一个基于 Blast 程序的多重序列对齐程序——Mblast*

孙焕东 张成岗** 贺福初

(军事医学科学院医学信息研究所，北京 100850) (军事医学科学院放射医学研究所，北京 100850)

摘要 核酸序列和蛋白质序列的相似性分析日益成为生物信息学研究的核心内容. NCBI 的 Blast 程序是进行此类分析的最有力工具. 虽然它提供了初步的将多条序列进行综合对齐的分析方案，但是实际效果却很不理想. 在对 Blast 程序的输出结果进行仔细分析的基础上，基于“求同存异”的思想，我们编制了一个多重序列对齐程序 Mblast. 该程序与目前流行的序列多重对齐程序相比，更容易检出序列的同源区.

关键词 Blast 程序，序列多重对齐

学科分类号 Q61

核酸序列和蛋白质序列的相似性分析是生物信息学研究中的一项重要内容，随着人类基因组计划的深入进行[1]，表达序列标签 (expressed sequence tag, EST)、cDNA 序列和基因组序列在不同实验室中被大量获得，从而使大规模的序列相似性分析显得日益重要，尤其是在药物的开发利用中越来越显示出其价值[2]. 核酸序列和蛋白质序列之间的两两比较虽然可以满足一些需要，然而很多工作需要基于多条序列之间的多重序列对齐，以发现序列的保守区和潜在的功能域. 目前流行的序列多重对齐软件主要基于 ClustalW 算法[3]进行分析，例如 ClustalW (http://www2.ebi.ac.uk/clustalw/)、Phylip (http://evolution.genetics.washington.edu/phylip/general.html)、MacVector (http://www.oxmol.com/prods/#bio)、Omiga[4]、DNAMAN (http://www.lynnon.com/) 等. 然而，所有这些软件的共同缺点在于将参与对齐序列的所有碱基或者氨基酸残基引入进行分析，在一定程度上弱化了多重序列对齐的效果，从而不利于潜在功能区域的检出. 美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 的 Blast 程序是一个对序列进行对齐的优秀工具[5]，虽然它提供了初步的将多条序列进行综合对齐的分析方案 (用参数“-m”定义)，但是实际效果却很不理想. 我们在对 Blast 程序的输出结果进行分析的基础上，抽取 Blast 程序输出结果中的同源区段进行拼接，间接地实现了基于 Blast 程序的序列多重对齐分析. 该策略将参与对齐分析的序列的不匹配部分舍去，而将所有的匹配部分进行综合分析，即“求同存异”，较 ClustalW 算法更易于检测出多条序列之间的保守区域，而且扩充了 Blast 程序的功能. 为此，我们开发了软件 Mblast，以 Blast 程序的结果作为输入数据，采用多种匹配与智能搜索技术，实现多重序列的对齐与统计，得到了多种处理结果.

1 基于 Blast 的序列多重对齐的方案

设计 Mblast，就是要充分利用 Blast 的计算结果，实现多重序列的分析与统计结果. 本文拟针对 Blast 程序第 2.0.12 版进行分析，适用于该软件在不同平台 (Windows/DOS、Linux、SGI 及 Internet 方式) 的输出结果. Blast 在序列计算与分析方面的功能极为强大，用户可以设定不同的选项，来形成不同的计算结果. 本文拟仅基于使用它对序列进行两两比较的分析结果，实现序列的多重对齐.

先分析一下 Blast 的输出结果，它可以分为五个部分：第一部分为版本与作者信息，第二部分为查询序列名称及其标识信息如下；第三部分为相似性序列列表；第四部分为查询序列与数据库相似性序列的具体对齐结果；第五部分为 Blast 程序运行

* 国家“863”计划 (863-102-10-04-04)、国家杰出青年科学基金 (39620514)、国家自然科学基金重点项目 (39730310, 39970247)、国家自然科学基金 (39900041, 39900074) 与军事医学科学院科技创新研究启动基金 (9905105) 部分资助项目.

