

改进的粒子群算法求解蛋白质结构预测问题

焉为家,郭雨珍

(南京航空航天大学理学院,江苏南京 210016)

摘要:蛋白质的生物学功能是由其空间结构决定的,因此,蛋白质结构预测就成为生物信息学领域中极具挑战性的问题之一。粒子群算法是一种新的群智能算法,优势在于简单容易实现,又有深刻的智能背景。在优化领域,粒子群算法适用于求解连续优化问题,而基于HP格点模型的蛋白质结构预测问题是一个离散问题。因此,文中通过借鉴单点调整算法的思想,引入了调整子和调整序的概念,重构了粒子群算法,并用改进的粒子群算法求解了这一典型的离散问题。数值模拟结果说明了算法的有效性。

关键词:蛋白质结构预测;粒子群算法;HP格点模型;组合优化

中图分类号:TQ937

文献标识码:A

文章编号:1673-629X(2011)12-0109-04

Modified Particle Swarm Optimization Algorithm for Protein Structure Prediction Problem

YAN Wei-jia, GUO Yu-zhen

(College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The biological functions of protein are determined by their dimensional folding structures, and protein structure prediction remains one of the most challenging problems in bioinformatics research. Particle swarm optimization (PSO) algorithm is a new group intelligent algorithm. The advantage of it is simple and easy to achieve, profoundly intelligent background. In the optimization field, PSO is suitable for continuous optimization, and protein structure prediction based on 2D HP lattice model is a discrete problem. Therefore, the concepts of adjustment operator and adjustment sequence is introduced to reconstruct PSO by using the ideas of node regulation algorithm. It is proposed to solve the typical discrete problem. The Numerical simulation results indicate that the algorithm is effective.

Key words: protein structure prediction; particle swarm optimization algorithm; HP lattice model; combination optimization

0 引言

蛋白质结构预测就是指从蛋白质的氨基酸序列预测出其空间结构。由于蛋白质的生物学功能在很大程度上依赖其空间结构,因而进行蛋白质结构预测对于理解蛋白质的结构与功能的关系,并在此基础上进行蛋白质复性、突变体设计以及基于结构的药物设计具有重要意义^[1]。解决蛋白质结构预测问题包括两个方面:一是找到一个简化的模型,使搜索的构象空间尽可能的小;二是找到一种有效的算法进行近似求解。Anfisen提出的蛋白质的氨基酸序列完全决定其空间结构的假说是蛋白质结构预测的理论基础。目前主要有两大类预测方法^[2]:一类是基于理论分析的方法,如从头预测法;另一类是基于知识的预测方法,如同源建

模法、折叠识别法。Dill等人提出的HP格点模型就是一个简化的精确模型^[3],化学家们用它来评估蛋白质结构预测方面的新假说,并且用它来检验一个新的折叠算法的有效性,实际上,这个模型已经成为测试折叠算法的一个标准。

粒子群算法(PSO)是Kennedy和Eberhart在1995年提出的,它是一种群集智能优化算法^[4,5],模拟了鸟类的觅食行为,可用于求解大量非线性、不可微和多峰值的复杂优化问题。PSO算法通用性强,只要求解问题是可计算的,无可微性及其他要求,它原理简单,实现容易,需要调整的参数较少,因而得到了学术界的广泛重视,已成为一种重要的优化工具,并成功应用在函数优化、模糊系统控制、神经网络训练和电网规划^[6]等领域。

目前,已有许多近似算法应用到HP模型中进行结构预测,如模拟退火算法、遗传算法^[7]、神经网络算法、弹性网算法^[8,9]等,这些算法各有各的优势,也有相对的不足,目前还没有一种算法完全优于其他算法的

收稿日期:2011-05-16;修回日期:2011-08-21

基金项目:工信部青年科技创新基金(Y1089-081)

作者简介:焉为家(1985-),男,硕士研究生,CCF会员,研究方向为生物智能优化;郭雨珍,副教授,硕士生导师,研究方向为生物信息的优化应用。

报告。粒子群算法在解决 TSP 问题上已经取得了很好的效果^[10],而基于 HP 模型的蛋白质结构预测问题与 TSP 问题属于同一类匹配问题,所以用粒子群算法可以有效地预测蛋白质结构;然而基于 HP 模型的蛋白质结构预测问题又与 TSP 问题有以下三个方面的不同:

- 1) 蛋白质的构象是链状,不要求首尾相连;
- 2) 在原来蛋白质链中相邻的氨基酸所占据的顶点在等距的网格上也相邻;
- 3) 蛋白质结构预测问题是求 H-H 对最多,而 TSP 问题是求路径最短。因此,需要对粒子群算法进行改进。

1 HP 格点模型简介

1.1 HP 格点模型

Dill 等人^[3]于 1995 年提出了一种疏水亲水模型 (Hydrophilic-Hydrophobic Model),它从几何结构、氨基酸字母表和氨基酸相互作用三个角度对蛋白质的结构进行了一定程度的简化。首先,从几何角度上仅考虑 C_α 原子的骨架结构;然后,从字母表角度上将组成蛋白质的 20 种氨基酸按照各自的亲水性和疏水性分为两组:疏水组,用 H 表示,亲水组,用 P 表示;最后,从氨基酸相互作用的角度上认为除相邻氨基酸形成的肽键外,其主要作用是不相邻疏水氨基酸之间的氢键作用,即 H-H 对,它的能量记为-1。这样,一种氨基酸的折叠构象的能量就是所有能形成能量为-1 的键的总数目,而它的天然结构就是平面等距网格中能量最低的折叠构象,即非肽键 H-H 对最多的折叠构象。

1.2 HP 格点模型的整数规划模型

设氨基酸的长度为 n , 网格格点的数目为 m , 当 $m > n$ 时,为非紧致的,当 $m = n$ 时,为紧致的,文中主要讨论 $m = n$ 的紧致情况。若第 i 个氨基酸占据第 j 个格点,则 $x_{ij} = 1$ (否则 $x_{ij} = 0$);令 $N(j)$ 表示格点 j 的网格相邻点去掉解中与 j 直接连接的网格相邻点的集合,则 $|N(j)| = 0, 1, 2, 3$, 表示集合 $N(j)$ 的元素数目,且 $i = H/P \Rightarrow f(i) = 1/0$;令 Y_i 表示第 i 个氨基酸的坐标,则 $Y = (Y_1, Y_2, \dots, Y_n)$ 表示问题的一个解,求 H-H 对的计算公式为:

$$J(Y) = \frac{1}{2} \sum_{j=1}^n \left[\sum_{i=1}^n f(i) x_{ij} \sum_{i \in N(j)} \sum_{i=1}^n f(i) x_{is} \right] \quad (1)$$

则二维 HP 格点模型可表示为如下的整数规划模型:

$$\begin{aligned} & \max J(Y) \\ & \text{s. t } \sum_{i=1}^n x_{ij} = 1, j = 1, 2, \dots, n \\ & \sum_{j=1}^n x_{ij} = 1, i = 1, 2, \dots, n \end{aligned}$$

$$\|Y_i - Y_{i-1}\| = 1, i = 2, 3, \dots, n$$

基于 HP 格点模型的蛋白质结构预测问题是一个组合优化问题,已经证明了它是 NP 难的^[11]。

2 用改进的 PSO 算法求解 HP 格点模型的整数规划

2.1 基本粒子群算法

在基本 PSO 算法中,在一个 D 维的目标搜索空间中有 m 个粒子组成一个群体,其中第 i 个粒子表示为一个 D 维的向量 $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, $i = 1, 2, \dots, m$, 它表示问题的一个解。第 i 个粒子的飞翔速度也用 D 维的向量表示,记为 $V_i = (v_{i1}, v_{i2}, \dots, v_{iD})$ 。第 i 个粒子迄今为止搜索到的最优位置记为 $P_{ibest} = (p_{i1}, p_{i2}, \dots, p_{iD})$, 整个粒子群迄今为止搜索到的最优位置记为 $P_{gbest} = (p_{g1}, p_{g2}, \dots, p_{gD})$ 。整个粒子群通过跟踪个体最优和群体最优来更新自己的速度和位置,在解空间中寻找最优解。PSO 算法迭代公式如下:

$$v_{id}^{t+1} = wv_{id}^t + c_1 r_1 (p_{id} - x_{id}^t) + c_2 r_2 (p_{gd} - x_{id}^t) \quad (2)$$

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \quad (3)$$

w 表示惯性权重因子; c_1, c_2 是学习因子,为非负常数; r_1 和 r_2 是两个独立的介于 $[0, 1]$ 之间的随机数; t 表示迭代次数。

2.2 改进的 PSO 算法

2.2.1 解的表示

设第 i 个粒子的位置向量表示为 $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, 其中 $x_{i1}, x_{i2}, \dots, x_{iD}$ 代表 n 个格点的编号,它表示一个解。

