ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.2, February 2007, pp.185−195
DOI: 10.1360/jos180185

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于最大权值路径算法的 DNA 多序列比对方法*

霍红卫+, 肖智伟

(西安电子科技大学 计算机学院,陕西 西安 710071)

A Multiple Alignment Approach for DNA Sequences Based on the Maximum Weighted Path Algorithms

HUO Hong-Wei+, XIAO Zhi-Wei

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

+ Corresponding author: Phn: +86-29-88202363, Fax: +86-29-88202837, E-mail: hwhuo@mail.xidian.edu.cn

Huo HW, Xiao ZW. A multiple alignment approach for DNA sequences based on the maximum weighted path algorithms. *Journal of Software*, 2007,18(2):185−195. http://www.jos.org.cn/1000-9825/18/185.htm

Abstract: For multiple sequences alignment problem in molecular biological sequence analysis, when the input sequence number is very large, many heuristic algorithms have been proposed to improve the computation speed and the quality of alignment. An approach called MWPAlign (maximum weighted path alignment) is presented to do global multiple alignment for DNA sequences. In this method, a de Bruijn graph is used to express the input sequences information, which is recorded in the edges of the graph. As a result, a consensus-finding problem can be transformed to a maximum weighted path problem of the graph. MWPAlign obtains almost linear computation speed of the multiple sequences alignment problem. Experimental results show that the proposed algorithm is feasible, and for a large number of sequences with mutation rate lower than 5.2%, MWPALign can obtain better alignment results and has lower computational time as compared to CLUSTALW (cluster alignments weight), T-Coffee and HMMT (hidden Markov model training).

Key words: multiple sequence alignment; de Bruijn graph; consensus sequence; maximum weighted path

摘 要: 针对生物序列分析中的多序列比对问题,当输入数据量比较大时,人们提出了很多启发式的算法来改善计算速度和比对结果.提出了用于进行全局 DNA 多序列比对的一种方法:MWPAlign(maximum weighted path alignment).该算法把序列信息用 de Bruijn 图的形式表示,并将输入序列的信息记录在图的边上,这样,就将求调和序列的问题转化为求图的最大权值路径问题,使多序列比对问题的时间复杂度降低到几乎线性.实验结果显示:MWPAlign 是可行的多序列比对算法,尤其对于变异率低于 5.2%的大量序列数据,相对于 CLUSTALW(cluster alignments weight),T-Coffee 和 HMMT(hidden Markov model training)有较好的比对结果和运算性能.

关键词: 多序列比对;de Bruijn 图;调和序列;最大权值路径
中图法分类号: TP301 文献标识码: A

* Supported by the National Natural Science Foundation of China under Grant No.69601003 (国家自然科学基金); the Natural Science Foundation of Shaanxi Province of China under Grant No.2005F33 (陕西省自然科学基金)

Received 2006-04-09; Accepted 2006-07-26

多序列比对是生物信息学中挑战性的问题之一,并在序列装配、序列注释、基因和蛋白质的结构和功能预测以及系统发育和进化分析等方面应用广泛.它是 SPS(sum-of-pairs scoring)意义下的 NP 完全问题.

大多数现有的多序列比对方法可以分为 4 类:精确比对方法、渐进比对方法、迭代比对方法、基于图论的比对方法[1].

精确比对方法完全基于动态规划算法,最为经典的是多维 Needlman-Wunsch 算法[2],但其可行的计算维数为 3.

渐进比对算法的基本思想是:迭代地利用两序列动态规划算法,先由两条序列的比对开始,逐渐添加新序列,直到所有序列都加入为止.但是,不同的添加顺序会产生不同的比对结果,所以,确定合适的比对顺序是渐进比对方法的一个关键问题.而两个序列越相似,人们对它们的比对就越有信心,因此,整个序列的比对应该从最相似的两个序列开始,由近至远逐步完成.基于这种方法的软件很多,主要有 CLUSTALW(cluster alignments weight)[3]和 T-Coffee[4]等,其中,CLUSTALW 是一个使用最广泛的渐进比对程序,它主要是根据两两比对的分值构建了一个进化树,然后根据进化树进行渐进比对.T-Coffee 类似于 CLUSTALW,利用扩展库取代 CLUSTALW 中的替代矩阵进行渐进比对,使得每一步渐进比对过程中用到的打分信息都取自于所有序列之间的关系信息,而不仅仅是当前要比对的序列,尤其是在比对初期,可以减少比对错误的几率.但是,T-Coffee 运行时间比 CLUSTALW 要慢.其他针对蛋白质或短 DNA 序列的基于渐进比对的多序列比对软件有 MULTALIGN(multiple alignment)[5],MAFFT(multiple sequence alignment by fast Fourier transform)[6],MUSCLE(multiple sequence comparison by log-expectation)[7],Align-m60[8]和 PROBCONS(probabilistic consistency)等[9].

迭代比对方法基于一个能产生比对的算法,并通过一系列的迭代方式改进多序列比对,直到比对结果不再改善为止.基于这种思想的方法很多,例如模拟退火、遗传算法、隐马尔可夫模型[10,11]等.其中,最有影响的多序列比对软件包 SAGA(sequence alignment by genetic algorithm)[12]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.

