一种基于遗传算法的DNA多序列比对方法

龚道雄，阮晓钢
（北京工业大学 电子信息与控制工程学院，北京 100022）

摘 要：为了克服遗传算法应用于多序列比对时所遇到的比对序列数受限制以及比对寻优速度慢的缺点，提出了一种基于遗传算法的DNA多序列比对方法(GAMA)；针对DNA多序列比对的特点，指出了传统遗传算法中的交叉操作将为序列比对带来沉重的计算负担；避开遗传算法通常所采用的遗传操作算子，设计了独特的遗传算子(插入删除算子和合并分离算子)、基于BLAST相似度评分方法和完全比对块加权的个体适应度值评价函数，采用了便于插入和删除操作以及相似度评分的基于字符和空位矩阵的染色体编码方案．本算法具有操作算子数量少，算子调用机制简明的特点．最后，给出了将GAMA应用于DNA多序列比对的算例，实验结果验证了本算法的可行性．

关键词：遗传算法；DNA多序列比对；遗传算子
中图分类号：TP 18 **文献标识码**：A **文章编号**：0254－0037(2003)01－0019－04

DNA多序列比对是现代生物序列分析的重要内容．通过多序列比对，可以预测新序列的结构和功能，可以分析序列之间的同源关系，以及进行系统发育分析．多序列比对是一个具有极高计算复杂度的组合优化问题[1]．遗传算法是一种建立在自然选择和进化进程概念基础之上的求解复杂系统优化问题的通用方法，该方法因具有不易陷入局部极小，能处理大型搜索空间，易于与其他优化方法结合，以及几乎适用于所有连续和离散优化问题的特点而获得了广泛的应用[2]．一些研究者就基于遗传算法的多序列比对方法进行了有益的探索[3-7]．Notredame 和 Higgins 的SAGA通过定义了22种遗传操作算子以及算子间的自动调用方法取得了较好的多序列比对质量，但当参与比对序列达到20个以上时比对的速度就非常慢[3]．Anbarasu 和 Sundararajan 研究了基于并行遗传算法(parallel adaptive genetic algorithm)的多序列比对方法[4]．Horng 和 Lin 等将动态规划方法和遗传算法相结合进行多序列比对，但只对于具有高度相似性的长序列在序列数量较少时才能取得较好的结果[6]．虽然上述研究在一些方面取得了满意的结果，遗传算法应用于多序列比对尚有许多问题有待于进一步的研究和探索．

1 问 题

多序列比对的目标是使得参与比对的序列中有尽可能多的列具有相同的字符，即使得具有相同碱基的位点位于同一列，这样便于发现不同序列之间的相似部分，从而推断它们在功能和结构上的相似性．

问题描述：DNA多序列比对过程可以表示成为一个五元组：

$$\mathrm{MAS} = \langle \Sigma, S, A, O, F \rangle \tag{1}$$

其中：1)$\Sigma = \Sigma' \cup \{\_\}$ 为多序列比对的符号集，$\Sigma' = \{A, T, C, G\}$ 为组成DNA的4种碱基，_ 为空位符，表示比对中插入的空位．2)S 为待比对的序列集，每个序列由数量不等的字符组成．$S = \{s^{(i)} \mid i = 1, 2, \cdots, m\}$，$s^{(i)} = (c_1^{(i)}, c_2^{(i)}, \cdots, c_l^{(i)})^{\mathrm{T}}$．其中，$l$ 为序列的长度，$c_j^{(i)} \in \Sigma'$ 为序列 $s^{(i)}$ 中的第 j 个字符．3)$A = (a_{ij})_{m \times n}$ 为多序列比对的结果矩阵，其中，$a_{ij} \in \Sigma$．矩阵的每一列为一个位点上的比对，矩阵的第 i 行对应

收稿日期：2002-06-20．
基金项目：国家自然科学基金资助重点项目(60234020)；国家自然科学基金资助项目(50274003)．
作者简介：龚道雄(1968－)，男，博士生．

于参与比对的第 i 个序列，序列中非空位字符的先后顺序在比对中不能改变. 4) O 为基本比对操作集，O = { insert_gap, delete_gap }，即插入和删除空位操作. 5) F 为在基本操作集之上实现最优比对的策略，它确定一系列在特定位点上的空位插入和删除操作.

F 即为所研究的中心问题. GAMA 算法可用一个四元组表示：

$$\text{GAMA} = \langle\ Sel, I, U, E\ \rangle \tag{2}$$

其中：1) Sel 为选择操作算子，实现从群体中选择个体进行繁殖的功能；2) I 为插入删除操作算子 Indel，实现空位的插入和删除功能；3) U 为合并和分离操作算子 Undiv，实现特定的子字符串的移动功能；4) E 为遗传算法的评价函数.

