文章编号: 1007- 2985(2005) 04- 0037- 04

基于模拟退火遗传算法的多序列比对方法

胡桂武[1], 曾 岫[2], 黄 辉[1]
(1. 广东商学院经济数学系, 广东 广州 510320; 2. 广州航海高等专科学校信息工程系, 广东 广州 510725)

摘 要: 针对 MSA 问题提出了将遗传算法与模拟退火算法结合在一起的混合算法. 该算法充分发挥了遗传算法和模拟退火算法的优越性, 可提高求解多序列比对 MSA 问题的计算精度和计算速度, 整个算法模拟了自然界进化的周期性, 较好的解决了群体的多样性和收敛深度的矛盾. 实验表明, 该方法算法是有效的.

关键词: MSA; 生物信息学; 遗传算法; 算子
中图分类号: TP301. 6; TP393 文献标识码: A

生物序列比对是计算生物学中最重要、也是最有挑战性的任务之一. 其中多序列比对问题(multiple sequence alignment 简称 MSA) 的求解已经证明多序列比对问题是一个 NP 完全问题[3], 想要找到复杂性为多项式的精确解算法是不可能的. 因此, 近年来研究者们致力于研究它的近似解, 但每一种方法都有其不足之处, 目前对多序列的研究还在不断的探索中. 笔者对该问题进行了适当的编码, 把该问题转换成搜索空间中的一个优化问题, 并提出了一种模拟退火算法和遗传算法的混合算法, 吸取 2 个算法的优点, 成为一种针对 MSA 问题的高效混合算法, 最后在 MSA 问题的求解上的得到了具体的测试. 结果表明了算法的有效性和可行性.

1 多序列比对(MSA) 描述

对于给定的 N 条序列, $S_1, S_2, \quad , S_N, S_k = S_{k1}S_{k2} \quad S_{kn_k}, k = 1, 2, \quad , N$. 这里 $S_{ij}, j = 1, 2, \quad , n_i, i = 1, 2, \quad , N$ 表示一个核苷酸或氨基酸残基. 有效的残基类型集合用 表示. 对 DNA 序列, = {A, G, C, T}; 对 *RNA* 序列, = {A, G, C, U}; 对蛋白质序列, 包含了 20 个字符, 每个字符代表一种氨基酸. 另外, 在比较几个相关序列时会出现中断(break) 现象, 这就产生了 间隙(gap) 问题, 间隙用 - 表示. 用 表示 {-}. 多序列比对定义[1] 如下:

定义 1 一个多重序列比对 $\boldsymbol{A}$ 是一个二维字符矩阵, 即 $\boldsymbol{A} = \{\boldsymbol{S}_{ij}\}, j = 1, 2, \quad L, i = 1, 2, \quad , N$. 其中: $S_{ij} = S_{ij}$ 或 - , max(n_i) L $\sum_{i=1}^{N} n_i$, 并满足( ) 序列的数目等于矩阵的行数; ( ) 如果移去每行中的 - 字符, 将得到原来的序列; ( ) 每一列中不允许同时为 - .

N 条序列比对的分值: 为 N 条序列中任意 2 条序列(共有 C_N^2 种可能) 的分值 V 之和, 用 SP 来表示: SP

收稿日期: 2004- 09- 07
作者简介: 胡桂武(1970-), 男, 湖南冷水江人, 广东商学院经济数学系讲师, 博士研究生, 主要从事数据挖掘、人工智能、生物信息学研究.

$= \sum_{i=1}^{C_N^2} V_i$. 已知 N 个序列的序列组 $S = \{S_1, S_2, \ , S_N\}$, 对于一个多序列的比对 , 可以定义不同形式的目标函数, 一种常用的目标函数是基于 SP 准则[1] 给出的, 其定义:

$$f(\ ) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{l=1}^{L} (s_{il}^{*}, s_{jl}^{*}), \tag{1}$$

其中 $(*, *)$: **R** 为二元实值函数. 称(1) 式的值为比对 的得分. 如果 S 的一个比对 满足条件

$$f(\ ) = \min f(\ ), \tag{2}$$

则称 是一个最佳比对. 求最佳比对的问题是一个 NP 完全问题.

