离散蛙跳算法预测 RNA 二级结构

林 娟,钟一文,张 骏

(福建农林大学 计算机与信息学院,福建 福州 350002)

[摘要] 针对 RNA 二级结构预测问题,提出了一种离散蛙跳算法,根据 RNA 分子折叠的特点,重新定义个体的移动距离和位置,并借鉴粒子群优化算法中的惯性权重加以改进,使算法在空间探索和局部求精间取得了很好的平衡. 与同领域中著名的预测软件进行了仿真比较,结果表明新的算法具有较高的预测精度.

[关键词] RNA 二级结构预测,离散蛙跳算法,最小自由能,茎区组合优化

[中图分类号] TP301.6; Q7 [文献标志码] A [文章编号] 1672-1292(2011)04-0063-07

Discrete Shuffled Flog Leaping Algorithm for RNA Secondary Structure Prediction

Lin Juan, Zhong Yiwen, Zhang Jun

(College of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou 350002, China)

Abstract: A discrete shuffled flog leaping algorithm is designed for the RNA secondary prediction problem. According to the characteristics of RNA folding, new search space and individual location updating rules are redefined to search the RNA secondary structure with minimal free energy in the combinatorial space of stems. The algorithm is modified by the introduction of inertia weight in particle swarm optimization algorithm (PSO) to get good balance between exploration and exploitation. The simulation results compared with some typical algorithms from the literature show that it can produce higher accuracy.

Key words: RNA secondary structure prediction, discrete shuffled flog leaping algorithm, minimal free energy, combinatorial optimization of stem

分子结构决定分子的性质和功能,RNA 的各种功能与其结构紧密相连,为进一步挖掘其功能,必须从了解 RNA 的结构入手. 然而,用实验方法确定 RNA 的三维空间结构花费高、难度大,且并非对所有分子都有效. 因此利用对已有分子结构和功能特性的认识,通过计算机模拟和计算来"预测"这些结构信息,可以用较低的成本和较短的时间获得具有一定可信度的结果. 生物信息研究者认为,RNA 和蛋白质的三级结构很难通过一级结构直接得到,预测二级结构是获取三级结构的必经之路[1]. 因此关于 RNA 二级结构的预测成为 RNA 结构研究的热点.

根据算法所基于的生物学原理,可将 RNA 二级结构预测算法分为两大类: 基于序列比对的算法和基于自由能最小的算法. 前者必须有一组具有较高序列相似性的 RNA 序列作为比对模板,不适用于对单条 RNA 序列进行预测; 后者从单条序列出发,通过计算 RNA 序列自身碱基配对产生的最小自由能来预测 RNA 的二级结构,具有更大的实用性. 其中,基于自由能最小的 RNA 二级结构预测算法又分为两类: 基于矩阵的动态规划算法和基于最小自由能的茎区组合优化算法. 动态规划算法的缺点是时间、空间复杂度较高,严重制约算法所能处理问题的规模. 茎区组合优化算法将 RNA 二级结构预测转化为离散空间的组合优化问题,被证明是 NP-困难问题[2],而基于种群的智能优化算法作为可以在较短时间内获得最优解(或次优解)的优化工具被引入到预测问题中,取得了一些成果. 其中应用最为广泛的是遗传算法: Van Baten-

收稿日期: 2011-08-10.
基金项目: 福建省自然科学基金项目(2008J0316)、福建农林大学青年教师科研基金(2010018).
通讯联系人: 林 娟,讲师,研究方向: 计算智能、生物信息学. E-mail: sannuo@126.com

burg[3]等讨论如何使用遗传算法进行预测并加以改进,分别采用茎长度及堆叠能量作为适应度函数,算法中包括三级结构、茎及茎的扰动信息,为早期遗传算法解决此类问题提供启示. 任清华[4]等提出用树表示RNA二级结构的方法,并给出混合遗传算法. 在该算法中,个体直接用茎序列编码,与用二进制串编码的同类型算法相比,在很大程度上缩短了个体编码长度. Kay C[5]等首先对比了两个热力学模型,并采用标准遗传算法,设计3种操作算子,并给出了并行版本P-RnaPredict.

