基于遗传退火算法的最大简约树构建的研究*

刘清雪,马志强,刘 磊
(吉林建筑工程学院 城建学院,吉林 长春 130111)

摘 要:针对最大简约法的搜索速度慢等特点,提出了一种遗传算法与模拟退火算法相结合的启发式搜索方法。利用模拟退火算法保障物种的多样性,克服了遗传算法的早熟现象,加快了实验后期的收敛速度。结果表明,该算法的准确性和运算效率都有较大提高。

关键词:种系发生树;最大简约法;遗传算法;模拟退火算法

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2010)15-0074-03

The study on maximum parsimony phylogenetic tree construction based on the genetic-annealing algorithm

LIU Qing-Xue, MA Zhi-Qiang, LIU Lei
(The City College of Jilin Architectural and Civil Engineering Institute, Changchun 130111, China)

Abstract: This paper proposed a new heuristic search method that the genetic algorithm and simulated annealing algorithm inspired by a combination of showing search. Use simulated annealing algorithm to protect the diversity of species, namely, overcome the premature convergence of genetic algorithms and speed up the convergence rate of the latter part of the experiment. The results show that the algorithm's accuracy and efficiency of operation has been greatly improved.

Key words: phylogenetic tree; maximum parsimony method; genetic algorithm; simulated annealing algorithm

系统发生(也称种系发生或系统发育,phylogenetic inference)是指生物形成或进化的历史[1],其基本思想是比较物种的特征,并认为特征相似的物种在遗传学上接近[2]。其研究的结果则以系统发生树(phylogenetic tree)表示,用它描述物种之间的进化关系。系统发生分析是根据某种标准,从给定的一组序列数据中推导出这些对象之间"最好"的系统发生树的过程。

简约法构建发生树,主要问题就是庞大的搜索空间。遗传算法是应用于搜寻各类问题最优解的一种方法,因此,基于遗传算法来寻找最大简约树是适合的。但该算法有两个严重的缺点,容易导致过早收敛、以及在进化后期搜索效率低[3]。

基于最优原则的最大简约法的启发式搜索,将模拟退火算法引入遗传算法群体更新的阶段,既保证群体多样性,又在后期逐步加快收敛速度,克服遗传算法早熟现象,最终目标是尽量使得最大简约树的树长最小、搜索时间最短。

1 最大简约法算法描述

最大简约法通过简约标准可以从现存后代的序列中客观地推测出祖先状态,不仅可以填补分子进化研究中的空白,更是对进化理论研究的重大贡献。对于系统发生树最直观的代价计算就是沿着各个分支累加特征变化的数目,而所谓简约就是使代价最小[4]。利用最大简约方法构建系统发生树,实际上是一个对给定分类单元所有可能的树进行比较的过程,针对某一个可能的树,首先对每个位点祖先序列的核苷酸组成做出推断,然后统计每个位点阐明差异的核苷酸最小替换数目。在整个树中,所有简约信息位点最小核苷酸替换数的总和称为树的长度或树的代价。通过比较所有可能的树,选择其中长度最小、代价最小的树作为最终的系统发生树,即最大简约树[5]。

2 遗传算法基本理论

遗传算法(genetic algorithm)由美国 HOLAND 教授于1975年首次提出,是一类通过模拟生物界自然选择和

*基金项目:国家自然科学基金项目(90304010);吉林建筑工程学院城建学院项目(院科字 2009111)

自然遗传机制的随机化搜索算法[3]。遗传算法首先对问题的解进行编码，然后从一组随机产生的初始解(称为群体)开始搜索过程。群体中的每个个体是问题的一个解，称为染色体。遗传算法主要通过交叉、变异、选择运算实现，染色体的好坏用适应度来衡量。根据适应度的大小从上一代和后代中选择一定数量的个体，作为下一代群体再继续进化，这样经过若干代之后，算法收敛于最好的染色体，它很可能就是问题的最优解或次优解。遗传算法中使用适应度的概念来度量群体中的每个个体在优化计算中达到最优解的优良程度[6]。

3 模拟退火算法基本理论

模拟退火算法来源于固体退火原理，将固体加温至充分高，再让其徐徐冷却，加温时，固体内部粒子随温度升高变为无序状，内能增大；而徐徐冷却时粒子渐趋有序，使每个温度都达到平衡态，最后在常温时达到基态，内能减为最小。根据 Metropolis 准则，将内能 E 模拟为目标函数值 f，温度 T 演化成控制参数 t，由初始解 i 和 $t0$ 开始，对当前解重复“产生新解→计算目标函数差→接受或舍弃”的迭代，并逐步衰减 t 值，算法终止时的当前解即为所得近似最优解。退火过程由冷却进度表控制，并具有以一定的概率接受恶化解的特点[7]。

