PHGA-COFFEE：多序列比对问题的并行混合遗传算法求解

刘立芳 霍红卫 王宝树

（西安电子科技大学计算机学院 西安 710071）

摘 要 设计了一个求解多序列比对问题的并行混合遗传算法（与之相应的软件称为 PHGA-COFFEE）. 该算法采用 COFFEE 函数作为个体的适应度函数，构造了六种遗传算子，特别是设计了两种新颖的变异算子，其中一种变异算子基于 COFFEE 的一致性信息设计，以改善算法的整体搜索能力. 另一种变异算子基于动态规划方法设计，以增强其局部搜索能力. 通过对 BAliBASE 中 144 个测试例的测试，证明该算法是有效的. 与已有的算法相比，该算法对处于朦胧区和具有 N/C 末端延伸的序列比对问题有更强的问题求解能力. 同时通过对算法并行化，其运行时间显著缩短.

关键词 生物信息学；多序列比对；并行混合遗传算法；动态规划

中图法分类号 TP18

PHGA-COFFEE: Aligning Multiple Sequences by Parallel Hybrid Genetic Algorithm

LIU Li-Fang HUO Hong-Wei WANG Bao-Shu

(*School of Computer Science and Technology, Xidian University, Xi'an* 710071)

Abstract A parallel hybrid genetic algorithm and an associated software package called PHGA-COFFEE are presented. The COFFEE function is used to measure individual fitness, and six genetic operators are designed, especially two novel mutation operators are proposed, one is designed based on the COFFEE's consistency information that can improve the global search ability, and another is realized by dynamic programming method that can improve individuals locally. Experimental results of the 144 benchmarks from the BAliBASE show that the proposed algorithm is feasible, and for datasets in twilight zone and comprising N/C terminal extensions, PHGA-COFFEE generates better alignment as compared to other methods. At the same time, the computation time of PHGA-COFFEE is remarkably reduced due to the parallel algorithm.

Keywords bioinformatics; multiple sequence alignment; parallel hybrid genetic algorithm; dynamic programming

1 引 言

多序列比对是进行生物序列分析的最基本的任务之一，它在发现序列模体（motif）和保守区域、系统发育分析、结构预测等方面具有重要的作用，是生物信息学当前研究的热点问题之一[1]. 已有的多序列比对算法大体分三类：精确比对算法、渐进比对算

收稿日期：2004-11-15；修改稿收到日期：2006-02-15. 本课题得到陕西省自然科学基金资助. 刘立芳，女，1972 年生，博士研究生，主要研究方向为生物信息学、并行计算. E-mail：lfliu@xidian.edu.cn. 霍红卫，女，1963 年生，博士，教授，主要研究领域为分布与并行计算、算法设计、生物信息学. 王宝树，男，1942 年生，教授，博士生导师，主要研究领域为智能信息处理、数据融合等.

法、迭代比对算法. 精确比对算法最为经典的是多维 Needlman-Wunsch[2] 算法,但其可行的计算维数为 3, Carrillo-Lipman[3] 算法通过减小计算空间,将计算维数提高到 10. 渐进比对算法由 Hogeweg[4] 首先提出,Feng[5] 和 Taylor[6] 又加以完善. 非常著名的、被广泛使用的多序列比对软件包 CLUSTAL W[7](其含窗口界面的版本为 CLUSTAL X)基于渐进比对思想构建. 近年来,迭代比对算法被越来越多地用于求解多序列比对问题,基于模拟退火、遗传算法、HMMs、Gibbs 抽样等的多序列比对算法被广泛应用于多序列比对问题的求解,其中多序列比对软件包 SAGA[8] 基于遗传算法构建,共设计了 22 种不同的遗传算子,采用动态调度的策略控制 22 种遗传算子的使用. PHGA[9] 采用了并行的混合遗传算法来求解问题,将问题表示成求解 k 维带权有向图的最短路问题,而设计了与之相应的一维染色体编码方式. 文献[10]对多种多序列比对方法和与之相应的软件的性能进行了详细的比较.