基于图模型的比对是近年来发展起来的方法,其主要代表就是偏序比对(partial order alignment,简称 POA)[13]——一种以有向无环图(directed acyclic graph,简称 DAG)的表示方式取代行列表示的全新多序列比对方法.在行列比对方法中,图总是单一的有向路径.然而,POA 方法扩展了图的结构,使得它是一个有向无环图.苛刻的线性行列表示法是过去 30 年来多序列比对的基础,POA 方法打破了这个限制,在多序列比对领域打开了一个全新的视角.后来,Yuzhen Ye 和 Adam Godzik[14]把偏序结构又进一步应用于蛋白质结构比对.2004 年,Pevzner 等人提出了新的基于图论的多序列比对方法 ABA(A-Bruijn alignment)[15].与 POA 不同的是,它把序列比对表示成可能含有环的有向图,使得 ABA 比 POA 具有更大的灵活性,尤其是对于含有交错或重复结构域的蛋白质序列比对和包含重复和倒位的 DNA 序列比对非常适用.其他基于图模型的比对方法还可参见文献[16].

上述方法各有其不同的优点,但它们中的大多数对于大量输入序列,其时空复杂度依然是实际应用的一个瓶颈,至少都是 $O(N^2L^2)$,其中 N 是序列条数,L 是序列平均长度.针对这个问题,本文提出了一种基于图模型的新方法,将 de Bruijn graph 方法应用到 DNA 全局多序列比对中,使多序列比对的时空复杂度降低到线性 $O(NL)$.

在序列装配算法中[17,18],序列装配问题被简化成求 de Bruijn graph 内的欧拉路径问题,其将序列信息记录在 de Bruijn graph 内.基于这种利用图记录序列信息的方法,我们提出基于最大权值路径算法的 DNA 多序列比对方法,本算法用 de Bruijn graph[19]的形式表示输入序列,将输入序列的信息记录在图的边上,定义边的权值为经过该边的序列的条数,则边的权值越大,说明此边越有可能代表输入序列的保守区域.将图中最大权值的边连接起来的最大权值路径,正好对应输入序列中保守区域的归并,也就是所求调和序列对应的路径.设想所有输入序列都是从一个祖先序列进化而来,我们要找的就是这个祖先序列.此过程不需要进行多序列比对,并且使寻找调和序列问题的时间复杂度大为降低,几乎是线性的.最后,利用得到的调和序列和每条输入序列进行两两比对得到比对结果.我们已经使用模拟数据对本算法进行了测试,并且和现有方法进行了比较,结果表明:MWPAlign (maximum weighted path alignment)是可行的 DNA 多序列比对方法,其时间复杂度优于现有的方法,并且在序列变异率较低时,比对结果优于 CLUSTALW,T-Coffee 和 HMMT(hidden Markov model training).

1 问 题

1.1 多序列比对问题

一条长度为 m 的生物序列是由 m 个字符组成的字符串,字符串中的字符取自于一个有限的字母表Σ,对于 DNA 序列,Σ包含 A,T,C,G 四个字母,分别代表 4 种不同的核苷酸,将其统称为碱基.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,将其统称为残基.给定 N 条序列组成的序列组 $S=(s_1,s_2,...,s_N)$,其中:$s_i=s_{i1}s_{i2}\ldots s_{iL_i}$ $(1\le i\le N)$,$s_{ij}\in\Sigma(1\le j\le L_i)$,$L_i$ 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.

$$S'=(s'_{ij}),\ 1\le i\le N,\ 1\le j\le L,\ \max(L_i)\le L\le\sum_{i=1}^{N}L_i.$$

该矩阵有如下特性:

① $s'_{ij}\in\Sigma U\{-\}$,其中,"–"代表空位;

② 如果删除空位"–",则 S'的每一行 $s'_i=s'_{i1}\ s'_{i2}\ \ldots\ s'_{iL_i}$ $(1\le i\le N)$与对应序列 s_i 相同;

③ S'中不存在只由空位"–"组成的列.在第 3.1 节中,图 2~图 5 列出了 10 条序列的不同多序列比对结果.

1.2 多序列比对结果的评判标准

目标函数用来评判序列比对结果的优劣.在多序列比对中,最常用的目标函数是 Sum-of-Pairs(SP)[20].根据 SP 目标函数,在比对结果的每一列中,将每对碱基给定一个分值 P_{score}(例如:$P_{score}(x,x)=2$,$P_{score}(x,y)=P_{score}(x,-)=P_{score}(-,x)=-1$ 和 $P_{score}(-,-)=0$.其中:"–"代表空位;x 和 y 代表两个不同的碱基),然后将这些分值 P_{score} 累加起来,得到每列的分值 C_{score},最后将每列的分值累加,即可得到 SP-Score.假定比对结果为 $S'=(s'_{ij})$,$1\le i\le N$,$1\le j\le L$,则 SP-Score 计算公式如下:

$$SP\text{-}Score(S')=\sum_{i=1}^{L}C_{score}(s'_{1i},s'_{2i},...,s'_{Ni}),$$

$$C_{score}(s'_{1i},s'_{2i},...,s'_{Ni})=\sum_{1\le p\le q\le N}P_{score}(s'_{pi},s'_{qi}).$$