4 基于遗传算法和模拟退火算法的最大简约法

虽然从发现了“早熟”现象，并对它所提出的改进策略有多种，但都是从遗传算子本身寻找改进方法，并没有根本解决“早熟”现象，它仍然是困扰遗传算法的一个问题，所以当把遗传算法应用在进化树构建中时，并没有达到令人完全满意的效果。遗传算法把握总体的能力较强，但局部搜索能力较差；而模拟退火算法具有较强的局部搜索能力，因此，为了克服遗传算法和模拟退火算法各自的缺点，发挥它们的优势，本文利用模拟退火算法对遗传算子进行改进，使遗传算法与模拟退火算法相结合，并应用在简约法构建进化树上。

(1)编码方式、适应度函数和种群的初始化

由于输入数据是核苷酸序列，由 A，C，G，T(U)所组合而成，因此直接使用这四个字母，将输入的每一个核苷酸序列看成一个编码，不需要进行额外操作。

使用简约法意义下的树长作为适应度函数，其值为进化树的适应值，为了加速算法的收敛，定义历史最大简约树为整个搜索过程中出现的具有历史最低适应值的树。

根据输入的物种序列，随机产生初始群体。

(2)选择退火算子

将生成的初始群体中的 PopSize 个个体进行适应度评价，个体适应度值越大，该个体被遗传到下一代的概率也越大。采用随机联赛选择方法，联赛规模 N 取值为 2。

①随机选择初始群体两个个体 P_i、P_j，计算其个体适应度值 $f(P_i)$和 $f(P_j)$。

②如果 $f(P_i)<f(P_j)$，则选择个体 P_i 遗传到下一代；否则，以概率 P 接受个体 P_i；

③重复①、②操作直到新的一代群体中也包含 PopSize 个个体。

(3)复制操作

对种群里的 PopSize 棵进化树依照适应值得分进行排序。由于适应值就是最大简约法的树长，得分越低，进化树差异越小，所以得分越低的进化树排序越靠前。将种群里的进化树排序完成后，复制过程也就结束了。

(4)交叉退火算子

挑选经复制过程后的第一棵进化树，也就是适应值最小的进化树，将这棵经由最优适应值进化树所产生的树标记为 α；再由复制过程产生的 PopSize 棵进化树中，随机挑选出一棵，复制此进化树，将其标记为 β；接着 β 再去与 α 进行交配：由 β 进化树随机选择分支，将此分支标记为 δ，接着将 α 树中移除 δ 分支所包含的所有物种，同时删除 α 树中多余的分支。将 δ 分支插入到 α 树中得分最高的位置。插入之后，就得到了一棵新的进化树。完成这样的一串动作之后，也就是完成了一次交配：α 树与 β 树经由交配结合而形成了一棵新的进化树。重复以上操作以概率 P_c 完成对父代的交叉操作。图 1~图5 所示为一个交叉过程。

分别计算父代和子代的适应度值，进行前文所述的退火操作。具体操作过程如下：

①选取适应值最小的树 α 及任意的树 β，并由 β 随机选择分支。

②将父代个体 α、β 进行交叉，生成子代个体 α，计算个体适应度 $f(\alpha)$，$f(\alpha')$。

③进行退火操作，如果 $f(\alpha)<f(\alpha)$，则用 α′代替 α，否则，以概率 P 接受 α′。

④循环步骤②、③，直到以概率 P_c 完成所有父代个体的交叉操作。

(5)变异退火算子

由复制过程产生的 PopSize 棵进化树中，随机挑选一棵。对挑出的这棵进化树随机选取两个不同的内部节

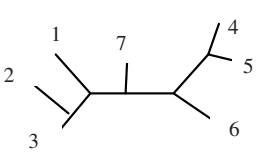

图 1 α 树

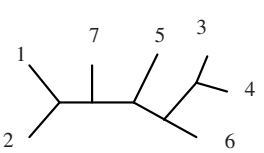

图 2 β 树

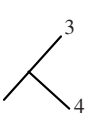

图 3 δ分枝

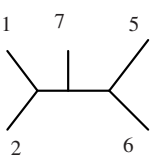

图 4 除去 δ 后的 α 树

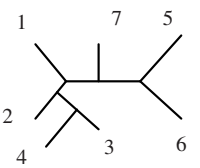

图 5 α 树与 β 树交叉之后的新树

点作为交换点，并交换这两个交换点，同时移动交换点以下的所有节点和分支。这样就完成了一次突变过程。图 6 所示为选择两个交换点，图 7 是两个点交换之后形成的新树。