本文针对多序列比对问题的 COFFEE[11] 优化模型,提出了一个新的基于并行混合遗传算法的求解算法(与之相应的软件称为 PHGA-COFFEE),共设计了 6 种遗传算子. 通过对 BAliBASE[10] 中的 144 个测试例的测试,其结果表明该算法是有效的,比对结果的准确性好于 SAGA,与 CLUSTAL W 和 PHGA 的相当,特别对处于朦胧区和具有 N/C 末端延伸的序列比对问题有更强的问题求解能力. 同时,并行处理算法的运行时间显著缩短.

2 问题描述

2.1 多序列比对问题

一条长度为 l 的生物序列是由 l 个字符组成的字符串,字符串中的字符取自于一个有限字母表 Σ,对于 DNA 序列,Σ 包含 A, C, G, T 4 个字母,分别代表 4 种不同的核苷酸,对于蛋白质序列,Σ 包含 20 个不同的字母,分别代表 20 种不同的氨基酸,将这些字母统称为残基. 给定 n 条序列组成的序列组 $S=(s_1, s_2, \cdots, s_n)$,其中 $s_i = s_{i1}\, s_{i2} \cdots s_{il_i}\ (1 \le i \le n)$,$s_{ij} \in \Sigma\ (1 \le j \le l_i)$,$l_i$ 为第 i 条序列的长度,则关于 S 的一个多序列比对可定义为一个矩阵 $\boldsymbol{A} = (a_{ij})$,其中 $1 \le i \le n$, $1 \le j \le l$, $\max(l_i) \le l \le \sum_{i=1}^{n} l_i$,并且该矩阵有如下特性:(1) $a_{ij} \in \Sigma \cup \{\cdot\}$,其中"·"代表空位;(2)如果删除空位"·",则 $\boldsymbol{A}$ 的每一行 $a_i = a_{i1}\, a_{i2} \cdots a_{il}\ (1 \le i \le n)$ 与对应序列 s_i 相同;(3) $\boldsymbol{A}$ 中不存在只由空位"·"组成的列. 一个 4 条蛋白质序列的多序列比对如图 1 所示.

```
kkdsnapkramtsfmffss.....dfrskhsdlsi.vemskaagaawkelgpeerkvyeemaekdkerykrem.........
.....kpkrprsayniyvscsfqcakddsaqgkl.....klvncawknlspcckqayiqlakddrirydncmkswccqmac
...adkpkrplsaymlwlnsaresikrenpdfkv.tevakkggelwrgl..kdkseweakaatakqnyiralqeyerngg.
..dpnkpkrapsaffvfmgcfrccfkqknpknksvaavgkaagerwkslscscskapyvakanklkgcynkaiaaynkgcsa
```

图 1 一个多序列比对例子

2.2 多序列比对问题优化模型

对于多序列比对问题,可以定义不同形式的目标函数,常用的有 SP 记分函数[2,3,7,9](weighted Sums-of-Pairs with affine gap penalties)、隐 Markov 模型(HMMs)[12]、COFFEE 记分函数[11](Consistency based Objective Function For alignmEnt Evaluation)等. SP 记分函数依赖于替代矩阵的选取、空位罚分和空位延伸罚分的设置、参数的选取并且设置复杂. HMMs 为一概率模型,需要大量的训练序列来保证其准确性. COFFEE 记分函数根据当前多序列比对 $\boldsymbol{A}$ 和双序列比对库之间的一致性信息来计算分值,可以利用已有的双序列比对程序生成双序列比对库,一旦建立了 n 条序列的双序列比对库,就不需要替代矩阵和空位罚分的选取. 本文所用的目标函数为 COFFEE 记分函数. 基于 COFFEE 记分函数的比对方法首先需要生成双序列比对库,即 n 条序列的两两最优比对,共有 $n(n-1)/2$ 个双序列比对. 基于此,一个多序列比对 $\boldsymbol{A}$ 的 COFFEE 记分函数定义如下[11]:

$$COST(\boldsymbol{A}) = \Big[\sum_{i=2}^{n}\sum_{j=1}^{i-1} w_{ij} SCORE(A_{ij})\Big] \Big/ \Big[\sum_{i=2}^{n}\sum_{j=1}^{i-1} w_{ij} LEN(A_{ij})\Big] \quad (1)$$

