原核表达中优化起始密码下游序列的软件设计与实现

Design and Implementation of DB Sequence Optimization Software

许 龙[1*],李 涛[2],周晓巍[1],黄培堂[1]

XU Long[1*], LI Tao[2], ZHOU Xiao-Wei[1] and HUANG Pei-Tang[1]

1 军事医学科学院生物工程研究所,北京 100071

2 军事医学科学院国家生物医学分析中心,北京 100850

1 *Beijing Institute of Biotechnology*, *Beijing* 100071, *China*

2 *National Center of Biomedical Analysis*, *Beijing* 100850, *China*

摘 要 在原核表达中影响外源基因表达效率的因素有很多,这其中翻译起始效率起了非常重要的作用。翻译起始效率又主要受 SD 序列、SD 序列与起始密码子之间的间距、DB(Downstream Box)序列、mRNA 翻译起始区(TIR)的二级结构和稀有密码子等因素的影响。主要针对 DB 序列和 5′端稀有密码子的优化设计了软件。通过计算机对序列进行分析比对后,按照匹配碱基数、匹配位置、密码子使用频率平均值的顺序进行排序,给出一些优化序列,并给出了软件算法。

关键词 DB 序列,16SrRNA,序列匹配性,基因表达,软件设计

中图分类号 Q786 **文献标识码** A **文章编号** 1000-3061(2006)06-1032-04

Abstract TIR (Translation Initiation Region) efficiency is very important in prokaryotic expression. The TIR's efficiency is highly dependent on SD (Shine-Dalgarno) sequence, distance between SD sequence and start codon, DB (Downstream Box) sequence, TIR's second structure, codon adaptation and so on. In this paper, we designed and implemented the software to optimize DB sequence and 5′ rare codons. It generated some optimization sequences by analyzing the target sequence and comparing it with 16S RNA. And the optimization sequences is sorting by number of base pairing, location of base pairing and codon adaptation. We drew up the algorithm and the core of code in this paper.

Key words DB sequence, 16SrRNA, base pairing, gene expression, software design

大肠杆菌由于遗传背景清楚,可进行大规模发酵培养,表达周期短,操作简单及有大量表达载体,而成为人们克隆和表达外源基因的首选。但在外源基因的表达过程中,常常会遇到表达效率不高或根本不表达等困难。以往的研究表明,影响外源基因表达效率的因素主要有密码子的选用、目的基因的量、mRNA 的稳定性、载体的选择、培养条件的控制、启动子的强度和翻译起始效率等。对于目的基因的量和启动子强度,都可以选用合适的载体来控制;而对于密码子选用、mRNA 稳定性和翻译起始效率却只能从基因序列上加以改造来控制。这其中翻译起始效率起了非常重要的作用,而翻译起始效率又主要受 SD 序列、SD 序列与起始密码子之间的间距、DB(Downstream Box)序列、mRNA 翻译起始区(TIR)的二级结构和稀有密码子等因素的影响[1]。

DB 序列在许多原核 mRNA 中有功能性作用,最

Received: April 10, 2006; Accepted: July 11, 2006.

* Corresponding author. Tel: 86-10-66948831; Fax: 86-10-66948833; E-mail: xul_long@126.com

初被认为是一个翻译增强元件,大约由 8 ~ 13 个核苷酸组成,位于起始密码子的下游,与基准序列(16S rRNA 的 + 1469 ~ + 1483 区间的序列)互补,并常常存在于许多高表达的大肠杆菌和噬菌体 mRNA 起始密码子的下游。有报道在 SD 序列存在的时候,DB 序列可促进 T7 基因的翻译[2]。DB 序列可与 SD 序列对翻译起始起协同增强作用,这两个序列均为翻译起始时核糖体需要结合的位点。但在没有 SD 序列存在的时候,DB 序列需要长达 12 ~ 13 个碱基来促进高效表达,而只有 8 ~ 11 个碱基并且中间有间断的时候,活性就大大降低了[2]。有学者在研究一个与 SD 类似的序列在介导柯萨奇病毒 B3 的 RNA 翻译起始过程中发现,不管是对 DB 区的碱基进行敲除或替换,都能显著影响翻译起始的效率[3]。还有学者通过对大量的序列表达进行分析,分析靶序列 5′末端 + 1 ~ + 50 局部区域和整个序列与基准序列间的序列匹配性,来寻找序列匹配性与表达效率间的关系,得出如下结论:基准序列与靶序列 5′端局部(+ 1 ~ + 50)匹配性越好,并且主要匹配区越靠近靶序列 5′端,则目的蛋白在原核系统内表达量越高;如果匹配性很好,但是主要匹配区远离靶序列 5′端,蛋白表达量则越低;如果基准序列与靶序列之间无明显匹配性,则无明显的蛋白表达[4]。通过改变 5′端局部序列来增加与基准序列的匹配性后,有的蛋白表达量占全菌蛋白量的比例由 0.03% 上升到 21%[5]。