少数文献[6-9]采用粒子群优化算法(Particle Swarm Optimization,PSO)解决RNA二级结构预测问题. 文献[6]提出的SetPSO算法通过加入和移除茎来代替传统公式,并通过增加随机元素以维持多样性. 文献[7]提出的HelixPSO算法为每一个粒子设定目标集合并利用遗传算法中的变异操作让粒子向目标集合靠拢. 文献[8]加入局部精英算子优化策略以解决粒子群易早熟问题. 文献[9]利用免疫替代算子提高单个PSO算法摆脱局部最优的能力,并设计了PSO算法的集成模型提高算法全局搜索能力.

何静媛等[10]设计了自适应人工鱼群算法的预测模型. Jun Yu等[11]提出改进的蚁群优化算法,采用启发式信息设计初始化信息素的规则以及更新策略.

随机蛙跳算法(Shuffled Flog Leaping Algorithm,SFLA)于2003年由Eusuff和Lansey最先提出,用以解决供水管网的组合优化问题[12],自出现以来受到广泛关注和应用. SFLA是一种基于种群的元启发式协同搜索算法,具有易理解、易编程实现及寻优能力强等优势[13],作为新兴的群智能优化算法,SFLA展示了较强的局部搜索能力和良好的全局搜索性能.

本文在SFLA算法的基础上,提出一种新的离散蛙跳算法(Discrete Shuffled Flog Leaping Algorithm,DSFLA),根据RNA二级结构预测问题的特点,对个体的位置、位置的相关运算规则和运动方程进行了重新定义,并设置合理的参数平衡DSFLA算法的空间搜索能力和局部求精能力,使算法具有较好的收敛性能及预测精度.

1 相关知识

RNA的二级结构是指RNA碱基序列通过自身回折形成碱基配对的茎区、茎区之间不配对的环区和末端的单链区等. 在RNA二级结构形成的过程中,通常只考虑AU、GC、GU 3种碱基配对. 连续的碱基配对相互堆积构成茎(stem),中间出现少数不配对的碱基形成突环(bulge loop)或内环(interior loop). 相邻连续的一段序列因两端互补而回折,形成像发卡一样的结构,称为发卡环(hairpin loop). 连接各发卡环而未能配对的区域叫做多分枝环(multi-branched loop),序列末端没有形成配对的单链叫做自由单链(unstructured single strand).

几个有关的RNA二级结构数学定义如下:

定义1 一个长度为n的RNA序列$R = r_1r_2\cdots r_n$的二级结构定义为配对碱基的集合$S = \{(r_i,r_j)\}$,其中(r_i,r_j)满足:

(1) $(r_i,r_j) \in \{(A,U),(U,A),(G,C),(C,G),(G,U),(U,G)\}$,$1 \leqslant i < j \leqslant n$,且$j - i > 3$;

(2) 若$(r_i,r_j) \in S$,且$(r_k,r_l) \in S$,则$i = k$当且仅当$j = l$;

(3) 若$(r_i,r_j) \in S$,且$(r_k,r_l) \in S$,$i < k$,它们只有串联和并联两种位置关系,即$i < k < l < j$或$i < j < k < l$.

该定义规定了算法所处理的RNA二级结构只能是常规碱基配对,发卡环环区单链至少为3个碱基长,并且不考虑三联碱基配对以及假结等情况.

定义2 设R是一个长度为n的RNA序列. $R_1 = r_ir_{i+1}\cdots r_{i+k-1}$和$R_2 = r_{j-k+1}r_{j-k+2}\cdots r_j$是$R$的两个子序列. 如果$R_1$和$R_2$的碱基依次互补配对(即$(r_{i+t},r_{j-t})$,$t = 0,1,\cdots,k-1$,且满足定义1中的性质),则称$R_1$和$R_2$在$R$的二级结构$S$中构成一个茎,记为$S(i,j,k)$,其中$i$和$j$分别表示茎在序列$R$中的5′端起始位置和3′端结束位置,$k$表示茎的长度.

定义3 给定两个茎$s_1(i_1,j_1,k_1)$和$s_2(i_2,j_2,k_2)$,若s_1和s_2既不发生重叠也不交叉,则称s_1和s_2相容.