其中,A_{ij} 为序列 s_i 和 s_j 在 $\boldsymbol{A}$ 中的双序列比对,$LEN(A_{ij})$ 是其比对长度,$SCORE(A_{ij})$ 是 A_{ij} 与库中 s_i 和 s_j 最优比对的一致性,其值等于在 A_{ij} 中和库中比对残基对的一致性数目,w_{ij} 为序列 s_i 和 s_j 的相同度. 如果 S 的一个比对 $\boldsymbol{A}'$ 满足 $COST(\boldsymbol{A}') = \max_{\boldsymbol{A}}(COST(\boldsymbol{A}))$,则

称 A' 是一个最优比对,此问题是一个 NP 完全问题[13].鉴于此,近年来人们致力于研究多序列比对问题的近似算法,已取得了不少有意义的成果[14].本文针对上述优化模型,提出了一个基于并行混合遗传算法的多序列比对问题求解算法,与 SAGA 不同的是,本文中的算法共设计了 6 种不同的遗传算子,算法中未使用 Guide Tree 信息指导变异算子的操作,而是充分利用了双序列比对库所带的一致性信息来指导变异算子的操作.

3　算法描述

3.1　种群初始化

矩阵 A 描述了一个多序列比对,对问题直接求解,染色体采用二维编码方式.初始种群的设定不是完全随机的,而是利用了双序列比对库的信息,将 n 条序列按其序号($1\sim n$)进行任意排列,则前 $\lfloor(n-1)/2\rfloor\times 2$ 条序列两两结合,将双序列比对库中双序列比对结果插入到序列所在位置,剩余的序列($\lfloor(n-1)/2\rfloor\times 2+1\sim n$)在其长度范围内任取一位置,插入一定数量的空位,个体的长度为变化后的 n 条序列中的长度的最大值,不足者补空位,并且保证个体的互异性.种群初始化时,每个染色体的最大长度设为 $w=\lceil 1.2\times l_{\max}\rceil$,$l_{\max}=\max(l_1,l_2,\cdots,l_n)$,迭代过程开始后,个体的最大长度为 $w=\lceil 2\times l_{\max}\rceil$,长度超过此值的个体,则舍弃.选择 1.2 作为乘数因子是基于观察一般多序列比对结果,其空位数很少超过 20%.

3.2　交叉算子

算法中设计了两种交叉算子,一点交叉和两点交叉,分别如图 2 和图 3 所示.对产生的两个新个体,只保留适值高的且与已生成的个体相异的一个个体.这样,在一定程度上保持了群体的多样性,但算法的收敛速度减慢.

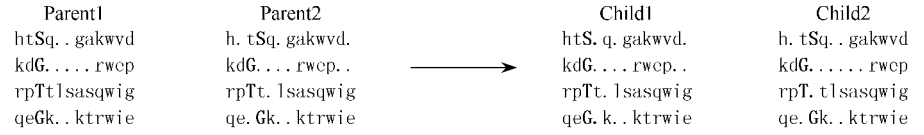


图 2　一点交叉示意

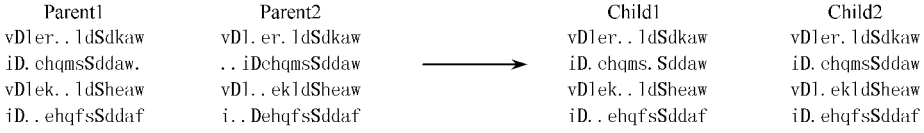


图 3　两点交叉示意

3.3　变异算子

算法中设计了两种变异算子,下面分别描述:

(1)ConsistencyShuffle.基于 COFFEE 函数一致性信息的目的,在 n 条序列中任选一条序列,在此序列中任选某一个残基,将双序列比对库中与该残基对齐的其它残基移到相应位置上,如图 4 所示.

