多序列比对的两种新方法

陈 莹,王能超

(华中科技大学数学系,湖北 武汉 430074)

摘要:对两种用于多序列比对的新方法——T-COFFEE 和 MAFFT 进行了研究,结果表明:T-COFFEE 是迄今为止准确性最高的方法,但此方法速度较慢;而 MAFFT 则将 FFT 运用于比对中,使速度大大提高,并且与 T-COFFEE 的准确程度相当.

关键词:多序列比对;T-COFFEE;MAFFT;BAliBASE 比对数据库

中图分类号:TP301.6 **文献标识码**:A **文章编号**:1000-1980(2004)02-0239-04

多序列比对就是把两条以上可能有系统进化关系的序列进行比对的方法.实际中一般运用的是几种经典的算法[1],如动态规划算法、Garrillo-Lipman 算法、遗传算法、隐马尔克夫模型(HMM)等.目前使用最广泛的多序列比对程序是 CLUSTALW[2].

本文介绍多序列比对的两种新方法——T-COFFEE 和 MAFFT,并对这两种方法进行了评价和比较.

1 T-COFFEE

1.1 基本思想

T-COFFEE[3] (Tree-based Consistency Objective Function For alignmEnt Evaluation),基于树的比对评估调和目标函数.这种算法提供了一种运用庞杂的数据源创建多序列比对的简单、易用方法.

建立比对的原始库和计算权重.在原始库中先用局部比对和全局比对方法建立双序列比对,局部比对和全局比对分别用 ClustalW 和 Lalign 程序进行;然后给每一对进行比对的残基都分配一个权重,定义每一对所得到的权重,等于在双序列比对中相同残基所占的百分比.例如,下面 2 个序列

SeqA:GARFIELD THE **L**AST **F**AT CAT

SeqB:GARFIELD THE **F**AST **C**AT ---

有 16 个残基匹配,2 个残基不匹配(黑体表示的),可以算得权重 $w(S_A,S_B)=\frac{16-2}{16}\times 100\%=88\%$.

库的连结和扩展.可以用简单程序将 ClustalW 和 Lalign 的基库集中在一起,把局部和全局比对信息有效地结合在一起.基库可直接用于计算多序列比对,并可找到最匹配的残基对的比对.另一方面,用打分系统对所有比对残基进行一致性检验,这样可增大库里信息的准确程度.接着对库进行扩展,令任意残基对经过处理的最后权重能反映整个库中的约束信息,以带权重方式来对信息进行连接.这里采用三联体的方法,即先运用一个已经比对的残基对,然后将其与序列中其余的残基组成三联体进行比对,比对中所取的权值为两两比对时权值的最小值.这样就建立了一个扩展库.

通过对扩展库渐进的比对,则可得到较为准确的结果.

1.2 方法评价

T-COFFEE 方法是现在所有的多序列比对方法中准确程度最高的一种方法.该方法将局部比对与全局比对方法相结合,预处理了所有双序列比对数据.通过预处理,较好地处理了来源于各种相异资源的比对信息.而且,该方法结合全局比对和局部比对的信息来创建库,这样的结果比别的单一方法更为准确.另一方面,运用打分系统,这样不但记录了 2 个序列之间的特殊残基,而且避免了许多局部最小化问题.

但是,这种方法的最大不足是速度较慢,一般情况下复杂度为 $O(N^3L)$(其中 N 为序列数,L 为序列的平

收稿日期:2003-06-19

作者简介:陈莹(1980—),女,广东湛江人,硕士研究生,主要从事高效算法设计研究.

均长度),最坏情况下复杂度将为 $O(N^3L^2)$.所以,T-COFFEE 是一种通过牺牲计算简便性来提高计算准确程度的方法.但当建立好某个固定库的比对信息后,再取库外序列与库内序列进行比较,所需时间将大大缩短,而且所得结果的准确度非常高.

2 MAFFT

2.1 基本思想

MAFFT[4](Multiple sequence Alignment based on Fast Fourier Transform),基于快速傅立叶变换的多序列比对.这种算法把序列信息看成信号,用 FFT 进行'信号处理",较快得到结果,而且准确度与 T-COFFEE 相当.