由上述定义可以得到以下两条性质:

性质1 设$S = \{s_1,s_2,\cdots,s_m\}$是RNA序列$R$中的一个茎的集合,若其中任意两个茎都相容,则$S$可以唯一地确定$R$上的一个二级结构. 该性质说明RNA的二级结构可由茎的组合唯一确定.

性质2　设 R 是一长度为 n 的 RNA 序列,S 是由 R 折叠而成的任一个二级结构,则 S 包含的茎个数不超过 $(n-2)/7$.

依据以上的定义和性质,可把 RNA 的二级结构预测问题形式化为一个 RNA 的相容茎区的组合优化问题,即: 对于给定的 RNA 序列,设茎区池 SP 表示该序列所有可能的茎区集合,$E_{\text{total}}(S)$ 表示一个二级结构 S 的总体自由能,则问题描述如下:

求茎区子集 $\{s_{i1}, s_{i2}, \cdots, s_{ik}\} \subset SP$,使得由该子集构成的二级结构 $S^* = s_{i1}, s_{i2}, \cdots, s_{ik}$ 有

$$E_{\text{total}}(S^*) = \min(E_{\text{total}}(S))$$

s. t. 茎区子集 $\{s_{i1}, s_{i2}, \cdots, s_{ik}\}$ 满足相容性条件.

其中,$E_{\text{total}} = E_{\text{stack}} + E_{\text{hairpin}} + E_{\text{bulge}} + E_{\text{internal}} + E_{\text{multi}}$,即一个 RNA 二级结构 S 的总体自由能等于其各结构单元自由能之和[14].

2　*SFLA* 算法

SFLA 的搜索从整个沼泽的青蛙中随机挑选出一个种群开始,种群随即被分为若干个并行的子群(memeplex),青蛙作为子群中文化进化单元(meme)的载体进行思想交流. 算法在每一个子群中朝不同方向进行独立搜索,使青蛙朝子群或种群中最好的方向进化,进化对应于一个跳跃步长.

初始种群 P,每只青蛙表示解空间的一个向量 $X = (x_1, x_2, \cdots, x_F)$,其中 F 表示变量的数目. 青蛙按照适应值降序排列. 整个种群分成 m 个子群,每一个子群里包含 n 只青蛙. 在分组过程中,第一只青蛙进入第一个子群,第二只青蛙进入第二个子群,第 m 只青蛙进入第 m 个子群,以此类推,则 $m+1$ 只青蛙进入第一个子群.

对于每一个子群,寻找适应值最好的和最坏的青蛙,将其分别定义为 X_b 和 X_w. 把整个种群中目前具有最好个体适应值的青蛙定义为 X_g,在每一次迭代中对最坏的青蛙进行类似 PSO 的操作,将最坏青蛙的位置进行如下调整:

青蛙移动的位置:

$$D_i = \text{rand}() * (X_b - X_w) \tag{1}$$

新的位置:

$$X_w = X_w + D_i, \quad D_{\max} \geqslant D_i \geqslant -D_{\max} \tag{2}$$

rand() 是 0 到 1 之间的随机数,$D_{\max}$ 是青蛙允许移动的最大范围. 如果以上操作会产生更好的解,用该解取代原先适应值最坏的青蛙,否则,将 X_b 替换成 X_g,继续使用式(1)和(2)产生新解. 如果适应值仍未得到提高,则随机产生一个新解取代原先最坏的青蛙.

以上操作重复一定迭代次数,直到满足迭代终止条件. SFLA 的算法流程简单、控制参数少,全局寻优能力强,并且具有柔性框架,易于编程实现,所以在很多领域都得到成功应用. 其中,对于组合优化问题,Eusuff 和 Lansey 首次将 SFLA 算法用于求解管道网络扩充中的管径尺寸问题,并在该算法基础上提出 SFLANET 模型[12]. 其后,Eusuff 等[13]采用 5 个标准离散函数测试 SFLA 算法的求解效果,给出了算法最佳参数组合,并用于求解地表水模型标定问题,通过与 GA 的对比显示了算法的有效性. Elbeltagi 等[14]对 5 个进化算法: GA、Memetic 算法、PSO、蚁群优化算法(Ant-Colony Optimization Algrithms,ACO)以及 SFLA 进行了研究,在连续和离散的优化问题上比较其执行时间、收敛速度和优化结果,为用进化算法解决优化问题提供了一些指导. 其他一些应用包括 TSP 问题、考试时间安排问题及零空闲流水线调度和批量无等待流水线调度问题等[15].