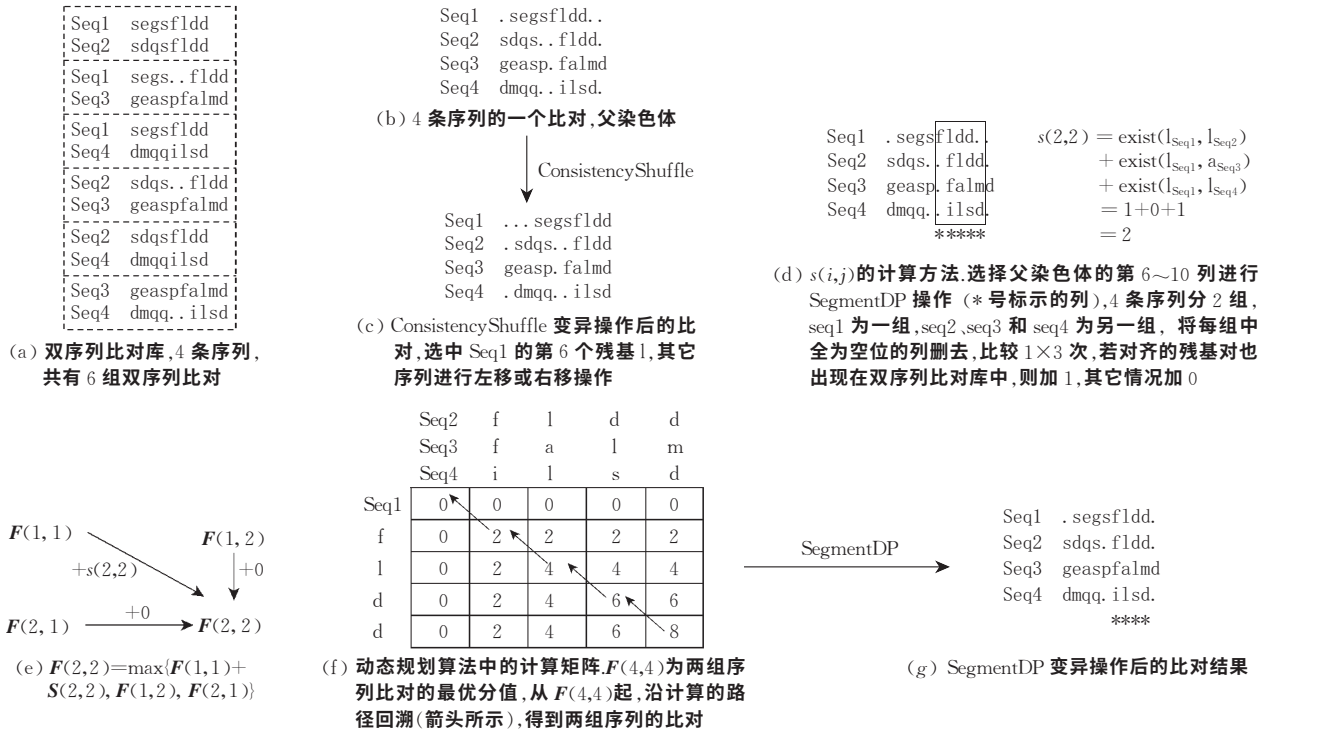
(2)SegmentDP.在父染色体上随机截取一段,其长度为 $2\leq l\leq 60$,将 n 条序列随机分为两组,其中一组只包含 $1\sim 2$ 条序列,这样当序列数多时,可减少计算量,将每组中全为空位的列删去,对这两组短片段运用动态规划(dynamic programming)的方法重新进行比对,然后连接其前段和后段,形成一个新的个体,如果新生成的个体不同于已有的个体,则将其保留,否则舍弃.下面给出计算方法.