2 算法的设计与分析

2.1 模拟退火算法

模拟退火算法[4] 起源于统计物理学中对固体退火过程的模拟. 它采用 Boltzmann 接受准则接收新解, 用一个冷却系统的参数控制算法进程, 是算法在多项式时间里给出一个近似最优解. 算法描述如下:

SETP 1: 初始退火温度 $T_k(k = 0)$, 产生随机初始解 X_0;

SETP 2: 对在 T_k 下重复执行如下操作, 直至达到温度 T_k 的平衡状态: (1) 在解 x 的领域中产生新的可行解 x^*, 计算新旧解的评价函数值的差 $f = f(x^*) - f(x)$; (2) 当 $\min\{1, \exp(- f / T_k)\}$ $random[0, 1]$ 接收新解, 其中 $random[0, 1]$ 是[0, 1] 区间内的随机数.

SETP 3: 令 $T_{k+1} = T_k, k$ $k + 1$, 其中 (0, 1). 若满足收敛条件, 则结束, 否则转 SETP 2.

2.2 混合算法

算法框架的设计是模拟自然界演化的周期性的特点, 算法对一个进化周期的设计是: 首先使用选择算子、自适应变异算子对群体进行进化, 然后用模拟退火算法对群体进行作用, 当群体经过一定的进化代数后, 如果陷入局部最优, 则使用突变算子增强算法的搜索能力. 经过一定的代数进化后, 仅仅保留最优解, 对最优个体所对应的序列组进行星比对, 比对后的序列组对应的染色体个体如果更优则取代最优解, 重新生成其余个体, 进入下一个周期, 宛如自然界种族灭绝后的再生. 这种策略并非退化, 而是尽快摆脱进化迟钝状态, 开始一个新的进化周期. 算法就是通过若干个这样的进化周期, 最后找到最优解的. 具体算法设计如下:

```
Procedure SA- GA Algorithm
begin
随机初始化群体 P; 计算 P 中个体的适应值;
    optimal _ indivi    P 中最优的个体;
    gen    0;
    while gen< LS do        LS 为进化周期数
      begin          一个进化周期开始
        k    0;
    while k< EG do        EG 是一个进化周期所含的进化代数
    begin
    对初始群体使用主搜索组合算子;
    if 陷入局部最优, 个体按适应值从小到大排序, 序号为偶数的个体使用突变算子.
    k    k+ 1;
    end;          保留最优个体
    if P 中最优个体好于 optimal _ indivi then optimal _ indivi    P 中的最优个体;          在进入下一个进化周期前进行重组
```

```
    S    {随机生成 N- 1 个体};        N 为种群规模
    P    S+ {个体 optimal _ indivi};
    gen  gen+ 1;
    end;
end;
```

对上述混合算法, 说明如下:

(1) 编码方式与适应度函数. 为了操作和评价方便, 算法针对的 MSA 问题采取十进制编码方式, 适应度函数由(1)式得.

对于已知 N 条序列的序列组 $S = \{S_1, S_2, \cdots, S_N\}$, 每一次比对用一条染色体表示, 染色体分成 N 段 $G_1, G_2, \cdots, G_N$, 第 G_k 段与序列 S_k 相对应. 其中 G_k 的长度表示在 S_k 中插入的空格数, G_k 中的每一个基因的值表示在 S_k 中插入空格的位置. 例如 $S = \{cfdaecd, feadace\}$, 染色体:

3	6	6	0	0	6	0	7	0	0	0	4

染色体分为2 段, 每段的长度为 6 则相应的比对为:

$$--cfd-aec---d$$
$$----fead-ace-$$

(2) 相关算子. 定义 2　变异算子定义为如下操作(假设父体是 S): 从 S 中随机取 2 个基因 g_1 和 g_2. 将 g_1 和 g_2 间的基因(含 g_1 和 g_2) 反序, 同时修改 S 的适应值.