3　预测 *RNA* 二级结构的离散蛙跳算法

基于茎区 RNA 二级结构的预测问题,问题的解空间是离散的. 针对茎区的组合优化问题,DSFLA 在经典的算法框架下,设计离散的表示方式表示个体,定义相应操作算子求解,并引入 PSO 惯性权重的调节机制,设计了合理的个体扰动策略.

3.1　个体表示

定义茎区全集 U,即 RNA 单链能够被找出的所有可能的茎区,每个茎区从 1 到 S_{num} 进行编号.

个体矢量的每一维表示一个茎,个体本身表示为一组两两相容的茎,它是一个 K 维向量,定义为 $X=(x_1,x_2,\cdots,x_i,\cdots,x_K)$,其中 $1\leqslant i\leqslant K$,$1\leqslant x_i\leqslant S_{\text{num}}$. 则种群初始化为随机产生的一定数目的个体.

3.2 个体的更新

在标准SFLA中,青蛙位置 X_w 的更新仅为抽象概念,表示青蛙所代表的解向量在连续解空间向其局部极值或全局极值逼近的向量运算. 对于离散的组合优化问题,需要设计具体的更新算子. 应用到RNA二级结构预测:每只青蛙是一个相容茎区的集合,位置的移动定义为集合内茎区的添加删除操作,通过添加 X_b 中存在的茎向好的解靠拢,同样由于茎与茎之间有相容性的约束条件,如果一味向集合内添加元素,在极短时间内将无法再添加新的茎,导致算法迅速收敛. 所以在添加之前,必须做删除的操作. 首先删除原先不存在于 X_b 中的茎,随后再选择 X_b 中的茎加入 X_w. 为该问题设计对应的 $D_{\max}$ 即青蛙最大跳跃步长的替代策略是:首先以概率 P_i 移除在 X_w 而不在 X_b 中的茎,再以概率 P_r 向 X_w 加入在 X_b 中的茎,则 $D_{\max}$ 对应于参数 P_i 和 P_r,通过向个体中加入在 X_b 中的茎元素向优质解靠拢,将式(1)和式(2)重新定义为:

$$X_w = X_w - O, \tag{3}$$

$$X_w = X_b + C, \tag{4}$$

其中,O 表示以 P_i 概率从 X_w 中选出的茎的集合,C 表示以 P_r 概率从 X_b 中选出的茎的集合,"-"表示移除操作,"+"表示添加操作. 如果以上操作会产生更好的解,用该解取代原先适应值最坏的青蛙,否则,将 X_b 替换成 X_g,继续用式(3)和(4)产生新解. 如果适应值还是没能得到提高,则随机产生一个相容茎区的集合取代 X_w.

3.3 个体的扰动策略

在DSFLA算法中,$D_{\max}$ 对应为移除劣质解中茎的概率 P_i 和从优质解中加入茎的概率 P_r. 如果将 P_i 设置为一个较大的固定值,即算法一直保持较大概率移除当前劣质解中的茎,造成较大的扰动,虽然可以在迭代初期增加多样性,但迭代时间随之增加. 且由于组成个体的茎与茎之间相容性约束条件,迭代到一定阶段,个体之间的相似度增加,虽然移除了大部分茎,但加入的茎变化不大,即增加了操作次数,但个体的适应度却得不到提高. 如果将 P_i 设置为较小的固定值,即算法仅从劣质解中移除少数茎,这样虽然可以加快收敛,但新的茎加入到个体中的机会将大大减少,算法极易早熟停滞.

在PSO算法中,惯性权重 ω 控制前面速度对当前速度的影响,为了使得算法在初始搜索时具有较好的全局搜索能力,在迭代后期也具备较强的局部搜索能力以提高收敛精度,通常将 ω 设置为线性下降. 为了平衡搜索精度和收敛速度,借鉴 ω 在PSO算法中的作用,DSFLA将 P_i 设定为线性下降,即迭代初期以较大概率移除劣质解中的茎,使得更多新的茎加入,扩大了搜索空间. 随着迭代次数增加,原先的劣质解朝优质解靠拢,迭代后期移除茎的概率相应减小,搜索在局部范围内进一步细化,提高了局部搜索的精度.