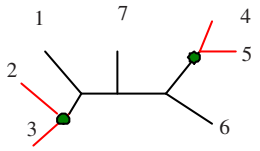

图 6 选择两个交换点

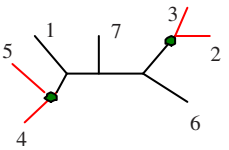

图 7 交换之后的新树

具体操作过程如下：

①随机选择个体 P_i 的两个内部节点做变异生成新个体 P_i'；

②计算个体适应度值 $f(P_i)$ 和 $f(P_i')$。如果 $f(P_i')<f(P_i)$，则用 P_i' 代替 P_i；否则，以概率 P 接受 P_i'；

③重复步骤①、②。

(6)更新初始群体

将复制过程、交配过程和突变过程产生的树作为新的族群，作为下一次迭代的初始群体。若当前群体中最佳个体的适应度比总的迄今为止的最好个体的适应度还要高，以当前群体中的最佳个体作为新的迄今为止的最好个体，同时用该个体替换掉当前群体中的最差个体。

(7)结束条件

①群体进化的代数超过最大代数值时；

②进化代数超过一定值而适应度值不再提高时，这个值为适应代数。

最后一代中适应度值最高的个体即为最优解。

5 实验结果

对于基于简约法的建树方法，可以从运行时间和树长两个指标来衡量。对遗传退火简约法，以 TreeBASE (http://treebase.org/treebase/) 所提供的序列资料作为测试数据的来源进行了数据实验和模拟实验。并选择了与 PHYLIP 软件做对比，本算法主要涉及的参数为：群体规模 PopSize=100，最大进化代数 MaxGeneration=200，交叉概率 $P_c=0.6$，变异概率 $P_m=0.1$，生成的初始群体用参数 P_0 表示，选择算子联赛规模 $N=2$。接受恶化解的概率公式 $P=\exp\left(\frac{f(P_j)-f(P_i)}{T}\right)$，实验结果如表 1 所示。

表 1 实验结果

名称	序列属性		最大简约法		遗传退火简约法	
	物种数	序列长度	运行时间/s	树长	运行时间/s	树长
Rbcl	55	1 315	271	5 433	252	5 221
CatLemurs	35	604	156	1 128	130	1 046
HIVenSweden	13	273	42	598	35	579
RNasep	13	492	59	674	54	657
lysozymeSmall	7	390	20	402	16	381

从表 1 可以看出，设计的遗传退火简约算法要比最大简约法具有更高的运算效率和准确性。

构建发生树的研究是生物信息学中的一个热点，已建立和发展了许多新的技术和方法，但由于问题的复杂性，目前还没有一种最优算法能在适当的时间内计算得到其精确解。本文中的改进算法相比原有算法性能上有了提高，但是仍有不足的地方，需要进一步地完善。

参考文献

[1] 吕宝忠，钟扬．分子进化与系统发育[M]．北京：高等教育出版社，2002．

[2] 张阳德．生物信息学[M]．北京：科学出版社，2004．

[3] 李敏强．遗传算法的基本理论与应用[M]．北京：科学出版社，2002．

[4] 钟扬，王莉，张亮．生物信息学[M]．北京：高等教育出版社，2003．

[5] 孙啸，陆祖宏，谢建明．生物信息学基础[M]．北京：清华大学出版社，2005．

[6] TIENG K, OPHIR F. Parallel computation in biological sequence analysis [J]. IEEE Transaction on parallel and Distributed Systems, 1998,9(3):21-25.

[7] 丁永生．计算智能—理论、技术与应用[M]．北京：科学出版社，2004．

(收稿日期：2010-02-09)

作者简介：

刘清雪，女，1977 年生，硕士，讲师，主要研究方向：生物信息学序列比对及系统发育分析。

马志强，男，1963 年生，博士，教授，主要研究方向：生物信息学、计算智能、嵌入式系统。

刘磊，女，1975 年生，硕士，讲师，主要研究方向：网格计算。