在生物上,氨基酸的取代频率依赖于其性质,特别是体积(volume)和极性(polarity)2 个指标[5]的区别.对于任意残基 a,用标准化形式 $\hat{v}(a)=\frac{v(a)-\bar{v}}{\sigma_v}$,$\hat{p}(a)=\frac{p(a)-\bar{p}}{\sigma_p}$表示.其中 $\bar{v}$,$\bar{p}$ 分别表示 20 种氨基酸体积和极性的均值,σ_v 和 σ_p 表示标准偏移量.这样每个氨基酸都被转换成向量 $a(\hat{v},\hat{p})$.

氨基酸序列关系的计算.定义 2 个序列之间的关系 $c(k)$:$c(k)=c_v(k)+c_p(k)$.其中:$c_v(k)$,$c_p(k)$分别是体积元和极性元的关系值,$c_v(k)=\sum_{\substack{1\leqslant n\leqslant N\\ 1\leqslant k+n\leqslant M}}\hat{v}_1(n)\hat{v}_2(n+k)$,$\hat{v}_1(n)$ 是长为 N 的序列 1 的第 n 个位置的体积;$c(k)$表示 2 个序列的相似程度,k 为偏移量.取 $c(k)$的最大值.然后,将 $\hat{v}_1(n)$和 $\hat{v}_2(n)$进行快速傅立叶变换得到 $V_1^*(n)$和 $V_2(n)$(其中 $*$ 代表复共轭).$c_p(k)$亦然.

找同源片段,分割同源矩阵.若相比较的 2 个序列有同源区域,那么它们将会和 $c(k)$中的某些峰值相对应.根据峰值对应的 k,对 2 个序列位置平移后进行比对,然后用打分系统找出同源区.建立矩阵 $S_{ij}(1\leqslant i,j\leqslant n,n$ 是同源片段的数量),若序列 1 的第 i 个同源片段与序列 2 的第 j 个同源片段相一致,那么 S_{ij}就为上面所计算的同源片段值,其余 S_{ij}的值设为 0.应用 DP 程序到 S_{ij}上,取得最优化路径.

扩展到多序列比对.考虑 $c_v(k)=\sum_{\substack{1\leqslant n\leqslant N\\ 1\leqslant k+n\leqslant M}}\hat{v}_1(n)\hat{v}_2(n+k)$,而 $\hat{v}_{\text{group1}}(n)=\sum_{i\in\text{group1}}w_i\hat{v}_i(n)$,其中 w_i 是权系数,可以用与 CLUSTALW 一样的方法进行计算.从而扩展到多序列比对.$c_p(k)$亦然.

在 MAFFT 中,采用渐进的方法——FFT-NS-1 和 FFT-NS-2,以及迭代修正的方法——FFT-NS-i 来编制程序.

FFT-NS-1:用 FFT 算法和上述标准化相似矩阵,输入的序列按照导向树的顺序进行渐进的比对,算法的 CPU 时间复杂度为 $O(N^2)$(N 为序列数量).当 k 很大时,可以用 Jones[6]的方法,并就实际做 2 个修改,即一是把氨基酸分成 6 个组,二是记下 6 个组的数量 T_{ij}(序列 i,j 共有的),然后把这个值转换成距离 D_{ij},$D_{ij}=1-\left[\frac{T_{ij}}{\min(T_{ii},T_{jj})}\right]$,再利用 UPGMA[7]方法,用距离矩阵创建导向树.

FFT-NS-2:在 FFT-NS-1 的基础上再次进行重复比对.

FFT-NS-i:把比对分成 2 组进行重复比对,直到没有更高分数出现时为止.

2.2 方法评价

MAFFT,这种方法把比对序列表示成向量序列,从而把比对信息作为信号进行处理,对同源区实现快速识别,其实质是利用 FFT 提高 CPU 的运行速度;而且,MAFFT 对于那些有大量插入、删除操作的相似长度的远源序列也适合.在准确度方面,它所得结果的准确程度与 T-COFFEE 的相当.特别是对于高度保守的序列,FFT 算法可以将 CPU 的时间复杂度减少到 $O(N)$.渐进的方法(FFT-NS-2)和迭代修正方法(FFT-NS-i),在实际操作中也有相当的优势.当输入序列多于 60 时,在相当的准确程度的前提下,FFT-NS-2 比 CLUSTALW 快得多,FFT-NS-i 比 T-COFFEE 快 100 多倍.