3.4 算法描述

DSFLA算法伪代码为:

设置参数,随机产生一组两两相容的茎集合构成初始种群;

在指定的迭代次数中,重复如下操作:

(1) 评价每个个体的适应值,按适应值大小降序排序并记录 X_g,将排序好的种群分成子群;

(2) 对每一个子群,在预先设定的迭代次数内重复下述操作:

(a) 判断子群中 X_b 和 X_w;

(b) 用式(3)和(4)对 X_w 进行更新;

(3) 将所有子群混合成新一代种群.

4 仿真结果与分析

4.1 测试序列及目标函数

从Genomic tRNA [9] 数据库中随机挑选10条tRNA序列,按序列长度进行编号,以short172为例,1表示第一条序列,长度为72 nt. 从Comparative RNA Website [10] 中挑选3条5SrRNA序列,分别是:Saccharmoyces cerevisiae 5SrRNA(X67579,117 nt),Haloarcula marismortui 5SrRNA(AF034620,122 nt),Thermus aquaticus 5SrRNA(X01590,123nt). 其中,序列名称后的字符是序列在数据库中的编号及序列的长度.

采用与著名动态规划软件 RNAstructure[17] 相同的标准自由能模型 INN-HB(Individual Nearest Neighbour-Hydrogen Bond) ,并采用相同热力学参数计算自由能作为目标函数以便比较.

4.2　参数设置

如其他群智能算法一样,SFLA 的参数选择非常重要. SFLA 主要的参数有: 青蛙的个数 F,子群的数目 m,每个子群内青蛙的个数 n,每个子群内的迭代次数 N 和允许移动的最大步长 $D_{\max}$. 首先,整个种群中青蛙的个数 F($F = m * n$) 是最重要的一个参数. F 与问题的复杂度有关,个体数目越多,搜索到全局最优解的可能性就越大,但是随着个体数目的增加,对适应值函数的计算量也随之增加. 当 F 确定,m 的选择必须确保 n 的值不能太小. 如果子群中的青蛙数目太少,会失去 SFLA 算法中局部进化的优势; 相应的,如果子群内迭代次数太少,青蛙间信息交换随之变慢,收敛速度也会降低. 如果 n 值太大,青蛙会受到大量不必要信息干扰,搜索时间延长. 对于迭代次数 N,如果设得太小,子群频繁地进行混合,局部范围内的信息交换就会变少,反之如果 N 设得太大,每个子群易陷入局部最优解. $D_{\max}$是青蛙跳跃的最大步长,如果设得太小,会降低全局搜索能力,如果设得过大,又会使得算法极易跳过实际上的最优解[15] .

对于参数的设置,目前没有一个指导性的原则,大部分根据实验测得. 根据文献[15] 及相应实验,本文对不同长度的序列设定不同参数: P_i 从 0.8 线性下降到 0.1,对 tRNA 的短序列,设置整个种群内个体个数为 $F = 50$,子群数目 $m = 5$,每个子群中青蛙的个数 $n = 10$,子群内迭代次数 $N = 10$,整个种群的进化次数 IterNum = 50,$P_r = 0.6$. 对 5SrRNA 的长序列,$F = 200$,$m = 20$,$N = 10$,IterNum = 200,其他参数不变.

4.3　仿真实验及结果

为说明算法性能,将实验结果与 RNAstructure、pknotsRG[18] 预测出的结果进行比较.

通常衡量一个算法预测 RNA 二级结构的准确性有如下指标[1]: TP 为正确预测碱基对的个数,FN 为真实结构中存在但未被正确预测出的碱基对个数,FP 为真实结构中不存在却被错误预测到的碱基. 敏感性 SE 指真实结构中所有碱基对中被正确预测到的百分比,且 $SE = TP/(TP + FN)$,特异性 SP 指在所有预测到的碱基对中正确预测的百分比,且 $SP = TP/(TP + FP)$.

对每条序列独立进行 20 次实验,取最好结果进行统计.