如果输入数据是标准比对库(例如 BALIBASE(benchmark alignment database))中的序列,即有一个标准的比对结果,我们就可以计算一个相对的 SP-Score,定义为 SPS.假定对于标准库的输入序列,标准库中比对结果为 S^*,某方法比对结果为 S',则 SPS 定义如下:

$$SPS=SP\text{-}Score(S')/SP\text{-}Score(S^*).$$

如果没有标准比对库,SPS 定义如下:

$$SPS=SP\text{-}Score(S')/(L\times N\times(N-1)/2).$$

显然,SPS 值反映了碱基对准确对齐的比率.为了反映所有序列准确对齐的比率,通常使用 CS(column score)[19]值来计算.CS 值计算策略为:如果一列上的所有碱基都相等,则 $c_i=1$;否则 $c_i=0$.同样,对于比对结果 S',CS 值计算公式为

$$CS=\sum_{i=1}^{L}c_i\Big/L.$$

基本上,SPS 值和 CS 值越高,说明比对结果越准确,越能反映序列的生物特性.在下面的实验中,将采用 SPS 和 CS 这两个值来评估本算法的比对结果.

2 算 法

MWPAlign 算法解决多序列比对问题的主要思想是:先求调和序列,然后用调和序列和每条输入序列进行两两比对,得到最终比对结果.所得调和序列是输入序列中保守区域的拼接,通过得到的调和序列和每条输入序列的两两比对,就很容易分辨输入序列中保守的碱基和变异的碱基,从而构造多序列比对结果.

2.1 算法描述

本算法用图的模型解决多序列比对问题,其形式化描述为:

算法 **1**. MWPAlign.

Input $S=\{s_1,s_2,\ldots,s_n\}$, each s_i has length l_i.

Output $S'=\{s'_1,s'_2,\ldots,s'_n\}$, each s_{ri} has length m.

1. use S to construct de Bruijn graph $G\leftarrow(V,E)$
2. eliminate cycles of G
3. get a maximum weighted path from G, then construct consensus sequence s_c from the path
4. **FOR** $i\leftarrow 1$ to n

   DO $s'_i\leftarrow$pairwisealignment (s_c,s_i)

construct output result $S'\leftarrow\{s'_1,s'_2,\ldots,s'_n\}$

本算法的核心部分为第 2 步,在去环过程中,尽量不能丢失每条边所代表的序列相似性信息,即要最大限度地保持序列之间的相似性,以保证在算法第 3 步取得与所有输入序列具有最大相似性的调和序列.下面对算法的每一步进行具体描述.

2.1.1 构造 de Bruijn 图

k 个连续的字母(在本文后续描述中,k 始终代表构造图的基本串长度)可以由两点一边组成的结构来表示.两个点分别代表 k 个字母中前后 $k-1$ 个字母,边代表 k 个字母本身,并且从代表前 $k-1$ 个字母的点指向代表后 $k-1$ 个字母的点.

显然,长度为 L 的字符串就可以表示为上述 $L-k+1$ 个结构,然后将每个结构添加到图中,就可以构造一个图 G.添加规则为:如果 G 中存在点或边代表的字符串与结构中的点或边代表的字符串匹配,那么就把匹配的点或边融合成一个点或边,而结构中未被融合的点或边将成为图中新的点或边.基于这样的思想,可将所有输入序列构造成一个图 $G=(V,E)$.

定义 $T(v)$为点 v 所代表的碱基串.构造图的过程形式化描述如下:

算法 **2**. Construct G.

Input $S=\{s_1,s_2,\ldots,s_n\}$, each s_i has length l_i.

Output $G=(V,E)$.

1. **FOR** $i\leftarrow 1$ **TO** n
2. **DO FOR** $j\leftarrow 1$ **TO** (l_i-k+1)
3. **DO** $T\leftarrow s_i(j,\ldots,j+k-1)$
4. $T_L\leftarrow s_i(j,\ldots,j+k-2)$
5. $T_R\leftarrow s_i(j+1,\ldots,j+k-1)$
6. $v_L\leftarrow -1$
7. $v_R\leftarrow -1$
8. **IF** ($hashtable(T_L)$)
9. **THEN** $v_L\leftarrow hashtable(T_L).node$
10. **IF** ($hashtable(T_R)$)
11. **THEN** $v_R\leftarrow hashtable(T_R).node$
12. **IF** $v_L\geq 0$ **AND** $v_R\geq 0$ **AND** there exists $e\in E$ in (v_L,v_R)
13. **THEN** add sequence info $\{i,j\}$ to e
14. **ELSE IF** $v_L<0$
15. **THEN** $v_L\leftarrow$new vertex v_{NL}
16. $V\leftarrow V\cup v_L$

17. $hashtable(T_L).node \leftarrow v_L$
18. **IF** $v_R<0$
19. **THEN** $v_R \leftarrow$new vertex v_{NR}
20. $V \leftarrow V \cup v_R$
21. $hashtable(T_R).node \leftarrow v_R$
22. new arc e_N in (v_L,v_R), add string T and sequence info $\{i,j\}$ to e_N
23. $E \leftarrow E \cup e_N$
24. **RETURN** $G \leftarrow (V,E)$

在算法中,将输入序列中的所有 k–1 个字符组成一个哈希表,将每 k–1 个字符在图中对应的点记录在表中.函数 $hashtable(T)$代表字符串 T 在哈希表中的位置,$hashtable(T).node$ 代表 T 在图中对应点的位置.