3 方法比较

本文用仿真程序 Rose[8]对 T-COFFEE,FFT-NS-2,FFT-NS-i 和 CLUSTALW 几种方法在相同条件下所需要的 CPU 时间进行计算,并用退化分析方法对其进行评估.

图 1 中实验对象共有 40 条输入序列,输入序列相一致的平均百分比为 35% ~ 85%(为高度保守序列).

图 1 中还给出了几种方法的退化系数 Y(与其相对应的是 CPU 的时间复杂度为 $O(X^Y)$,X 为序列长度).由图 1 可见,CLUSTALW 所需的时间与序列长度的平方成正比,TCOFFEE 所需时间则更长,而对于 MAFFT 来说,基本上为 $O(N)$(N 为序列长度).在序列长度不断增大时,MAFFT 方法的优势更为明显.

图 2 中,输入序列的平均长度为 300,输入序列相一致的平均百分比为 35% ~ 85%,还计算出几种方法的退化系数 Y.由图 2 可见,T-COFFEE 的时间复杂度为 $O(K^3)$(K 为序列数量),当序列数量很大时,这种方法就变得不可行,而 MAFFT 花费的 CPU 时间不到 $O(K)$,在大数据量的计算中比较容易实现.

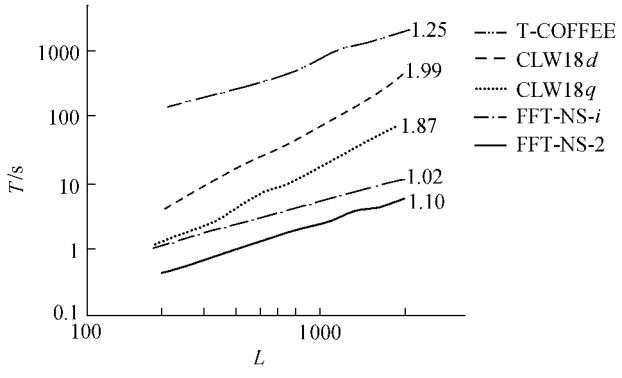

图 1　输入序列长度 L 和 CPU 时间 T 的关系

Fig.1　Relationship between the length of input sequences and CPU time

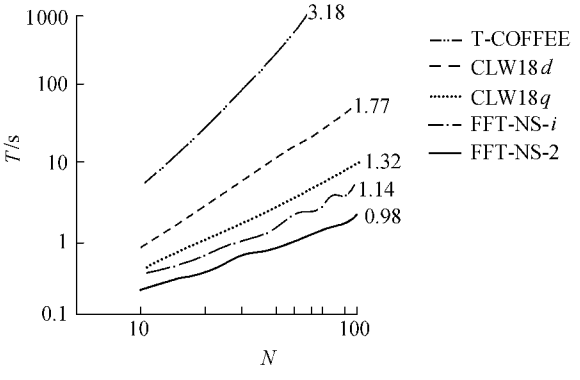

图 2　输入序列数量 N 和 CPU 时间 T 的关系

Fig.2　Relationship between the amount of input sequences and CPU time

运用系统的比对程序对 BAliBASE[9] 比对数据库进行了计算,结果见表 1.BAliBASE(Benchmark Alignment dataBASE)数据库是基于三维结构重叠的'正确'比对数据库,第 1 版的数据分为 5 类,表 1 按此 5 类进行分析.所给出的值为序列对的和分数与列分数(中间用/分开).表中还列出了序列比对所用的 CPU 时间.通过对比,可以看出 MAFFT 这种方法对于远源序列(类 1,2)没有较明显的作用,但在类 3,4,5 中,MAFFT 在速度和准确程度上都大大优于经典算法.总的来看,在众多的方法中,T-COFFEE 的准确程度最高,同时,FFT-NS-i 的准确程度也与之相当,而速度则要快得多.