表 1　DSFLA 算法与 RNAstructure、pknotsRG 的结果比较

Table 1　Performances compare of DSFLA and RNAstructure, pknotsRG

序列名称	算法名称	*TP*	*FN*	*FP*	*SE*/%	*SP*/%	序列名称	算法名称	*TP*	*FN*	*FP*	*SE*/%	*SP*/%
	DSFLA	16	4	4	80	80		DSFLA	21	3	4	87.5	84
short172	RNAstructure	12	8	15	60	44	short882	RNAstructure	12	12	16	50	42.3
	pknotsRG	12	8	10	60	54.5		pknotsRG	17	7	9	70.8	65.4
	DSFLA	19	1	0	95	100		DSFLA	21	3	3	87.5	87.5
short273	RNAstructure	5	15	17	25	22.7	short984	RNAstructure	18	6	8	75	69.2
	pknotsRG	5	15	15	25	25		pknotsRG	16	8	9	66.7	64
	DSFLA	21	0	0	100	100		DSFLA	20	4	5	83.3	80
short373	RNAstructure	21	0	0	100	100	short1086	RNAstructure	21	3	5	87.5	80.8
	pknotsRG	17	4	3	81	85		pknotsRG	12	12	16	50	42.9
	DSFLA	21	0	0	100	100		DSFLA	33	4	5	89.2	86.8
short473	RNAstructure	21	0	0	100	100	S. cerevisiae	RNAstructure	32	5	7	86.5	82.1
	pknotsRG	21	0	0	100	100		pknotsRG	28	9	14	75.7	66.7
	DSFLA	19	1	0	95	100		DSFLA	31	7	7	81.6	81.6
short573	RNAstructure	5	15	16	25	23.8	H. marismortui	RNAstructure	31	7	7	81.6	81.6
	pknotsRG	5	15	15	25	25		pknotsRG	31	7	5	81.6	86.1
	DSFLA	21	0	2	100	91.3		DSFLA	29	11	7	72.5	80.6
short676	RNAstructure	21	0	2	100	91.3	T. aquaticus	RNAstructure	29	11	14	72.5	67.4
	pknotsRG	21	0	2	100	91.3		pknotsRG	8	32	28	20	22.2
	DSFLA	21	0	0	100	100							
short777	RNAstructure	15	6	8	71.4	65.2							
	pknotsRG	12	9	8	57.1	60							

从表 1 可以看出,DSFLA 对 tRNA 类型短序列的预测精度非常高,有 3 条序列(分别是: short373、short474、short777) 的敏感性和特异性都达到 100% ,即算法能准确无误地预测出真实结构,对于其他序列最低的指数也能达到 80% 以上.

对 5SrRNA 的较长序列,DSFLA 的预测精度也较高. 对于 S. cerevisiae 这条序列,敏感性和特异性分别达到 89.2% 以及 86.8%,注意到 S. cerevisiae 的真实结构中带有 2 个 CU 对,未被列入预测的碱基对范围(GC、AU、GU),这是制约对该序列预测精度的因素,如果能够提供 CU 碱基对相应的自由能参数,应该能得到更好的结果.

与其他算法的性能比较,对于长度在 100 nt 内的短序列,DSFLA 仅对 short1086 这条序列的预测精度不如 RNAstructure,其他预测结果均高于 RNAstructure 和 pknotsRG. 超过 100 nt 的长序列中,仅对于 H. marismortui 的特异性不如 pknotsRG,其余序列的预测精度均超过两者或与之相当.

为观察 DSFLA 求解性能,以 short473(敏感性和特异性均为 100%)和 S. cerevisiae 序列为例,观察算法的收敛情况.

从图 1 和图 2 收敛曲线可以看出,DSFLA 算法具有较高的收敛速度,可以较快地找到最优解. 对每条序列独立运行算法 30 次左右,观察其求解过程,其他序列的收敛情况基本相似,说明 DSFLA 不仅收敛速度快,而且稳定性好,是一种有效解决 RNA 二级结构预测问题的算法.