** 通讯联系人.

Tel: 010-66931590, E-mail: zhangcg@nic.bmi.ac.cn

收稿日期：2000-07-12，接受日期：2000-08-23

时所使用的参数描述. 要实现多重序列的对齐, 我们主要采用第四部分进行分析. 具体方法是先读入 Blast 程序的输出文件, 经过内部对齐重组, 形成一对多的碱基序列矩阵. 该矩阵首行记录标准序列, 后续各行依次为各序列与标准序列的对齐结果. 当出现与标准序列不一致的碱基时, 插入空位实现对齐.

2 多重比较的实现技术

由 Blast 的输出结果实现多重对齐, 看似简单, 实为不易. 我们采用的技术是:

第一步, 读入 Blast 的输出结果. 读入文件采用两次扫描方式, 每次读入都要实现文件中各部分的分离. 读入时要区分是基于核酸序列库, 还是蛋白质序列库的计算, 同时还区分是本地化计算, 还是网络化计算. 这些均由程序启动时指定不同参数实现, 同时程序也兼有一定的自动识别能力. 不同的序列库与不同的计算方式, 会输出稍有差别的计算结果, 这在我们的程序中都一一作了区别处理. 第一次扫描读入主要提取查询标准序列的标识名称, 库中搜索序列数, 提取序列的最大长度, 序列对齐的定位, 并产生输出文件 blasts. out (见下文 **3**).

第二步, 为下一步处理作准备, 先分配序列的公共缓冲区 QuerySequence、QuerySt、SjbctSt、QueryBuf 和 SbjctBuf. QuerySequence 是一字串, 用以存贮标准查询序列. QuerySt 和 SjbctSt 为每一对序列对齐时的存贮区, 其结构如下:

```
struct {
int Start, End; //序列的起始位置, 结束位置
int data1; //统计数据
double data2; //统计数据
PSTR sequence; //序列串, 初始为空
}SEQUENCE;
```

QuerySt 存贮该对序列中的标准查询串的部分片段, SbjctSt 存贮该对序列中的对齐串的部分片段. QueryBuf 和 SbjctBuf 分别用于刚读入每一对对齐序列的字串存贮区.

第三步, 进行第二次扫描读入, 直接定位到文件的第四部分, 依次读入每一对对齐序列于 QueryBuf 和 SbjctBuf 中. 读完每一对序列后, 将它们的起始位、结束位、序列串一起存于 QuerySt 和 SbjctSt 的 Start、End 和 sequence 中. 根据每次读入的序列对, 形成标准查询串 QuerySequence. 在 QuerySequence 中记录了标准查询串的标识名, 每一对序列对齐中标准串的起始位置, 结束位置, 及多重对齐后的标准序列串, 这一步是该程序中最关键、最复杂的一步. 当每读入一对序列时, 它要根据标准序列的先后位置插入进 QuerySequence 中, 并对放入 QuerySequence 中的序列及 SbjctSt. sequence 中的对应序列进行对齐, 如果发现由于多重对齐需移位的字母前插入 ":", 作为移位符. 在 QuerySequence 中, 若发现有多个标准查询串重叠时, 通过标记每段的起始位与结束位, 而不是重复存贮该段序列.

第四步, 输出多重对齐结果文件blastcod. out . 在第三步后, 本算法已经将多重序列对齐后的字串存放于它建立的存贮区中, 然而这些存贮区中的数据要想输出还需进行必要的处理. 在 QuerySequence 的标准串是:

Query(S, E)> s1, e1x1> s2, e2x2 …… …> sn, enxn

其中 S 和 E 分别是查询序列的总起始位置和总结束位置, s1 和 e1, s2 和 e2, ……, sn 和 en 分别是各对齐序列标准串的起始位置和结束位置, x1, x2, xn 代理序列的部分片段, > 表示一个新片段的开始. 多重对齐文件的输出, 必须以该字串、QuerySt 和 SbjctSt 中的内容进行分析形成. 输出以 QuerySequence 开始, 先在文件的第一行输出标准序列, 接下来输出各个序列. 为了实现多重对齐的结果, 要不断通过加入空格调整每个序列的开始位置. 在输出多重对齐结果的同时, 还生成了两个统计结果文件blaststa. out 和 blaststb. out (详见 **3**), 以获得序列的构成与分布数据.