另外,稀有密码子的影响因素也是相当重要的。大肠杆菌对编码同一种氨基酸的各种密码子的使用频率并不相同,甚至相差很大[6]。稀有密码子的存在可大大降低蛋白质合成的速率,使蛋白质的表达量降低,甚至使蛋白合成中途停止,形成截短蛋白[7]。通过改变 5′端稀有密码子同时也会降低翻译起始区的二级结构稳定性,从而有利于核糖体的接近与结合,来促进表达量的提高[8]。因此,运用大肠杆菌常用密码子替换稀有密码子,不但可提高重组蛋白的表达水平,甚至在 IPTG 诱导后可促进宿主菌的生长[9]。

因此本文针对 DB 序列及 5′端稀有密码子优化设计了相应的软件。通过计算机对序列进行分析比对后,给出一些优化后的序列,按照匹配碱基数、匹配位置、密码子使用频率平均值的顺序进行排序。

1　软件算法流程图

软件算法的流程图见图 1。

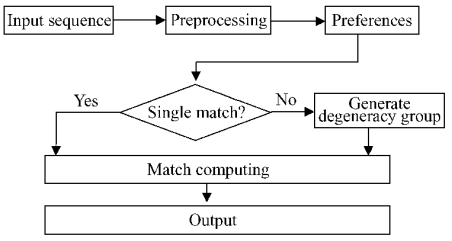

图 1　软件算法流程图

Fig. 1　Algorithm flowchart of the software

2　算法实现

2.1　总体架构

开发系统:Windows XP;开发语言[10]:Borland Delphi 6.0 数据库:Paradox 7 适用系统:Windows 98/ME/2000/XP。

2.2　应用流程

2.2.1　输入序列:用户输入将要筛选的序列,确定后,系统会对输入序列字串进行格式化。包括字符格式化和长度格式化。系统会自动剔除输入序列中所有非 'A'、'C'、'G'、'T' 的字符并要求输入序列字串长度必须是 3 的倍数,即整数个密码子。该步骤容错规则是如果不符合 3 的倍数就进行末尾剪切至满足规则的最大长度。系统默认读码相位是从第一位开始。系统将对输入序列字串中的所有密码子进行遍历检测,如果发现终止密码子则提示用户。由于本程序应用范围限定,靶序列长度不宜过长,超过 30bp 系统将提示用户可能耗费大量时间。

2.2.2　相关设置:程序对用户输入序列进行格式化之后,用户可以进行相关设置。包括选择匹配序列。因为各菌中的基准序列均不一样,所以选择基准序列提供了两种接口:用户可以选用大肠杆菌的基准序列,也可以直接输入其它菌的基准序列。该设计为程序的功能做了拓展。无论用户如何选择,程序执行时都是先将用户的要求转译成要在输入序列字串中搜索的子串然后进行搜索。即用户欲在靶序列中搜索基准序列 ACCGTT,计算机预算的时候就是转换成在靶序列中搜索 AACGGT。

2.2.3　主要算法介绍:

(1)"简并序列生成":计算机先根据密码子的简并性,用每个密码子的所有简并密码子替换一遍靶序列中的密码子。每一次替换产生一个新的序列。我们设计了递归算法实现计算出一个序列的所有简并序列。比如:我们给定序列 seq1:GAATTCCCGGGGATCCGTCGACCTGCAGCC,计算出来该序列有 442 368 条简并序列,临时数据文件有

200M 之多。

(2)“匹配运算”:是指基准序列与靶序列首位对齐后,基准序列以 1bp 为步长在靶序列上滑动直到两序列末端对齐为止,记录每一次的匹配分数。规则是在对应位置上,相同的碱基每个记 1 分,不同则记 0 分。对于每个序列我们记录得分最高的位置以及分值。所有序列匹配运算完毕后,程序进行排序。按照用户的要求输出得分最高的序列。比如 seq1 有 30bp,如果搜索一个 14bp 的基准序列,那么对于每个序列要比对 30 – 14 + 1 = 17 次(就是使两个序列首位对齐到末端对齐的所有可能性)。所以对于 seq1 的所有简并序列我们总共要进行匹配运算 17 × 442 368 = 7 520 256 次。