定义3　突变算子定义为如下(假设父体是 S): 确定一个自然数 K, K 不大于染色体的长度, 把染色体分成 K 段. 从染色体 S 中每一段染色体片段 S_i 中随机取 1 个基因 c_i; 随机给每个 c_i 赋可能的值, 同时修改 S 的适应值.

定义 4　主搜索组合算子: (1) 用轮盘赌选择法选择再生个体, 按一定的交叉概率(P_c) 使用交叉算子, 生成新的个体, 交叉算子使用单点交叉;

(2) 按一定的变异概率(P_m) 使用变异算子, 生成新的个体, 变异算子即定义 2. 其中:

$$P_m = \begin{cases} P_m/(1-F) & F \quad a, \\ P_m & \text{other.} \end{cases}$$

其中: $F = fit_{over}/fit_{max}$, 0　P_m　0.5, 0.5　a　1.

(3) 对变异后个体使用模拟退火算法, 产生新一代的种群.

3　实验与结论

3.1 实验

为了检验新的算法的有效性, 笔者已将给出的混合算法用 c++ 编程实现, 并将其中用于多序列(DNA, RNA 和氨基酸序列) 比对问题, 测试结果显示该方法性能良好, 下面仅仅以 4 组来自数据库 NCBI 的 RNA 序列为实验例子与著名的多序列比对程序 CLUSTAL 进行比较, 实验表明混合算法所得到的结果优于 CLUSTAL 的结果, 同时说明了该算法在求解 MSA 问题上的可行性和优越性.

Test case	Nseq	Length	Clustal score	AN- GAscore
RNA(1)	6	185	1 287	1 283
RNA(2)	7	653	5 057	4 996
RNA(3)	17	672	4 772	4 662
RNA(4)	23	352	7 761	6 545

注　Test case　比对序列, Nseq　序列数, Length　序列平均长度, Score　比对后的罚分.

3.2 结论

从理论分析和实例可见,笔者提出的混合算法在 MSA 问题求解上得到较满意的结果,充分发挥了二者的优势.为了适应生物信息的海量性特征以及生物序列爆炸性增长的趋势,研究和设计速度更快、精度越高的算法极为重要,这也是笔者下一步的工作.

参考文献:

[1] 塞图宝,梅丹尼斯,朱 浩,等.计算分子生物学导论[M].北京:科学出版社,2003.

[2] WATERMAN M S. General Methods of Sequence Comparison [J]. Bull. Math. Biol., 1984, 46: 473- 500.

[3] WANG L, JIANG T. On the Complexity of Multiple Sequence Alignment [J]. J. Comput. Biol., 1994, (1): 337- 348.

[4] 康立山.非数值并行算法(第一册) 模拟退火算法[M].北京:科学出版社,1997.

[5] 刘 勇,康立山,陈毓屏.非数值并行算法(第二册) 遗传算法[M].北京:科学出版社,1995.

[6] LIPMAN D J, ALTSCHUL S F, KECECIOGLU J D. A Tool for Multiple Sequence Alignment [J]. Proc natn Acad Sci, 1989, 86: 4 412 - 4 415.

[7] JIANG Tao, KEARNEY P, LI Ming. Some Open Problems in Computational Molecular Biology [J]. J of Algorithms, 2000, 34: 194- 201.

An Algorithm Based on the Simulated Annealing Genetic Algorithm for Multiple Sequence Alignment

HU Gui-wu[1], ZENG Xiu[2], HUANG Hui[1]

(1. Department of Mathematics, Guangdong Commercial College, Guangzhou 510320, China;
2. Information Technology Department of Guangzhou Maritime College, Guangzhou 510725, China)

Abstract: The mixed algorithms of genetic algorithms and simlutated annealing algorithm are put forward. The new algorithm not only sufficiently exerts the advantages of the two algorithms, but also improves the computing precision and speed. The algorithm simulates the recurrence of nature evolution process, and solves the contradiction between the diversity of population and the convergence speed. The experiment shows that the algorithm is effective.

Key words: multiple sequence alignment; bioinformatics; genetic algorithm; operator