在本算法中，每代的群体规模是固定的. 每代所产生的个体均加入到群体中与父代共同竞争，适应度值低的个体被淘汰.

2 算 法

2.1 编码方法

为了遗传操作和评价的方便，在本算法中直接应用由字符和空位组成的矩阵表示一个比对方案，作为遗传算法中的一个个体(individual).

$$A = (\text{Indiv}_{ij})_{m\times n} \qquad \text{Indiv}_{ij} \in \Sigma \tag{3}$$

其中：m 为参与比对的序列数；n 为序列长度，短序列通过在序列的尾部插入空位补足. 例如，两序列比对问题中，S_1 = ACAATG, S_2 = TCAACTATC，则个体的初始编码为：

A C A A T G _ _ _
T C A A C T A T C

2.2 选择算子

为了防止因为超级个体的存在而使算法过早收敛，本算法在选择个体进行繁殖时采取如下策略：高于当代平均适应度值一个标准偏差的个体给定两次繁殖机会；低于当代平均适应度值一个标准偏差的个体给定零次繁殖机会；其余的给定一次繁殖机会.

2.3 遗传操作算子

因为多序列比对不允许改变比对序列中非空位字符串的相对顺序，所以当采用交叉操作并且参与操作的两个个体中有一个的交叉点选定之后，另一个就不能再任意的选取，而且在另一个个体中确定交叉点时还需要进行多次串匹配和比较操作. 设参与比对的序列数为 m，序列的最大长度为 n，遗传算法的群体规模为 P，遗传操作的最大代数为 G，交叉概率为 p_x，则在一次遗传操作中将进行 $k = (1/2)^*G \times P \times p_x \times m$ 次平均字符串长度为 $n/2$ 的字符串比较. 多序列比对的这个特点使得采用交叉算子将耗费大量的计算，极大地增加了计算复杂性. 交叉算子的使用是使得遗传算法在求解多序列比对问题时速度慢和可比对的序列数受限制的主要原因. 因此，在本文中取消了交叉操作算子.

2.3.1 Indel 算子 Indel 算子包括插入 (insert) 和删除 (delete) 两个算子. 插入和删除操作单对空位而言. 当满足插入概率时，向序列中随机地插入一个空位. 当满足删除概率时，从序列中删除随机选定的空位. 插入和删除的概率相对较大.

在一个序列中插入或删除空位之后，其他序列也必须各随机插入或删除一个空位. 即无论是插入还是删除操作，向每一个序列中插入或从每一个序列中删除的空位数应该相同，以便维持比对矩阵的每一行字符串长度相等. 若个体中有一列全为空位，则删除该列.

2.3.2 Undiv 算子 Undiv 算子包括合并 (union) 和分离 (divide) 两个算子，是插入和删除算子的复合运算. 这两个算子的操作对象只能是纯粹的字符串或空位串，而不能是字符与空位的混合串. Undiv 算子的操作概率相对较小.

合并算子：所选择的空位(或字符)串向前或向后移动，越过相应的字符(或空位)串直到与最邻近的空位(或字符)串合并为一体. 所选择串的移动方向随机决定.

分离算子：若所选择的空位串或字符串长度大于2，则随机地截断为两部分，随机地选择前(或后)段向前(或后)移动一定的距离．该移动距离至少为一个位置，至多为整个相邻的相异串的长度．

2.4 评价函数

对于一个比对结果A，比对的评价函数定义为A所有列中的两两字符比对评分之和减去A中的空位罚分之和

$$E(A)=\sum_{j=1}^{n}W(x_j)-\Sigma g_p \tag{4}$$

其中

$$W(x_j)=\sum_{p<q}w(x_j^p,x_j^q) \tag{5}$$

$W(x_j)$为比对结果中一列的评价值；$w(x_1,x_2)$为两两字符间比对的评分函数，w为$\Sigma'\times\Sigma'\to R$，具体的值可根据特定的字符间的取代矩阵求得；$g_p$为空位罚分，其罚分规则为

$$g_p(k)=e\times(k-1)+g_0 \tag{6}$$

其中：$g_0=10$，为空位开放罚分；$e=2$，为空位延伸罚分；k为空位串的长度．位于序列首尾的空位不罚分．

字符两两比对的评分采用如下的BLAST相似度评价公式

$$w(x_j^p,x_j^q)=\begin{cases}0 & \exists x\in\{x_j^p,x_j^q\}\ x='\_'\\ +5 & x_j^p=x_j^q\\ -4 & x_j^p\neq x_j^q\end{cases}\Bigg\}\forall x\in\{x_j^p,x_j^q\},x\neq'\_' \tag{7}$$