第五步, 考虑到按第四步生成的多重对齐结果可能很宽, 达几千甚至上万个字符的宽度, 无法打印输出. 在软件的最后一步中, 又利用 blastcod. out 的输出文件, 形成了可限定宽度的多重对齐结果 blastcop. out. 同时, 还对多重对齐中每一列中出现的各碱基数进行了统计, 并产生了输出文件 blaststc. out.

上述算法不仅可以计算 Blast 的单块数据, 还可计算多块数据.

3 算法产生的输出结果

本算法可产生 6 个输出文件: blasts. out、blaststa. out、blaststb. out、blaststc. out、blastcod. out 和 blastcop. out . 实际上, blasts. out 是 Blast 原结果文件的简化, 是将原结果文件中的对齐阵简化为

只包含对齐序列的首尾信息，以便快速分析序列相似程度.

Blaststa.out 是核酸序列比较中，各碱基（字母）出现的数目统计及其百分比.

Blaststb.out 是 Blast 的一些统计数据（如 Length、Score、Expect、Identities、Strand 及序列的起始与结束位置）输出，并以表格的形成输出.

Blaststc.out 是多重序列对比时每列上的字母统计，每行输出的是每列上各个字母出现的次数的百分比.

Blastcod.out 是多重序列对齐字母矩阵，即对序列对齐结果进行显示，第一行是列号，中间各行是参与对齐的序列对齐结果，最后一行是各列的一致性序列（Consensus）.

Blastcop.out 是多重序列比对结果按指定宽度输出的结果，可供折行打印.

基于上述结果，用户可以很容易根据最后一行的一致性序列判断参与对齐的多重序列之间的保守区域.

4 讨 论

本文报道了基于“求同存异”的思想抽取 Blast 程序输出结果中的同源区段进行拼接，间接地实现基于 Blast 程序的序列多重对齐分析方案. 该策略更易于检测出多条序列之间的保守区域，而且扩充了 Blast 程序的功能.

显然，核酸序列和蛋白质序列的多重对齐分析，将是后基因组时代即功能基因组时代进行生物信息学研究的有力工具. 目前人类基因组的序列草图已经完成，那么，对于 4 种碱基和 20 种氨基酸残基的排列顺序所含有的生物学意义的研究，将对核酸序列和蛋白质序列的多重对齐分析提出更为苛刻的要求. 正如本文所述，目前流行的序列多重对齐软件. 由于将参与对齐序列的所有碱基或者氨基酸残基引入进行分析，在一定程度上弱化了多重序列对齐的效果，从而不利于潜在功能区域的检出. 本文所提出的基于 Blast 程序的多重序列对齐方案. 则充分利用了 Blast 程序可以“抽取”出参与对齐序列的“共同”区段，并将其整理后输出而达到了多重对齐的目的. 实际上体现了“求同存异”的思想，即在多重序列对齐结果中求出所有序列的共同点，而“隐藏”其不同点，显然利于所有参与对齐序列的共同保守区段的获得.

为了便于跨平台运行，我们用 C 语言进行了程序设计，目前已在 Windows/DOS、Linux 等平台中调试通过，并正在将其用于大量新发现基因功能区域的检测过程中[6].

参 考 文 献

1 Collins F S, Patrinos A, Jordan E, *et al*. New Goals for the U. S. Human Genome Project: 1998~2003. Science, 1998, **282** (5389): 682~689

2 Cane D E. Biosynthetic pathways: biosynthesis meets bioinformatics. Science, 2000, **287** (5454): 818~819

3 Thompson J D, Higgins D G, Gibson T J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, **22** (22): 4673~4680

4 Calvet J P. SOFTWARE: Comprehensive sequence analysis. Science, 1998, **282** (5391): 1057~1058

5 Altschul S F, Madden T L, Schäffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, **25** (17): 3389~3402

6 Zhang C G, Yu Y T, He F C, *et al*. Characterization, chromosomal assignment, and tissue expression of a novel human gene belonging to the ARF GAP family. Genomics, 2000, **63** (3): 400~408