根据上述规则,所有输入序列的信息将会被记录在 de Bruijn 图内,其中:每 k 个连续的碱基对应图中唯一的一条边;每 k–1 个连续的碱基对应图中唯一的一个点;每条输入序列唯一对应了图中的一条路径.显然,图中每条边表示了一个序列片断,这个片断有可能是某一条序列的片断,也可能是多条序列共有的片断.而我们的目的就是要找到尽可能多的多条序列共有的片断.因此,如果将每条边的权值设为经过这条边的序列条数,则拥有最大权值的路径也就是输入序列共有片断的归并.为了得到最大权值路径,先对图进行去环.

2.1.2 转化图

原始输入序列存在很多重复序列片断,故初始的图存在很多环.为了去环,采用深度优先搜索算法找到环,然后根据边所代表的序列信息去环.

众所周知,在深度优先搜索过程中,图中的点被着色,以显示它们在搜索过程中的状态.起初,所有点都是白色,当其第一次访问时,将其设为灰色.而当某个点的所有邻接点都设为黑色后(在有向图中,当所有以此点为弧尾的边所指向的点都已经设为黑色),此点将被设为黑色.显然,所有灰色的点将会组成一条路径.如果当前要被访问的点出现在路径中,就说明此路径中存在一个以此点开始并以此点结束的环.

对于图中的一个环,去环的方法是选取环中一点,将环从此点切断.选择切断点的方法为:将环中每相邻的两边之间的联系信息记录下来,将联系信息最少的两边之间的点设为切断点.

切断环的方法为:设切断点为 v,在环路径中,以切断点 v 为弧头的边设为 e_L,以切断点 v 为弧尾的边设为 e_R.新建点 v'将边 e_L 指向点 v'.然后在整个图中,设所有以切断点 v 为弧尾的边集为Σ,对于Σ中的每条边 e,如果 e 与 e_L 有联系信息,则从边 e 中把联系信息分离出来,并以此信息新建一条以点 v'为弧尾的新边.在新建的这些边中,如果存在边 e_N 是由 e_R 中分离出的信息构成,则设边 e_N 所指向的点为新切断点,边 e_N 为新的 e_L,重新开始切断环过程.如此重复,直到不存在由 e_R(即指每次切断环过程中的边 e_R,而非某个具体的边)中分离出信息构成的边为止.经过多次上述过程后,环即可被去除.图 1 举例说明了去环过程(其中,图 1(a)为原始的环,图 1(b)~图 1(d)为去环过程中的 3 个步骤).

如图 1(a)所示,在由点 v_1,v_2 和 v_3 组成的环中,有 3 条边 e_1,e_2 和 e_3.其中:边 e_1 和 e_2 的联系信息有序列 s_1;边 e_2 和 e_3 的联系信息有序列 s_1 和 s_3;边 e_3 和 e_1 的联系信息有序列 s_1 和 s_2.因此,将边 e_1 和 e_2 之间的点 v_2 设为切断点.

在环路中,以 v_2 为弧头的边为 e_1,以 v_2 为弧尾的边为 e_2.如图 1(b)所示,新建点 v_{N1},将边 e_1 指向点 v_{N1}.在整个图中,以点 v_2 为弧尾的边集中只有一条边 e_2,并且经过边 e_1 的序列 s_1 也经过边 e_2,因此,将边 e_2 中与 e_1 有联系的序列 s_1 信息分离出来,利用此信息构造新边 e_{N1}.显然,边 e_{N1} 的信息是由组成环的边 e_2 分离出来的信息构成的,所以,须开始新的去环过程,图 1(c)为以点 v_3 为切断点的去环结果.图 1(d)为以点 v_1 为切断点的去环结果.此时,环就被去除.

显然,上述例子中的环已经被去掉.此时,只要将深度优先搜索的路径回退到查找到环的点,即可继续进行深度优先搜索,查找新的环进行处理.在上述例子中,回退到点 v_1 即可.