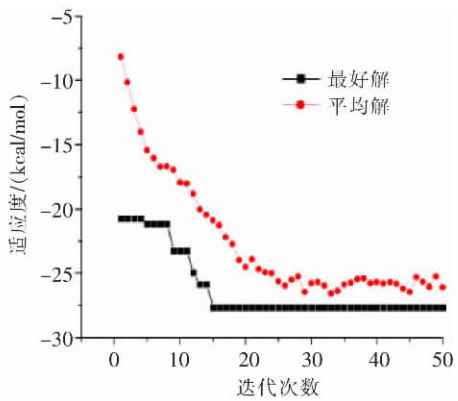

图 1　short473 最好解和平均解的收敛曲线

Fig.1　The convergence curve with short473

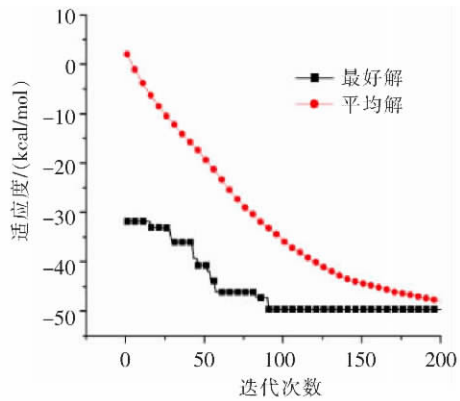

图 2　S.cerevisiae 最好解和平均解的收敛曲线

Fig.2　The convergence curve with S.cerevisiae

5　结论

本文在传统蛙跳算法模型基础上,提出了一种有效的离散蛙跳算法,根据问题的特点,重新定义迭代更新过程,建立求解模型. 通过实例对 DSFLA 进行了实验仿真,结果表明该算法预测精度高、收敛快、需要进行的迭代次数少,对于 tRNA 的短序列的结构预测十分有效,对 5SrRNA 类型的序列也有较好的预测性能,为解决 RNA 二级预测问题提供了新的思路.

[参考文献](References)

[1] 邹权,郭茂祖,张涛涛. RNA 二级结构预测方法综述[J]. 电子学报,2008,36(2):331-336.
Zou Quan, Guo Maozu, Zhang Taotao. A review of RNA secondary structure prediction algorithms [J]. Acta Electronica Sinica, 2008, 36(2):331-336. (in Chinese)

[2] Lyngsa R B, Pedersen C N. Pseudoknots in RNA secondary structures [C] // Computational Molecular Biology (RECOMB'00). New York: ACM, 2000: 201-209.

[3] van Batenburg F H D, Gultyaev A P, Pleij C W A. An APL-programmed genetic algorithm for the prediction of RNA secondary structure [J]. Theor Biol, 1995, 174: 269-280.

[4] 任清华,莫忠良,陶玉敏. 预测 RNA 二级结构的一种遗传模拟退火算法[J]. 武汉大学学报:理学版,2004,50(1):23-28.
Ren Qinghua, Mo Zhongliang, Tao Yumin. A genetic simulated annealing algorithm for RNA secondary structure prediction [J]. Journal of Wuhan University: Nature Science Edition, 2004, 50(1):23-28. (in Chinese)

[5] Kay C W, Alain A D, Andrew G H. RNA Predict: An evolutionary algorithm for RNA secondary structure prediction [J]. Transactions on Computational Biology and Bioinformatics, 2008, 5(1): 25-41.

[6] M Neehling, A P Engelbrecht. Determining RNA Secondary Structure Using Set-based Particle Swarm Optimization: Evolution-