2.2.2 速度表示

PSO 算法的速度迭代公式(2)已不适合求解基于 HP 格点模型的蛋白质结构预测这一离散问题,为此文中借鉴单点调整算法思想^[12]提出了调整子和调整序的概念,重新构造了速度公式,下面首先介绍一下用到的几个概念^[10, 13]:

定义 1: 设 n 个格点的 HP 模型的解为 $X = (x_i)$, $i = 1, 2, \dots, n$, 定义调整子 $T(i_1, i_2)$ 为将解序列中的顶点 x_{i_1} 从 i_1 位置剔除,并插入到解序列的第 i_2 位置,则产生一个新的解 X' , 例如: 解序列为 $X = (1, 2, 3, 6, 5, 4, 7, 8, 9)$, 调整子为 $V = T(5, 2)$, 则 $X' = X + V = (1, 5, 2, 3, 6, 4, 7, 8, 9)$, 这里“+”表示按照调整序列对解序列进行调整。

定义 2: 一个或多个调整子的有序队列构成一个调整序列,称为调整序,记为 $ST, ST = (T_1, T_2, \dots, T_n)$, 式中 T_1, T_2, \dots, T_n 是调整子,它们之间的先后顺序是有意义的,不满足交换律,调整序作用于一个解上等价于这个序列中的所有调整子依次作用于该解上。

定义 3:不同的调整序作用于同一解上有可能产生相同的新解,所有具有相同效果的调整序的集合称为调整序的等价集。

定义 4:若干个调整序可以合并成一个新的调整序,定义 \oplus 为两个调整序的合并算子。

定义 5:在调整序的等价集中,拥有最少调整子的调整序称为该等价集的基本调整序。

例如:可按如下方法构造一个基本调整序,记给定两个解 A 和 B ,需要构造一个基本调整序 ST ,使得 $B + ST = A$,以 $A: (1, 2, 3, 6, 5, 4, 7, 8, 9)$, $B: (5, 6, 3, 2, 1, 4, 7, 8, 9)$ 为例,可以看出

$A(1) = B(5)$,则第一个调整因子就是 $T_1(5, 1)$, $B_1 = B + T_1(5, 1) = (1, 5, 6, 3, 2, 4, 7, 8, 9)$,

$A(2) = B_1(5)$,则第二个调整因子就是 $T_2(5, 2)$, $B_2 = B_1 + T_2(5, 2) = (1, 2, 5, 6, 3, 4, 7, 8, 9)$,

$A(3) = B_2(5)$,则第三个调整因子就是 $T_3(5, 3)$, $B_3 = B_2 + T_3(5, 3) = (1, 2, 3, 5, 6, 4, 7, 8, 9)$,

$A(4) = B_3(5)$,则第四个调整因子就是 $T_4(5, 4)$, $B_4 = B_3 + T_4(5, 4) = (1, 2, 3, 6, 5, 4, 7, 8, 9) = A$,

所以基本调整序 $ST = (T_1, T_2, T_3, T_4)$,此时 $B + ST = A$ 。

2.3 改进的 PSO 算法求解二维 HP 格点模型的思想

首先,初始化粒子群,即给群体中每个粒子赋一个随机的初始解和一个随机的调整序,根据公式(1)计算每个粒子的适应度值,对每个粒子分别比较它的适应度值和它经历的最好位置的适应度值以及群体所经历的最好位置的适应度值,如果更好,更新 P_{ibest} 和 P_{gbest} ;

然后,根据公式 $X_i^{(t+1)} = X_i^{(t)} + V_i^{(t)}$,计算每个粒子的下一个解,若得到的解不满足距离要求,需进行调整,调整策略如下:从链中找出满足要求的最长链 P ,剩余格点组成集合 S ,对 S 中的每个格点 s_j 分别计算将其插入到 P 中所有格点的代价,从中选取代价最小的位置,将 s_j 插入到相应的位置,直到 S 中所有的格点都插入到 P 中为止,注意,可能需要不止一次的调整,当解满足距离要求时,记 $(ST)_i = X_i^{(t+1)} - X_i^{(t)}$,令 $V_i^{(t)} = (ST)_i$;