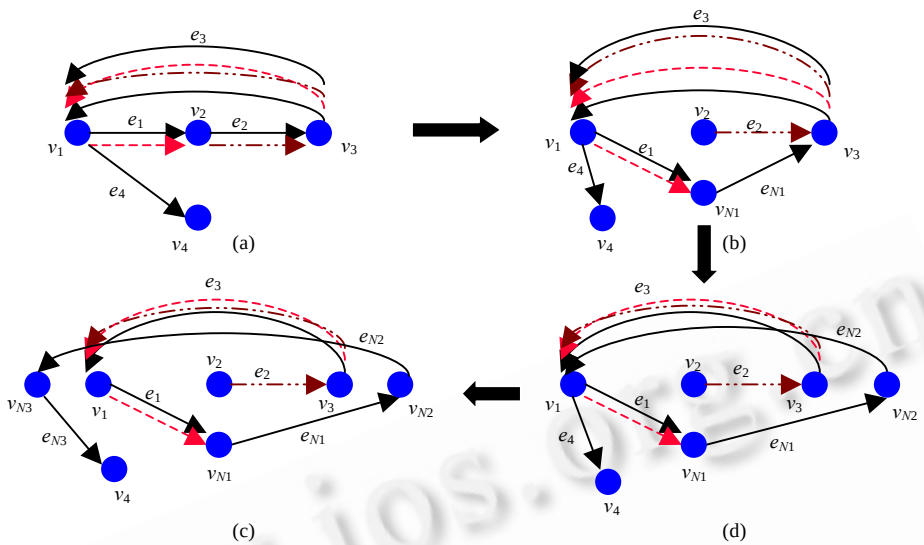

Fig.1 Eliminating cycle

图 1 去环过程

2.1.3 从图中提取调和序列

经过上述去环过程后,已经得到一个有向无环图.为了适应求最大权值路径的需要,定义边的权值为非重叠碱基的个数,假设经过边 e 的序列条数为 N,则权值:$W(e)=N\times(L(e)-k+1)$.然后将每条边的权值取其对应的负值,则找最大权值路径的问题就可以转化为求图的最短路径问题.具体求最短路径的方法描述如下:

首先对有向无环图进行拓扑排序[21,22],按照拓扑顺序,对图中每个点记录其邻接点的路径值,然后从最小路径值的点进行回溯,即可得到最短路径.此算法与 Bellman-Ford 算法类似,其形式化描述如下:

算法 **3**. DAG-Shortest-Path.

Input $G=(V,E)$.

Output P.

1. topologically sort the vertices of G
2. **FOR** each vertex u, taken in topologically sorted order
3. **DO FOR** each vertex $v\in Adj[u]$
4. **DO** $RELAX(u,v,E(u,v))$
5. follow the π information, trace back from the minimum vertex, and construct path P

算法 **4**. $RELAX(u,v,e)$.

1. **IF** $d[v]>d[u]+w(e)$
2. **THEN** $d[v]=d[u]+w(e)$
3. $\pi[v]\leftarrow u$

其中:$Adj[u]$表示以点 u 为弧尾的所有边指向的点集;$d[u]$表示点 u 当前的路径值;$w(e)$表示边 e 的权值;$\pi[v]$记录点 v 在路径中前一个点的信息.具体可参考文献[21].求得最大权值路径后,显然,此路径对应了一条序列,这条序列不一定是输入序列集中的某一条,但它就是所要求的调和序列.

2.1.4 两两比对

得到调和序列以后,可以用动态规划算法进行调和序列与每条输入序列之间的两两比对,每条输入序列的

两两比对结果就是这条序列的最终比对结果.因为得到的调和序列与所要比对的序列之间已经具有一定的相似性,所以,我们采用带有限制的动态规划算法,即只选择在两条序列中的位置距离在一个常数范围内的一对碱基进行比较.具体描述如下:假设限制常数为 L,动态规划矩阵为 M,且矩阵中 X 轴的下标用 i 表示,Y 轴的下标用 j 表示,则动态规划矩阵可初始化为

$$M(i,j)=-\infty \textbf{ if } |i-j|\geq L;\ M(i,j)=0 \textbf{ if } |i-j|<L.$$

通过这样的矩阵初始化设置,即可使所要比对的两个碱基的距离限制在常数 L 范围内,而比对过程与 Needleman-Wunsch 算法相同.在具体比对中,为了减少比对次数,可以对距离超过 L 的两个碱基不进行比对,这样可以使整个比对过程达到几乎线性的时间复杂度.根据实验所得结论,限制常数 L 一般取值为 $\max(60, 2\times|L_1-L_2|)$,其中,$L_1$ 和 L_2 为所要比对的两条序列的长度.

2.2 算法分析

MWPAlign 算法在具体实现中还存在如下一些问题:

(1) k 值大小的选择问题.k 值选择是否合理直接影响最后的比对结果.当 k 值较小时,可以得到更多具有相似性信息的边,但是增加了图的复杂性,使去环过程变得非常复杂,而且易丢失序列相似性信息,从而导致最终比对结果质量下降;当 k 值较大时,具有相似性信息的边就会减少,这样就会把具有较短相似序列串的信息丢掉.经过实验得出结论,一般对于 DNA 序列,k 值应在 10~20 之间,当序列相似度比较大时,取值可以稍大($k\geq 16$);当相似度比较小时,取值可以稍小($k\leq 13$).

(2) 去环过程中的问题.由于输入序列的复杂性,图中存在环不可避免.而在去环过程中,分离一些边的相似性信息,同时也就损失了序列的相似性.所以,如何在去环过程中使相似性信息不丢失,或尽可能少丢失,就是要进一步解决的问题.