Mblast: A Multiple Alignment Program Based on Blast Program*

SUN Huan-Dong

(*Institute of Medical Information, Academy of Military Medical Science, Beijing* 100850, China)

ZHANG Cheng-Gang**, HE Fu-Chu

(*Institute of Radiation Medicine, Academy of Military Medical Science, Beijing* 100850, China)

Abstract The Blast program developed by National Center for Biotechnology Information (NCBI) is one of the powerful tools for sequence analysis including both nucleotide and amino acid sequences. Although it could be used for multiple sequences alignment, the result is not always available. After analysis of the blast result, a new algorithm was designed to optimize the multiple alignment of Blast. A program named "Mblast" was then

developed for application. Result demonstrated that the program is useful in identifying conserved region of multiple sequences.

Key words sequence analysis, multiple sequence alignment, Blast program, bioinformatics

* Partially supported by Chinese High-tech Program (863-102-10-04-04), Chinese National Distinguished Young Scientist Award (39625014), Chinese National Natural Sciences Foundation Key Project (39730310, 39970247), Chinese National Natural Science Foundation General Program (39900041, 39900074) and Initiative Foundation for Scientific and Technological Innovation of Academic Military Medical Science (9905105).

** Corresponding author. Tel: 86-10-66931590, E-mail: zhangcg@nic.bmi.ac.cn

Received: July 12, 2000 Accepted: August 23, 2000

肿瘤坏死因子家族的新成员——TRANCE

叶传忠[1)] 张芳林[2)] 陈常庆[3)]

([1)]复旦大学附属中山医院泌尿外科，上海 200032; [2)]上海第二医科大学附属瑞金医院，上海市内分泌研究所，上海 200025; [3)]中国科学院上海生物工程研究中心，上海 200233)

肿瘤坏死因子相关激活诱导因子 (TNF-related activation-induced cytokine, TRANCE) 是肿瘤坏死因子家族的一个新成员，人TRANCE基因首先由Wong等克隆，它是一种完整的Ⅱ型转膜糖蛋白，由316个氨基酸组成，胞外区与同是TNF家族成员的TRAIL、FasL和TNF等有较高的同源性，TRANCE表达的蛋白质是免疫系统、骨的发生和保持平衡的重要调节因子. 最近，Nagai等通过5′-RACE法获得了分泌型的TRANCE (sTRANCE, secreted form TRANCE) 编码区cDNA片段，并证实其具有TRANCE相同的活性.

TRANCE与T细胞功能及免疫反应有关，可提高成熟树突状细胞 (dentric cell, DC) 的存活. DC是TRANCE在免疫系统中主要的靶向细胞，TRANCE通过提供存活信号及细胞因子调节信号给DC，从而增加其数量、功能并延长其存活，进而增加其T细胞刺激容量. 在将来的 *ex vivo* (回体移植) 实验中，使用DC加上重组的TRANCE能促进T细胞依赖的免疫反应，可用于恶性肿瘤的治疗.

TRANCE还是钙调节激素、成骨细胞以及破骨细胞生成之间的枢钮，TRANCE通过增加成熟破骨细胞的能动性、速度及存活，从而增加其重吸收骨的能力，骨的稳态依赖于TRANCE和OPG的局部浓度的平衡. 因此一些骨的疾患可能是由于TRANE或OPG的异常调节引起，这提示将来有望使用TRANCE信号或OPG功能的拮抗剂或类似物来治疗.

目前我们已成功地克隆了sTRANCE，并利用基因重组技术在大肠杆菌内表达了sTRANCE与MBP的融合蛋白，利用Amyrose resin亲和层析技术对其进行了纯化并证实其具有增殖DC的作用. sTRANCE cDNA的克隆和表达，为研究探讨sTRANCE的生物学性能和开发抗肿瘤药物创造了条件.