA 算法通过对能量函数的优化,能够快速找到疏水能量函数最小值对应的疏水核心最大化的蛋白质构象。可见,肽链中疏水核心结构对形成稳定紧密的蛋白质构象有重要意义。研究表明二维 HP 模型的蛋白质折叠模拟在真实蛋白质折叠研究中是可行和有效的,将为在分子水平进行蛋白质设计等研究提供参考方法与数据,增强我们对蛋白质结构形成过程及蛋白质结构与功能之间关系的理解,促进蛋白质在生物制药等领域的广泛应用。

参考文献

[1]刘赟,王存新,王宝翰,等.基于格子模型的蛋白质设计方法[J].生物化学与生物物理进展,2004,31(2):172-176.
LIU Yun, WANG Cunxin, WANG Baohan, et al. A Protein design procedure based on the lattice model[J].Progress in Biochemistry and Biophysics,2004,31(2):172-176.

[2]解伟,王翼飞.蛋白质折叠的计算机模拟[J].上海大学学报(自然科学版),2000,6(2):145-149.
XIE Wei, WANG Yifei. Computer simulation for protein folding[J]. Journal of Shanghai University (Natural Science), 2006,6(2):145-149.

[3]UNGER R, MOULT J. Genetic algorithm for protein folding simulations[J].Journal of Molecular Biology,1993,231(1):75-81.

[4]倪红春,王翼飞.基于遗传算法的蛋白质折叠模拟系统[J].上海大学学报(自然科学版),2001,7(4):359-364.
NI Hongchun, WANG Yifei. A system for protein folding simulation based on genetic algorithms[J]. Journal of Shanghai University(Natural Science), 2001,7(4):359-364.

[5]陆恒云,杨根科,潘常春,等.改进的蚁群算法求解蛋白质折叠问题[J].计算机工程与设计,2010,31(8):1786-1816.
LU Hengyun, YANG Genke, PAN Changchun, et al. Improved ant colony optimization algorithm for 2D HP protein folding[J]. Computer Engineering and Design, 2010, 31(8):1786-1816.

[6]SHMYGELSKA A,HOONS H H. An ant colony pptimization algorithm for the 2D and 3D hydrophobic polar protein folding problem[J]. BMC Bioinformatics, 2005,6(1):30-51.

[7]YAN S, Wu G. Analysis on folding of misgurin using two-dimensional HP model[J]. Proteins,2011,10(3):764-773.

[8]阎隆飞,孙之荣.蛋白质分子结构[M]. 北京: 清华大学出版社,1999.
YAN Longfei,SUN Zhirong.Molecular structure of protein[M]. Beijing:Press of Tsinghua University, 1999.

[9]王翼飞,史定华.生物信息学[M].北京: 化学工业出版社,2006.
WANG Yifei, SHI Dinghua. Bioinformatics[M]. Beijing: Press of Chemistry and Technology, 2006.

[10]DILL K A. Principles of protein folding:A perspective from simple exact models[J].Protein Science, 1995, 4(4):561-602.

[11]BERGER B, LEIGHT T. Protein folding in the hydrophobic hydrophilic (HP) model is NP-complete[J]. Journal of Computational Biology, 1998,5(1): 27-40.

[12]SZUSTAKOWSKJ J D,WENG Z.Protein structure alignment using a genetic Algorithm[J].Proteins,2000,38(4):428-440.