设 F 为一个二维的运算矩阵,$F(i,j)$ 为两组片段分别从 $1\sim i$ 列和从 $1\sim j$ 列的最优比对分值,按式(2)进行计算:

$$F(i,j)=\max\{F(i-1,j-1)+s(i,j),F(i-1,j),F(i,j-1)\}\quad(2)$$

其中 $s(i,j)$ 为第 1 组的第 i 列和第 2 组的第 j 列的比对分值,当第 1 组第 i 列的一个残基和第 2 组第 j 列的一个残基相对齐,并且该残基对也出现在双序列比对库中时,加 1 分,其它情况加 0 分,若第 1 组有 m 条序列,第 2 组有 n 条序列,则要进行 $m\times n$ 次比较.具体计算过程如图 4 所示.

SegmentDP 作用于长度 $2\leq l\leq 60$ 的一个短片段,其计算时间控制在一定范围内,而与染色体的长度无关.



(a) 双序列比对库，4 条序列，共有 6 组双序列比对

(b) 4 条序列的一个比对，父染色体

(c) ConsistencyShuffle 变异操作后的比对，选中 Seq1 的第 6 个残基 l，其它序列进行左移或右移操作

(d) $s(i,j)$的计算方法.选择父染色体的第 6～10 列进行 SegmentDP 操作（* 号标示的列），4 条序列分 2 组，seq1 为一组，seq2、seq3 和 seq4 为另一组，将每组中全为空位的列删去，比较 1×3 次，若对齐的残基对也出现在双序列比对库中，则加 1，其它情况加 0

(e) $\boldsymbol{F}(2,2)=\max\{\boldsymbol{F}(1,1)+\boldsymbol{S}(2,2), \boldsymbol{F}(1,2), \boldsymbol{F}(2,1)\}$

(f) 动态规划算法中的计算矩阵.$\boldsymbol{F}(4,4)$为两组序列比对的最优分值，从 $\boldsymbol{F}(4,4)$起，沿计算的路径回溯（箭头所示），得到两组序列的比对

(g) SegmentDP 变异操作后的比对结果

图 4 ConsistencyShuffle 和 SegmentDP 变异操作

3.4 选择算子

算法采用了两种选择方式：最佳个体保存法、赌轮选择法，将上一代个体的适值按大小排序，将前 10％的个体保留到下一代，其余 90％的个体采用赌轮选择法选取父体，经由交叉或变异操作产生.

3.5 迁移算子

并行遗传算法的实现采用了粗粒度模型. 子种群之间的个体迁移结构为环状拓扑，迁移策略为一传一，即每个处理器按赌轮法选择出个体作为迁移对象，将选出的个体传送给其相邻处理器，并且接收相邻处理器传送的个体.

3.6 双序列比对库的生成

可以使用已有的序列比对程序来生成双序列比对库，每次将两条序列进行比对，n 条序列，则需进行 $n(n-1)/2$ 次比对. 在 PHGA-COFFEE 中，是采用 Needlman-Wunsch 算法[2] 进行双序列比对，并借鉴 CLUSTAL W[7] 的替代矩阵的选取和罚分方法. 对于蛋白质序列的比对，是选用了 BLOSUM 系列矩阵[15] 作为替代矩阵，并根据序列之间的进化距离来选择，如：80～100％选用 BLOSUM80，60～79％选用 BLOSUM62，30～59％选用 BLOSUM45，0～29％选用 BLOSUM30.