ary Computation [M]. New York: IEEE Press, 2006: 1 670-1 677.
[7] Michael Geis, Martin Middendorf. A particle swarm optimizer for finding minimum free energy RNA secondary structures [C] // Swarm Intelligence Symposium 2007. New York: IEEE Press, 2007: 1-8.
[8] 陈自郁,何中市,何静媛. 预测 RNA 二级结构离散粒子群优化算法 [J]. 深圳大学学报: 理工版, 2009, 26(3): 272-277.
Chen Ziyu, He Zhongshi, He Jingyuan. Discrete particle swarm optimization algorithm for RNA secondary structures [J]. Journal of Shenzhen University: Science and Engineering Edition, 2009, 26(3): 272-277. (in Chinese)
[9] 胡桂武,彭宏. 基于免疫粒子群集成的 RNA 二级结构预测算法 [J]. 计算机工程与应用, 2007, 43(3): 26-29.
Hu Guiwu, Peng Hong. Algorithm based on immune PSO ensemble for predicting RNA secondary structure [J]. Computer Engineering and Applications, 2007, 43(3): 26-29. (in Chinese)
[10] 何静媛,邹东升,何中市. RNA 二级结构预测的自适应鱼群算法模型 [J]. 系统仿真学报, 2010, 22(6): 1 370-1 374.
He Jingyuan, Zou Dongshen, He Zhongshi. Self-adaptive artificial fish swarm algorithm for RNA secondary structure prediction [J]. Journal of System Simulation, 2010, 22(6): 1 370-1 374. (in Chinese)
[11] Jun Yu, Changhai Zhang, Yuanning Liu, et al. Simulating the folding pathway of RNA secondary structure using the modified ant colony algorithm [J]. Bionic Engineering, 2001(7): 382-389.
[12] Muzaffar M, Eusuff, Kevin E Lansey. Optimization of water distribution network design using the shuffled frog leaping algorithm [J]. Journal of Water Resources Planning and Management, 2003, 129(3): 210-225.
[13] 韩毅, 蔡建湖, 周根贵, 等. 随机蛙跳算法的研究进展 [J]. 计算机科学, 2010, 37(7): 16-19.
Han Yi, Cai Jianhu, Zhou Gengui, et al. Advances in shuffled frog leaping algorithm [J]. Computer Science, 2010, 37(7): 16-19. (in Chinese)
[14] 王翼飞,史定华. 生物信息学-智能优化算法及其应用 [M]. 北京: 化学工业出版社, 2006: 178-210.
Wang Yifei, Shi Dinghua. Bioinformatics-Intelligent Optimization Algorithms and Its Application [M]. Beijing: Chemical Industry Press, 2006: 178-210. (in Chinese)
[15] Muzaffar Eusuff, Kevin Lansey, Fayzul Pasha. Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization [J]. Engineering Optimization, 2006, 8(2): 129-154.
[16] Elbeltagi E, Hegazy T, Grierson D. Comparison among five volutionary-based optimization algorithms [J]. Advanced Engineering Informatics, 2005, 19(1): 43-53.
[17] Muker M, Mathews D H, Turner D H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide in RNA biochemistry and biotechnology [M] // Barciszewski J, Clark B F C, eds. NATO ASI Series. Dordrecht, NL: Kluwer Academic Publishers, 1999: 11-43.
[18] Jens Reeder, Robert Giegerich. Design, implementation and evaluation of a practical pseudoknots folding algorithm based on thermodynamics [J]. BMC Bioinformatics, 2004(5): 104.

[责任编辑: 严海琳]

(上接第 62 页)

[9] 李明仑,李凡长. DFL 的格结构及其应用研究 [J]. 计算机应用与软件, 2007, 24(9): 145-150.
Li Minglun, Li Fanzhang. Research on lattice structure of DFL and its application [J]. Computer Applications and Software, 2007, 24(9): 145-150. (in Chinese)
[10] 李明仑. 基于动态模糊格的决策树理论及应用研究 [D]. 苏州: 苏州大学计算机科学与技术学院, 2008.
Li Minglun. Research and application on theory of decision tree based on dynamic fuzzy lattice [D]. Suzhou: School of Computer Science and Technology, Soochow University, 2008. (in Chinese)
[11] Iun H Witten, Eibe Frank. Data Mining Paratical Machine Learning Tools and Techniques with Java Implementations [M]. Beijing: China Machine Press, 2003.
[12] Zoubin Ghahramani, Michael I Jordan. Supervised learning from incomplete data via an EM approach [J]. Advances in Neural Information Processing Systems, 1993(6): 120-127.
[13] 刘鹏,雷蕾,张雪凤. 缺失数据处理方法的比较研究 [J]. 计算机科学, 2004, 31(10): 155-156, 174.
Liu Peng, Lei Lei, Zhang Xuefeng. A comparison study of missing value processing methods [J]. Computer Science, 2004, 31(10): 155-156, 174. (in Chinese)

[责任编辑: 严海琳]