算法 MWPAlign 的时间复杂度在 4 步中几乎都是线性的.假设输入序列条数为 N,序列的平均长度为 L.第 1 步,构造图时,总的结构个数为 $N\times(L-k+1)$,则这一步的时间复杂度为 $O(NL)$;第 2 步,转化图时,深度优先搜索的时间复杂度为点数和边数的总和,而图中点数最多为 $N\times(L-k+2)$,边数最多为 $N\times(L-k+1)$,所以,这一步的时间复杂度也是 $O(NL)$;第 3 步,求最大权值路径时,拓扑排序的时间复杂度为点数和边数的总和,是 $O(NL)$.在求最短路径时,对于每个点有一个循环,而其循环次数取决于以此点为弧尾的边的条数,但总的来看,循环的总次数为图中边的数目.因此,求最短路径的时间复杂度也是图中点数和边数的总和,即 $O(NL)$,则这一步总的时间复杂度也是 $O(NL)$;第 4 步,两两之间应用动态规划进行比对,因为是受限的动态规划,所以每次比对的时间复杂度为 $O(L)$,总的比对次数为 N,则这一步的时间复杂度为 $O(NL)$.整个算法每一步都是线性的复杂度 $O(NL)$,故其总的时间复杂度几乎可以达到线性的 $O(NL)$.

3 实 验

MWPAlign 已经在 Linux 系统上用 C 语言实现,为了验证其有效性,使用模拟数据进行了多序列比对实验.生成模拟数据的规则为:以一随机序列为基础,分别以 5.2%,10.5%和 16.4%的概率发生变化,这种变化可以是变异、插入和删除这 3 种中的任意一种,选择某种变化的概率是相等的.测试机器为 PC(Intel Pentium 4 1.5GHz, 256M main memory).评价测试结果的标准为 SPS(第 2 种计算方法),CS 值以及各自的运行时间(单位为 s),并且分别与 CLUSTALW1.83(应用最广泛的多序列比对软件)、T-Coffee(基于相容的优化目标函数的渐进比对方法软件)、HMMT(用马尔可夫模型解决多序列比对问题的软件)的实验结果进行比较.另外,还应用 SAGA(基于遗传算法的迭代比对方法软件)针对实验数据进行了测试,其运行时间太长,无法忍受,故本文没有将其比对结果列出.

3.1 实验结果

实验中,选取序列平均长度为 100~500,序列条数为 10~500,序列变异率为 5.2%,10.5%和 16.4%的数据进行测试,并且对每个测试用例都连续运行程序 30 次,最后取其结果的平均值.表 1~表 3 分别描述了 MWPAlign 与

CLUSTALW1.83,T-Coffee 和 HMMT 在不同序列变异率条件下的实验结果比较.由于 T-Coffee 软件只能对条数少于 60 的序列数据进行测试,故当测试序列的条数大于 60 时该算法失效,用"F"表示.图 2~图 5 分别列出了 MWPAlign,CLUSTALW,T-Coffee 和 HMMT 这 4 种方法在序列变异率为 5.2%、序列条数为 10、序列平均长度为 97 时的比对结果.

从图 2~图 5 中可以看出:MWPAlign 比对结果的序列长度为 109,比对上的列数为 67;CLUSTALW 比对结果的序列长度为 107,比对上的列数为 57;T-Coffee 比对结果的序列长度为 107,比对上的列数为 56;HMMT 比对结果的序列长度为 101,比对上的列数只有 11.显然,MWPAlign 的比对结果优于 CLUSTALW,T-Coffee 和 HMMT.

Table 1 Multiple alignment comparison among MWPAlign, CLUSTALW, T-Coffee, and HMMT with sequence mutation rate 5.2%

表 1 MWPAlign,CLUSTALW,T-Coffee 和 HMMT 在序列变异率为 5.2%时,多序列比对结果的比较

N	*L*	MWPAlign			CLUSTALW			T-Coffee			HMMT		
		SPS	CS	T/s	SPS	CS	T/s	SPS	CS	T/s	SPS	CS	T/s
10	100	0.857	0.664	1	0.844	0.581	1	0.830	0.542	2	0.515	0.103	1
10	200	0.831	0.630	2	0.823	0.561	2	0.806	0.519	4	0.510	0.078	1
10	500	0.846	0.657	2	0.833	0.584	4	0.816	0.543	9	0.530	0.095	2
50	100	0.746	0.191	3	0.728	0.171	4	0.703	0.093	100	0.522	0.008	2
50	200	0.741	0.158	7	0.732	0.148	9	0.706	0.064	223	0.730	0.013	3
50	500	0.740	0.193	17	0.700	0.149	55	0.671	0.065	623	0.461	0.010	7
100	100	0.688	0.100	6	0.655	0.098	10	F	F	F	0.474	0.014	2
100	200	0.714	0.107	13	0.666	0.105	35	F	F	F	0.490	0.004	5
100	500	0.700	0.101	36	0.641	0.094	185	F	F	F	0.401	0	20
500	500	0.640	0.090	227	0.561	0.091	4 656	F	F	F	0.273	0	90