3.7 算法描述

算法中对遗传算子的调度采用如下的方法：选择两个个体，以概率 P_c 进行交叉运算，如果进行交叉运算，则又以概率 P_{ct} 进行两点交叉；否则进行一点交叉，交叉运算产生的两个新个体，保留适应度高的一个个体，对新个体以概率 P_{mcs} 进行 ConsistencyShuffle 变异运算；交叉运算之后，重新选择一个个体，以概率 P_{mdp} 进行 SegmentDP 变异运算；按给定的参数 *Send_rate* 和 *Send_best* 进行子种群间的迁移运算，*Send_rate* 为向相邻处理器迁移个体的频率，*Send_best* 为每个迁移步发送个体的数量. 迭代终止条件为达到规定最大代数 g_{max} 或在规定的代数 $g_{unimproved}$ 内，最大适值无改变. 算法描述如下.

```
Procedure PHGA-COFFEE
Begin
  1. Build pairwise library.
  2. Initialize population.
  While (Generation<g_max and
         g_current-unimproved<g_unimproved) do
  Begin
  3. Keep 10% parents to next generation.
     While (children number≠population size) do
     Begin
  4.   Tournament selection (parent1,parent2).
  5.   Crossover.
       If (random()<P_c)
         If (random()<P_ct)
           Two point crossover.
         Else
           One point crossover.
```

```
6.  ConsistencyShuffle mutation (child).
    If (random()<P_mcs)
      ConsistencyShuffle.
7.  Tournament selection (parent1).
8.  SegmentDP mutation.
    If (random()<P_mdp)
      segmentDP.
    End
9.  Migration.
    If (Generation MOD Send_rate==0)
      Send emigrants and receive immigrants according to the number of Send_best.
  End
End
```

4　测试结果

PHGA-COFFEE 是基于联网微机构成的 Windows 2000 机群、MPI(Message Passing Interface)平台实现. 基准多序列比对库 BAliBASE 1.0[10] 包含 144 个蛋白质多序列比对测试例,通过对测试例的比对,可以检验多序列比对程序的性能. 144 个测试例分为 5 类:(1)等进化距离近似长度的序列,82 例. 这 82 个测试例按照其残基相同度又划分成 3 类(V1(相似度<25%),V2(相似度在 20～40%之间),V3(相似度>35%));(2)一个家族相关序列和 1～3 条孤儿(orphans)序列,23 例;(3)多个家族的等进化距离的序列,12 例;(4)具 N/C 末端延伸的序列,15 例;(5)具内部插入的序列,12 例. 采用 SPS(Sum-of-Pairs Score)和 CS(Column Score)分值来衡量多序列比对程序的性能[10],SPS 分值表示残基对准确对齐的比率,CS 分值表示所有序列准确对齐的比率(即对齐多少列). 使用 BaliScore[10] 程序来计算 SPS 和 CS 分值. BAliBASE 和 BaliScore 可从 ftp://ftp-igbmc.u-strasbg.fr/pub/BAliBASE 下载. 我们从 ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/ 下载了 CLUSTAL X1.83,并对 144 个测试例进行了比对,得到其 SPS 和 CS 分值. SAGA 0.95 和 PHGA 的数据来源于文献[9],为对 BAliBASE 中 141 个测试例的计算值. 表 1 给出了 PHGA-COFFEE,CLUSTAL X,SAGA 和 PHGA 的每一类的平均 SPS 和 CS 分值(PHGA-COFFEE 的参数设为 *Send_rate*=20,*Send_best*=2).

表 1　类 1～类 5 的 SPS 和 CS 分值表(SPS/CS)

方法	SPS 和 CS 分值						
	类 1(82 例)	类 2(23 例)	类 3(12 例)	类 4(15 例)	类 5(12 例)	Avg. 1	Avg. 2
PHGA-COFFEE	0.874/0.812	0.825/0.315	0.727/0.458	0.761/0.509	0.847/0.669	0.840/0.660	0.807/0.553
CLUSTAL X	0.866/0.796	0.857/0.404	0.724/0.499	0.731/0.480	0.837/0.621	0.836/0.661	0.803/0.560
SAGA	0.767/0.666	0.786/0.223	0.607/0.275	0.546/0.175	0.758/0.573	0.737/0.510	0.693/0.382
PHGA	0.854/0.779	0.842/0.364	0.737/0.448	0.760/0.415	0.871/0.746	0.835/0.649	0.813/0.550

注:Avg. 1 为 144 个测试例的平均值,Avg. 2 为 5 个类的平均值.

根据类 1 中的分类,进一步给出 PHGA-COFFEE,CLUSTAL X 的平均 SPS 和 CS 值,如表 2 所示(文献[9]中未给出 SAGA 和 PHGA 的更详细的分值).