定义由连续的整列具有相同字符的列所组成的块为完全比对块，因为像完全比对块这样连续的完全比对成功的列是我们所期望的，所以在评价函数中应加大其评价权重．定义完全比对块(见图1)的评价值计算方法为

$$S_{\text{block}}=k\times\sum_{i=1}^{n}S^{(i)} \tag{8}$$

F	V	Q	R	E	L	F	Q
R	F	Q	R	E	L	E	T
I	L	Q	R	E	L	G	T
T	L	Q	R	E	L	Y	K

图1　一个4列的完全比对块

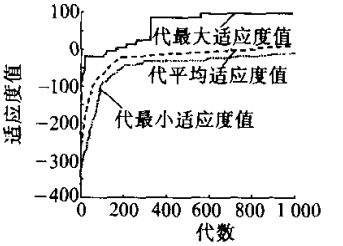
其中：权值$k=\begin{cases}n & n>1\\ 1.5 & n=1\end{cases}$；$n$为完全比对块的列数；$S^{(i)}$为单列的评价值．

3 算　例

图2为应用本算法进行DNA多序列比对的一例．该例中的遗传算法参数取值如下：群体规模为100；最大代数为1 000；插入、删除、合并和分离概率分别为0.8、0.4、0.1和0.1．由图可见，本算法获得的比对质量是令人满意的．图中遗传操作过程每一代的适应度值的变化趋势显示，算法的收敛速度较快，尤其是算法首达最优解的时间较早(图中最大值的较长水平线部分说明了这一点)，而该最优解正是所要求的．本例说明，GAMA 算法能够取得良好的比对质量．

```
A C A A T G A C C G A T C _ _ _ _ _ _ _ _ _
T C A A C T A T _ C A C C _ G A T C A C G A
A C _ _ _ _ A _ C _ A G C A G A A T C A A C
_ _ _ _ _ _ A _ G A A T C C T A A G A _ _ _
A C C _ _ _ _ _ _ G A T C C T A _ T C A A C
```

(a) GAMA的比对实例



(b) GAMA操作中适应度值变化曲线

图2　DNA多序列比对实例及遗传操作适应度值的变化曲线

4 结束语

作者对DNA多序列比对问题进行了一些探索,提出的方法可以取得良好的比对质量. 进一步提高算法寻优的质量和效率,是下一步要进行的工作.

参考文献:

[1] BBAXEVANIS A D, FRANCIS OUELLETTE B F. 生物信息学:基因和蛋白质分析的实用指南[M]. 李衍达, 孙之荣, 译. 北京: 清华大学出版社, 2000.

[2] MICHALEWICZ Z. 演化程序:遗传算法和数据编码的结合[M]. 周家驹, 何险峰, 译. 北京: 科学出版社, 2000.

[3] NOTREDAME C, HIGGINS D G. SAGA: sequence alignment by genetic algorithm[J]. Nuc Acids Res, 1996, 24(8): 1515-1524.

[4] ANBARASU L A, SUNDARARAJAN V. Multiple sequence alignment using parallel adaptive genetic algorithm[A]. Simulated Evolution and Learning[C]. Berlin: Springer, 1999. LNAI1585, 130-137.

[5] WAYAMA M, TAKAHASHI K, SHIMIZU T. An approach to amino acid sequence alignment using a genetic algorithm[J]. Genome Informatics, 1995, 6: 122-123.

[6] HORNG J T, LIN C M, LIN B J. Applying genetic algorithm to multiple sequence alignment[A]. In Proc of the Genetic and Evolution Computation Conference[C]. Las Vegas: Morgan Kauf mann Publishers, 2000. 883-890.

[7] HANADA K, YOKOYAMA T, SHIMIZU T. Multiple sequence alignment by genetic algorithm[J]. Genome Informatics, 2000, 11: 317-318.

DNA Multiple Sequence Alignment Method Based on Genetic Algorithm

GONG Dao-xiong, RUAN Xiao-gang

(College of Electronic Information & Control Engineering, Beijing University of Technology, Beijing 100022, China)

Abstract: In order to solve the problems of both the alignment sequences number limitation and time-consuming which genetic algorithm will encounter in multiple sequence alignment, the authors propose a DNA multiple sequence alignment method based on genetic algorithm (GAMA). Aimed at the characteristics of DNA multiple sequence alignment, they point out that the traditional crossover operation will dramatically aggravate the computation burden of GA, and they adopt two new kinds of genetic operators (Indel & Undiv operators) in GAMA instead of the crossover and mutation operators used in normal GA. The BLAST similarity score matrix and an absolute alignment block weighed fitness evaluation function are adopted. The characters and gaps composed matrix, which is convenient for genetic operation as well as individual evaluation, are also adopted to coding the chromosome of population. There are only a few genetic operators in the scheme of this paper, of which the scheduling mechanism is also simple. This paper present an example of applying GAMA to DNA multiple-sequence alignment, the experiment results validate the feasibility of this algorithm.

Key words: genetic algorithm; DNA multiple sequence alignment; genetic operators