Table 2 Multiple alignment comparison among MWPAlign, CLUSTALW, T-Coffee, and HMMT with sequence mutation rate 10.5%

表 2 MWPAlign,CLUSTALW,T-Coffee 和 HMMT 在序列变异率为 10.5%时,多序列比对结果的比较

N	*L*	MWPAlign			CLUSTALW			T-Coffee			HMMT		
		SPS	CS	T/s	SPS	CS	T/s	SPS	CS	T/s	SPS	CS	T/s
10	100	0.722	0.409	1	0.732	0.419	1	0.680	0.304	2	0.417	0.028	1
10	200	0.717	0.395	1	0.723	0.401	2	0.734	0.364	4	0.407	0.014	1
10	500	0.708	0.389	2	0.711	0.393	4	0.703	0.333	9	0.348	0.021	2
50	100	0.653	0.107	3	0.642	0.097	4	0.572	0.008	98	0.386	0	2
50	200	0.676	0.104	5	0.653	0.089	7	0.586	0.008	220	0.397	0	3
50	500	0.655	0.109	16	0.618	0.094	42	0.668	0.003	614	0.352	0	6
100	100	0.628	0.089	6	0.562	0.092	7	F	F	F	0.347	0	2
100	200	0.624	0.089	11	0.559	0.089	27	F	F	F	0.341	0	6
100	500	0.619	0.090	33	0.627	0.090	168	F	F	F	0.271	0	21
500	500	0.553	0.090	223	0.442	0.091	4 650	F	F	F	0.244	0	92

Table 3 Multiple alignment comparison among MWPAlign, CLUSTALW, T-Coffee, and HMMT with sequence mutation rate 16.4%

表 3 MWPAlign,CLUSTALW,T-Coffee 和 HMMT 在序列变异率为 16.4%时,多序列比对结果的比较

N	*L*	MWPAlign			CLUSTALW			T-Coffee			HMMT		
		SPS	CS	T/s	SPS	CS	T/s	SPS	CS	T/s	SPS	CS	T/s
10	100	0.612	0.221	1	0.659	0.239	1	0.623	0.168	2	0.392	0.028	1
10	200	0.594	0.211	1	0.611	0.215	2	0.573	0.138	4	0.374	0.029	1
10	500	0.591	0.188	2	0.599	0.196	3	0.559	0.117	9	0.301	0.006	2
50	100	0.501	0.087	3	0.498	0.091	3	0.447	0	95	0.372	0.002	3
50	200	0.481	0.089	5	0.482	0.088	8	0.432	0	210	0.364	0	4
50	500	0.498	0.092	20	0.492	0.090	43	0.442	0	615	0.241	0	10
100	100	0.471	0.086	6	0.467	0.088	8	F	F	F	0.347	0	4
100	200	0.470	0.089	15	0.450	0.089	28	F	F	F	0.321	0	11
100	500	0.463	0.091	41	0.429	0.090	162	F	F	F	0.209	0	26
500	500	0.423	0.090	220	0.341	0.090	4 648	F	F	F	0.195	0	103

Fig.2 Multiple alignment from MWPAlign

图 2 MWPAlign 多序列比对结果

Fig.3 Multiple alignment from CLUSTALW

图 3 CLUSTALW 多序列比对结果

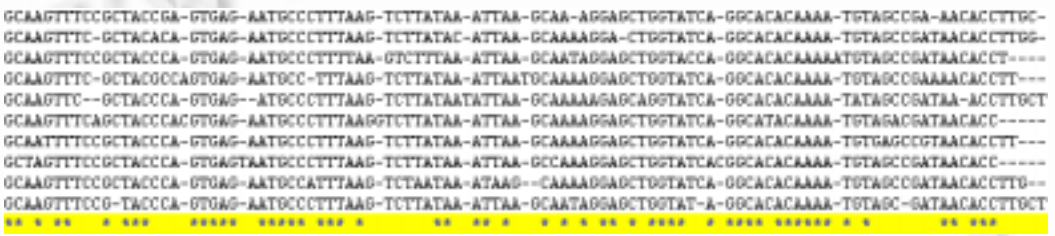

Fig.4 Multiple alignment from T-Coffee

图 4 T-Coffee 多序列比对结果

Fig.5 Multiple alignment from HMMT

图 5 HMMT 多序列比对结果

3.2 结果分析

基于表 1~表 3 的实验结果,下面分别从比对结果和运行时间两个方面对实验结果进行分析.