表 2　类 1 的 SPS 和 CS 分值表(SPS/CS)

方法	SPS 和 CS 分值		
	V1(相似度<25%)(23)	V2(相似度在 20～40%之间)(31)	V3(相似度>35%)(28)
PHGA-COFFEE	0.654/0.506	0.944/0.904	0.978/0.962
CLUSTAL X	0.651/0.495	0.928/0.876	0.975/0.955

测试结果表明,PHGA-COFFEE 的准确性好于 SAGA,与 CLUSTAL X 和 PHGA 的相当,特别是序列比对的朦胧区(twilight zone)V1,其准确性与 CLUSTAL X 相当,而多序列比对的难点是要提高相似度小于 20～25%的序列比对的准确性. 文中算法采用 COFFEE 函数作为记分函数,对于双序列比对库,一般可使用已有的序列比对程序生成双序列比对库,如 CLUSTAL W 双序列比对库,或多种方法结合生成的双序列比对库,随着双序列比对库准确性的提高,PHGA-COFFEE 比对的准确性可进一步提高.

上述 4 种方法中,CLUSTAL X 属于渐进比对方法的范围,PHGA-COFFEE、SAGA 和 PHGA 属于迭代比对方法的范围,且都基于遗传算法. 渐进比对的方法计算时间少,而迭代比对的方法计算时间长,这是由方法本身的性质决定的. 表 3 给出了用 PHGA-COFFEE,SAGA 和 PHGA 三种方法得到上述结果时,对所有测试例的计算时间(SAGA 和 PHGA 的数据来自文献[9]).

表 3 PHGA-COFFEE, SAGA 和 PHGA 的 CPU 运行时间

方法	运行时间(s)	机器类型
PHGA-COFFEE (144 例)	4553	具有四个节点的机群,节点为一台具有 1 个 Intel 奔腾 IV3.0GHz 处理器、1GB 内存的 Acer 计算机
PHGA-COFFEE (144 例)	16320	一台具有 1 个 Intel 奔腾 IV3.0GHz 处理器、1GB 内存的 Acer 计算机
SAGA (141 例)	207700	一台具有 4 个 450MHz 处理器、1GB 内存的 Sun Ultra 80 计算机
PHGA (141 例)	31540	一台具有 4 个 450MHz 处理器、1GB 内存的 Sun Ultra 80 计算机

从表 3 中看到,计算时间长是迭代比对方法的一个弱点,特别是基于遗传算法的求解方法,就 PHGA-COFFEE 而言,其在有 4 个计算结点的机群系统上的运行时间与在 1 个计算结点上的运行时间相比显著缩短. 在种群规模固定的情况下,其运行时间与计算结点的个数和算法参数的设置有关,表 4 给出了在有 4 个计算结点的系统上,随参数 *Send_rate* 和 *Send_best* 的不同设置,PHGA-COFFEE 的运行时间.

表 4 *Send_rate* 和 *Send_best* (SR/SB) 不同值设置下 PHGA-COFFEE 的运行时间

Send_rate/*Send_best*	运行时间(s)
5/1	3,711
10/2	4,308
20/2	4,553
30/2	5,006
40/2	5,002

在表 4 中所示的 *Send_rate* 和 *Send_best* 的 5 种不同设置中,20/2 的比对准确性最好,其它情况下,类 2 和类 3 的比对准确性有所下降,*Send_rate* 太小,算法收敛速度快,但解的精度下降,*Send_rate* 太大,趋于串行,算法收敛速度与解的精度均不能提高.

基于遗传算法的迭代比对方法与渐进比对方法的结合使用(即将渐进比对法得到的比对结果再进行迭代比对,如将 CLUSTAL X 产生的比对结果,作为一个种子加入到初始种群中),经遗传算法作用后,可得到更加准确的比对结果.