最后,分别计算 P_{ibest} 和 $X_i^{(t)}$ 的差 A , $A = P_{ibest} - X_i^{(t)}$,计算 P_{gbest} 和 $X_i^{(t)}$ 的差 B , $B = P_{gbest} - X_i^{(t)}$,其中 A 和 B 都是一个基本调整序,根据公式 $V_i^{(t+1)} = V_i^{(t)} \oplus r_1(P_{ibest} - X_i^{(t)}) \oplus r_2(P_{gbest} - X_i^{(t)})$ 计算得到 $V_i^{(t+1)}$,其中 $r_1(P_{ibest} - X_i^{(t)})$ 表示以概率 r_1 保留 A ,即保留 A 中 $r_1(P_{ibest} - X_i^{(t)})$ 个调整子,取整数, $r_2(P_{gbest} - X_i^{(t)})$ 表示以概率 r_2 保留 B ,即保留 B 中 $r_2(P_{gbest} - X_i^{(t)})$ 个调整子,取整数,迭代下去,直到达到最大迭代次数为止。

3 算法分析

为了验证算法的有效性,从氨基酸的序列库中选取几个序列进行模拟。

算例 1:将长度为 20 的序列 HHHHHHPHHHHHH-PHHHHHPHH^[14] 嵌入到 4×5 的网格中,网格第一行从左到右记为 1~5,以此类推,以格点 13 作为坐标原点建立直角坐标系,经程序运行得到了多个适应度值为 12 的最优构象,其中两个如图 1 所示:即 12-13-8-7-2-1-6-11-16-17-18-19-20-15-14-9-10-5-4-3 和 13-8-7-12-17-16-11-6-1-2-3-4-5-10-9-14-15-20-19-18。

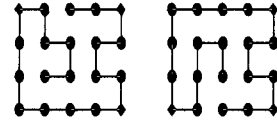


图 1 序列 HHHHHHPHHHHHH-PHHHHHPHH 的最优折叠构象

(图 1~3 中给出的最优构象中,圆点代表疏水性氨基酸分子,而方点代表亲水性氨基酸分子,黑线代表相邻氨基酸的共价键)

算例 2:将长度为 36 的序列 HPHHHHHPHHPPHPH-PHHHPHPHPPHPPHPPHHHHHH^[14] 嵌入到 6×6 的网格中,网格第一行从左到右记为 1~6,以此类推,以格点 21 作为坐标原点建立直角坐标系,经程序运行后得到了多个适应度值为 21 的最优构象,其中两个如图 2 所示:即 35-36-30-29-28-34-33-27-26-32-31-25-19-20-21-15-14-13-7-1-2-8-9-3-4-10-11-5-6-12-18-17-16-22-23-24 和 5-6-12-11-10-4-3-9-8-2-1-7-13-14-15-21-20-19-25-31-32-26-27-33-34-28-29-35-36-30-24-18-17-23-22-16。

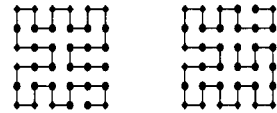


图 2 序列 HPHHHHHPHHPPHPH-PHHHPHPHPPHPPHPPHHHHHH 的最优折叠构象

算例 3:为了验证算法的有效性,构造了一个长度为 25 的序列 HHPPPPHPPPPHPPPPHPPPPH,将它嵌入到 5×5 的网格中,网格第一行从左到右记为 1~5,以此类推,以格点 13 作为坐标原点建立直角坐标系,经程序运行后得到了多个适应度值为 8 的最优构象,可以严格证明,如果不考虑空间上的相似变换,如上的例子的最优解是唯一的,最优构象如图 3 所示:即

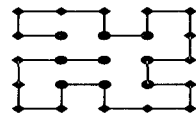


图 3 序列 HHPPPPHPPPPHPPPPHPPPPH 的最优折叠构象

13-12-11-16-21-22-17-18-23-24-25-20-19-14-15-10-5-4-9-8-3-2-1-6-7。

三个算例的数值模拟实验结果如表 1 所示。

表 1 算例结果

算例序号	序列长度	最大迭代次数	最优适应度值	平均用时(s)
1	20	1000	12	18
2	36	2000	21	37
3	25	1000	8	13

结果表明,改进的算法具有较强的可行性和有效性,并且在时间上有了较大的优势。

4 结束语

文中提出了一种改进的 PSO 算法,并借此来求解基于二维 HP 格点模型的蛋白质结构预测问题,数值模拟实验表明,改进的算法具有较强的可行性和有效性,且在时间上有较大的优势。改进的 PSO 算法在解蛋白质结构预测问题时,初始解以及初始调整序的选择决定了下一代解的好坏以及能否很快收敛到最优解。因此,如何恰当地解决个体间的协作和个体经历最优及群体经历最优对个体的影响是解决这一问题的关键,也是进一步研究的问题。

参考文献:

- [1] 王菲露,宋杰,宋杨. BP 神经网络在蛋白质二级结构预测中的应用[J]. 计算机技术与发展,2009,19(5):217-218.
- [2] Luo L F. The time scale of protein folding simple model of chaperous[J]. Acta SNON,1994(25):52-56.
- [3] Dill K A, Bronnberg S, Yue K, et al. Principles of protein folding—a perspective from simple exact models[J]. Protein science,1995(4):561-602.