(1) 比对结果.总体上,MWPAlign 的比对结果明显都优于 HMMT;但是,相对于 CLUSTALW 和 T-Coffee,针对不同数据集的比对结果有优有劣.CLUSTALW 和 T-Coffee 的比对结果相当;但是,T-Coffee 的时间复杂度明显高于 CLUSTALW.在序列变异率为 5.2%时,MWPAlign 的比对结果整体上都优于 CLUSTALW,尤其在序列条数较多时,SPS 值明显优于 CLUSTALW;CS 值与 CLUSTALW 相当.当序列变异率为 10.5%时,在序列条数较少的情况下(10 以下),MWPAlign 比对结果稍差于 CLUSTALW;而当序列条数增大时,其结果又优于 CLUSTALW.当序列变异率为 16.4%时,在序列条数较多的情况下(100 以上),MWPAlign 比对结果优于 CLUSTALW;而当序列条数

较少时,其结果相对于CLUSTALW较差.所以可总结出,MWPAlign针对大量的低变异率的序列数据有较好的比对结果.

(2) 运行时间.MWPAlign 的运行时间总体上少于 CLUSTALW 和 T-Coffee,比 HMMT 多.当序列条数较少时,MWPAlign 的运行时间稍少于 CLUSTALW;但当序列条数较多时,其运行时间明显少于 CLUSTALW.从MWPAlign 的运行时间上可以看出,相对于原始输入序列数据,其运行时间几乎呈线性时间增长,这也进一步证明了本算法解决多序列比对问题有几乎线性的时间复杂度.

当序列之间变异率较大时,保守区域很不明显,也有可能根本就不存在保守区域,所得调和序列不能代表所有输入序列的共同特性,故比对结果较差.所以,在序列之间变异率较大时,改善调和序列的质量,是下一步要解决的问题.

4 结 论

本文提出了一种新的算法 MWPAlign,用图结构解决 DNA 多序列比对问题,其最大的特色有两点:① 不需要进行多序列比对就可以得到包含了所有输入序列中保守区域的调和序列;② 对于大量数据有较好的比对结果和较优的时间复杂度.此算法相对于其他方法可以明显降低时间复杂度,并且在序列变异率较低时取得了很好的比对结果.但是,此算法也有一些不足之处有待改进:当序列之间变异率较大时,比对结果较差;并且,算法本身在去环过程中存在丢失序列相似性信息的情况.这些问题都有待进一步解决.

References:

[1] Batzoglou S. The many faces of sequence alignment. Briefings in Bioinformatics, 2005,6(1):6−22.

[2] Needlman SB, Wunsch CD. A general method application to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970,48(3):443−453.

[3] Thompson JD, Higgins DG, Gibson TJ. 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.

[4] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology, 2000,302(1):205−217.

[5] Barton GJ, Sternberg MJE. A strategy for the rapid multiple alignment of protein sequences. Journal of Molecular Biology, 1987,198(2):327−337.

[6] Katoh K, Misasa K, Kuma K, Miyata T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 2002,30(14):3059−3066.

[7] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 2004,32(5): 1792−1797.

[8] Van Walle I, Lasters I, Wyns L. Align-m—A new algorithm for multiple alignment of highly divergent sequences. Bioinformatics, 2004,20(9):1428−1435.

[9] Do CB, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple alignment of amino acid sequences. Genome Research, 2005,15(2):330−340.

[10] Lukashin AV, Engelbrecht J, Brunak S. Multiple alignment using simulated annealing: Branch point definition in human mRNA splicing. Nucleic Acids Research, 1992,20(10):2511−2516.

[11] Hernández-Guía M, Mulet R, Rodríguez-Pérez S. A new simulated annealing algorithm for the multiple sequence alignment problem: The approach of polymers in a random media. Physical Review E, 2005,72(3):1−7.

[12] Notredame C, Higgins DG. SAGA: Sequence alignment by genetic algorithm. Nucleic Acids Research, 1996,24(8):1515−1524.

[13] Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics, 2002,18(3):452−464.

[14] Ye YZ, Godzik A. Multiple flexible structure alignment using partial order graphs. Bioinformatics, 2005,21(10):2362−2369.

[15] Raphael B, Zhi D, Tang H, Pevzner P. A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Research, 2004,14(11):2336−2346.

[16] Zhang Y, Waterman MS. An eulerian path approach to global multiple alignment for DNA sequences. Journal of Computational Biology, 2003,10(6):803−819.

[17] Idury RM, Waterman MS. A new algorithm for DNA sequence assembly. Computational Biology, 1995,2(2):291−306.

[18] Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc. of the National Academy of Sciences USA, 2001,98(17):9748−9753.

[19] Annexstein FS, De Bruijn G. Sequences: An efficient implementation. IEEE Trans. on Computers, 1997,46(2):198−200.

[20] Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acid Research, 1999,27(13):2682−2690.

[21] Goodrich MT, Write; Huo HW, Trans. Algorithm Design Foundations, Analysis and Internet Examples. Beijing: Posts and Telecommunications Press, 2006 (in Chinese).

[22] Huo HW. Exercises & Solutions on Algorithms. Beijing: Higher Education Press, 2004.

附中文参考文献:

[21] Goodrich MT,著;霍红卫,译.算法分析与设计.北京:人民邮电出版社,2006.

霍红卫(1963－),女,陕西蒲城人,博士,教授,主要研究领域为算法分析与设计,生物信息学算法,进化算法,并行与分布计算.

肖智伟(1980－),男,硕士,主要研究领域为算法分析与设计,生物信息学算法.