5 结束语

本文针对多序列比对问题,设计了一个适用于该问题的并行混合遗传算法,以 COFFEE 函数作为该算法的适应度函数,在此基础上设计了 ConsistencyShuffle 和 SegmentDP 两个变异算子,使得该算法的整体搜索能力和局部搜索能力大大提高. 通过对 BAliBASE 的测试,验证了该算法的有效性,与已有的算法相比,该算法对处于朦胧区和具有 N/C 末端延伸的序列比对问题有更强的问题求解能力.

通过对 PHGA-COFFEE 的分析,如果将双序列比对的全局比对和局部比对的结果结合起来,可使序列功能区域的比对更加准确,从而使双序列比对库的比对准确性得以提高,使算法的性能进一步提高.

参 考 文 献

1 Atwood T. K. *et al*. Luo Jin-Chu *et al*. translate. Introduction to Bioinformatics. Beijing: Peking University Press, 2001 (in Chinese)
(Attwood T. K., Parry-Smith D. J. 著,罗静初等译. 生物信息学概论. 北京:北京大学出版社, 2002)

2 Needleman S. B., Wunsch C. D.. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970, 48(3): 443～453

3 Carrillo H., Lipman D. J.. The multiple sequence alignment problem in biology. SIAM Journal on Applied Mathematics, 1988, 48(5): 1073～1082

4 Hogeweg P., Hesper B.. The alignment of sets of sequences and the construction of phylogenetic trees: An integrated method. Journal of Molecular Evolution, 1984, 20(2): 175～186

5 Feng D. F., Doolittle R. F.. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution, 1987, 25(4): 351～360

6 Taylor W. R.. A flexible method to align large numbers of biological sequences. Journal of Molecular Evolution, 1988, 28(1～2): 161～169

7 Thompson J. D., Higgins D. G., Gibson T. J.. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 1994, 22(22): 4673～4680

8 Notredame C., Higgins D. G.. SAGA: Sequence alignment by genetic algorithm. Nucleic Acids Research, 1996, 24(8): 1515～1524

9 Nguyen H. D., Yoshihara I.. Aligning Multiple Protein Sequences by Parallel Hybrid Genetic Algorithm. Tokyo, Japan: Universal Academy Press, 2002, 123～132

10 Thompson J. D., Plewniak F., Poch O.. A comprehensive

comparison of multiple sequence alignment programs. Nucleic Acids Research, 1999, 27(13): 2682～2690

11 Notredame C., Holm L., Higgins D. G.. COFFEE: An objective function for multiple sequence alignment. Bioinformatics, 1998, 14(5): 407～422

12 Eddy S.. Multiple alignment using hidden Markov models. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology. Cambridge, England: AAAI/MIT Press, 1995, 114～120

13 Wang L., Jiang T.. On the complexity of multiple sequence alignment. Journal of Computational Biology, 1994, 1(4): 337～348

14 Notredame C.. Recent progresses in multiple sequence alignment: A survey. Pharmacogenomics, 2002, 3(1): 131～144

15 Henikoff S., Henikoff J. G.. Amino acid substitution matrices from protein blocks. In: Proceedings of the National Academy of Sciences of the USA, Washington, USA, 1992, 10915～10919

LIU Li-Fang, born in 1972, Ph. D. candidate, lecturer. Her research interests include parallel computing, pattern recognition and bioinformatics.

HUO Hong-Wei, born 1963, Ph.D., professor. Her research interests include algorithm design and analysis, parallel computing and bioinformatics.

WANG Bao-Shu, born 1942, professor. His research interests include intelligent information processing, pattern recognition and intelligent control.

Background

Multiple sequence alignment is a basic tool in various aspects of molecular biological analyses ranging from detecting key functional residues to inferring the evolutionary history of a protein family. Genetic algorithm(GA) based methods have been used successfully as a practical way to solve the problem. When used alone, however, GA-based methods have the drawbacks of relatively poor quality and slow speed. This paper presents a new GA-based method for more efficient multiple sequence alignment. First, an appropriate cost function is used for the method. Next, dynamic programming method is investigated and incorporated into the algorithm. Finally, the algorithm is parallelized to reduce its run time. Extensive simulation results prove the feasibility of the algorithm.