表 1　各方法对 BAliBASE 比对数据库的计算结果

Table 1　Results of various alignment methods for BAliBASE benchmark tests

方　法		CPU 时间/s	类 1(82)	类 2(23)	类 3(12)	类 4(15)	类 5(12)	平均(144)
渐进方法	CLW18d	2202	0.871/0.792	0.856/0.329	0.754/0.490	0.745/0.417	0.852/0.617	0.844/0.639
	CLW18q	1657	0.871/0.790	0.871/0.790	0.754/0.490	0.728/0.402	0.887/0.709	0.847/0.644
	FFT-NS-2	227	0.849/0.761	0.849/0.761	0.779/0.486	0.797/0.532	0.951/0.826	0.845/0.652
迭代修正方法	T-COFFEE	12065	0.876/0.797	0.856/0.343	0.777/0.497	0.811/0.555	0.961/0.901	0.865/0.683
	FFT-NS-i	1466	0.864/0.787	0.853/0.363	0.789/0.518	0.799/0.534	0.956/0.835	0.857/0.675

注　类后括号内的数为该类比对序列数量.

用 Iwabe[10] 对 RNA 聚合酶的最大子集序列进行了比对(共有 11 块高度保守区域),结果见表 2.由表 2 可知,MAFFT 与 T-COFFEE 的准确程度最高,而 MAFFT 速度比 T-COFFEE 快得多,在某些情况下可快 100 多倍.

表 2　对 RNA 聚合酶的最大子集序列运用几种方法结果的比较

Table 2　Comparison of various methods using the largest subunit sequences of RNA polymerase

样本:76 条序列 × 1182 – 2890 sites			样本:24 条序列 × 1206 – 2890 sites		
方法	CPU 时间/s	正确比对块的数量	方法	CPU 时间/s	正确比对块的数量
T-COFFEE	—	—	T-COFFEE	745.3	11
CLW18d	675.5	10	CLW18d	100.1	9
CLW18q	159.4	10	CLW18q	50.8	9
FFT-NS-2	18.2	11	FFT-NS-2	7.15	11
FFT-NS-i	173.1	11	FFT-NS-i	46.00	11

参考文献:

[1] 郭卫斌,施保昌,王能超. 多重生物序列对准及其算法综述[J]. 高科技通讯,2001,(6):96—102.

[2] 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[J]. Nucleic Acids Res,1994,27:2682—2690.

[3] NORTREDAME C,HIGGINS D G,HERINGA J. T-COFFEE:a novel method for fast and accurate multiple sequence alignment[J]. J Mol Biol,2000,302:205—217.

[4] KAZUTAKA K,KAZUHARU M,TAKASHI M. MAFFT:a novel method for rapid multiple sequence alignment based on fast Fourier transform[J]. Nucleic Acids Research,2000,30:3059—3066.

[5] MIYATA T,MIYAZAWA S,YASUNAGA T. Two types of amino acid substitutions in protein evolution[J]. J Mol Evol,1979,12:219—236.

[6] JONES D T,TAYLOR W R,THORNTON J M. The rapid generation of mutation data matrices from protein sequences[J]. Comput Appl Biosci,1992,18:275—282.

[7] SOKAL R R,MICHENER C D. A statistical method for evaluating systematic relationships[J]. University of Kansas Scientific Bulletin,1958,28:1409—1408.

[8] STOYE J,EVERS D,MEYER F. Generating benchmarks for multiple sequence alignments and phylogenetic reconstructions[J]. Proc Int Conf Intell Syst Mol Boil,1997,5:303—306.

[9] THOMPSON J D,PLEWNIAK F,POCH O. BAliBASE:a benchmark alignment database for the evaluation of multiple alignment programs[J]. Bioinformatics,1999,15:87—88.

[10] IWABE N, KUMA K,KISHINO H,*et al*. Evolution of polymerases and branching patterns of the three major groups of archaebacteria[J]. J Mol Evol,1991,32:70—78.

Two novel methods for multiple sequence alignments

CHEN Ying, WANG Neng-chao

(*Department of mathematics, Huazhong University of Science and Technology, Wuhan* 430074, *China*)

Abstract: The methods of T-COFFEE and MAFFT for multiple sequence alignments are studied. The results show that T-COFFEE provides alignments of the highest accuracy among the methods to date, but its rate is low. As for MAFFT, owing to the application of FFT, the CPU time is greatly reduced, and its accuracy is comparable to that of T-COFFEE.

Key words: multiple sequence alignment; T-COFFEE; MAFFT; BAliBASE alignment database