(上接第 108 页)

参考文献:

- [1] Franses P H. 商业和经济预测中的时间序列模型[M]. 北京:中国人民大学出版社,2002.
- [2] 张善文. Matlab 在时间序列分析中的应用[M]. 西安:西安电子科技大学出版社,2007.
- [3] de Gooijer J G, Hyndman R J. 25 years of time series forecasting[J]. International Journal of Forecasting,2006,22:443-473.
- [4] 于俊年. 计量经济学[M]. 第 2 版. 北京:对外经贸大学出版社,2007.
- [5] Kumar K, Jain V K. Autoregressive integrated moving averages (ARIMA) modeling of a traffic noise time series[J]. Applied Acoustics,1999,58:283-294.
- [6] Tseng Fang-Mei, Yu Hsiao-Cheng, Tzeng Gwo-Hsiung. Combining neural network model with seasonal time series ARIMA

- [4] Kennedy J, Eberhart R C. Particle Swarm Optimization[C]//Proceedings of IEEE International Conference on Neural Networks. Piscataway:IEEE,1995:1942-1948.
- [5] Shi Y, Eberhart R C. Empirical study of particle swarm optimization[C]//Proceedings of the IEEE Congress on Evolutionary Computation. Piscataway:IEEE,1999:1945-1990.
- [6] 唐俊. PSO 算法原理及应用[J]. 计算机技术与发展,2010,20(2):215-216.
- [7] 李绍新,张延娇. 改进的遗传算法在蛋白质结构预测中的应用[J]. 华南师范大学学报(自然科学版),2009(1):56-60.
- [8] Guo Yuzhen, Feng Enmin, Wang Yong. Exploration of two-dimensional HP lattice model by combining local search with elastic net algorithm[J]. The Journal of Chemical Physics,2006,125(15):154102.
- [9] Guo Yuzhen, Feng Enmin, Zhao Jingcheng, et al. Optimal HP Configurations of Protein by Combining Local Search with Elastic Net Algorithm[J]. Journal of Biochemical and Biophysical Method,2007,70(3):335-340.
- [10] 王翠茹,张江维. 改进粒子群优化算法求解旅行商问题[J]. 华北电力大学学报,2005,32(6):47-51.
- [11] Hart W E, Istrail S. Robust proofs of NP-hardness for protein folding general lattices and energy potentials[J]. Journal of Computational Biology,1997,4(1):1-22.
- [12] 刘任任. 算法设计与分析[M]. 武汉:武汉理工大学出版社,2003:121-124.
- [13] Wang K P, Huang L, Zhou C G, et al. Particle swarm optimization for traveling salesman problem[C]//The 2nd International Conference on Machine Learning and Cybernetics. Xi'an: [s. n.],2003:1583-1586.
- [14] Yanikoglu B, Erman B. Minimum energy configurations of the 2-dimensional HP-model of proteins by self-organizing networks[J]. Journal of Computation Biology,2002,9(4):613-620.

model[J]. Technological Forecasting & Social Change,2002,69:71-87.

- [7] Spence M. Job Market Signaling[J]. Quarterly Journal of Economics,1973,87(3):355-374.
- [8] 王琼,刘国祥. 金融时间序列的趋势路径的提取[J]. 南京师大学报(自然科学版),2002,25(2):105-109.
- [9] 刘小君,张立臣. 基于 UML 的实时系统开发[J]. 微机发展(现更名:计算机技术与发展),2003,13(5):81-83.
- [10] 赵纪涛,马莉,王现君. 一种自适应的模糊关联规则挖掘算法[J]. 计算机技术与发展,2008,18(5):64-66.
- [11] 辛治运,顾明. 基于最小二乘支持向量机的复杂金融时间序列预测[J]. 清华大学学报,2008,48(7):1147-1149.
- [12] 但志平,郑胜. 最小二乘向量机在说话人识别中的应用[J]. 计算机技术与发展,2007,17(5):30-32.