(3)“排序”:由于数据量巨大,我们相应设计了基于数据库过滤数据的排序方法。比如用户设置输出 N 个匹配分值最大的结果,那么我们搜索数据库中最大的结果,记录其位置,建立索引,然后通过标志字段的值将其过滤去,进而再搜索最大的……循环往复,直至 N 个最大结果产生,取消数据库过滤设置,根据索引找出这 N 个匹配分值最大的结果输出。

(4)“密码子排序”:程序在对每个序列匹配运算的同时还计算这个序列所有密码子使用频率平均值,在匹配分数和匹配位置一样的情况下,优先列出密码子使用频率平均值最高的序列。同时标记出其中的稀有密码子,进行人工取舍。

程序输出结果以匹配得分高低为依据,密码子的使用频率计算结果供用户参考平衡选择。这样用户在使用的时候能够按照要求对靶序列进行改造,在不改变氨基酸序列的情况下,进行与相应序列匹配的优化选择。但是按照笔者的理解及实际经验,此处的‘不改变氨基酸序列’的前提也不是绝对的。因为现在很多序列的 5′端是信号肽,只要不影响该信号肽的生物功能,可以适当根据基准序列改变该信号肽上的氨基酸序列。根据研究报道,适当改造信号肽结构可提高外源蛋白的分泌效率。增加信号肽 N 端的正电荷或增加信号肽疏水核心 H 区的疏水性或长度,有利于提高信号肽的加工效率。Sagiya 等[11] 利用提高 N 端正电荷和 H 区疏水性的方法,使金枪鱼生长激素产率提高了 10 倍,终产量达 240mg/L。因此在软件给出结果后,我们还可以根据软件的结果,进一步进行人工的筛选和优化。

2.2.4 结果输出:以下是一个输出结果的一部分,右边为注释。

```
共筛选了 442368 条序列:                          筛选简并序列总数
以下是匹配值最大的前 8 个序列信息:
>>序列:1 :GAATTTCCAGGAATTAGAAGGCCTGCAGCA
>>匹配序列是:CATGAATTACAAAG
>>最大匹配值:11/14  最佳匹配位置:8                在第 8 位出现最大匹配信息分值 11
>>密码子使用频率平均值:0.264                      所有密码子平均使用频率(原核)
>>密码子使用频率最小值:0.03 位置:19               最小使用频率密码子位置
>>匹配模式:

          *     *  *                             星号表示该密码子低于用户设定值
GAATTTCCAGGAATTAGAAGGCCTGCAGCA                    输入序列
      CATGAATTACAAAG                              匹配序列
      ^^ ^^^^^^^ ^^ ^                             匹配的碱基
12345678911234567892123456789 3                   位置编号
```

3 软件创新点和适用范围

该软件创新之处:(a) 应用了数据库支持数据处理,快速、便捷、有序。(b) 大量数据排序的时候,使用数据库过滤功能,避免大量使用内存。(c) 序列的预处理以及容错功能设计。

该软件适用于:(a) 如果表达载体上不带信号肽,通过表达序列上的起始密码子来启动翻译,则可以用该软件优化表达序列的 5′端;(b) 如果表达载体上携带有信号肽或者某些前导序列,可以进行载体优化改造;(c) 也适用于新表达载体的构建。

REFERENCES(参考文献)

[1] Sprengart ML, Porter AG. Functional importance of RNA interaction in selection of translation initiation codons. *Mol Microbio*, 1997, **24**(1): 19 – 28

[2] Sprengart ML, Fuchs E, Porter AG. The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*. *EMBO J*, 1996, **15**(3): 665 – 674

[3] Yang DC, Paul Cheung, Sun YH *et al*. A Shine-Dalgarno-like sequence mediates in vitro ribosomal internal entry and subsequent scanning for translation initiation of *Coxsackievirus* B3 RNA. *Virology*, 2003, **305**(1): 31 – 43

[4] Jing BQ, Yuan Y. Enhanced translation of 16S rRNA pairing region of mRNA. *Journal of North Sichuan Medical College*(**川北医学院学报**), 2000, **15**(2): 1 – 4

[5] Bucheler US, Werner D, Schirmer RH. Random silent mutagenesis in the initial triplets of the coding region: a technique for adapting human glutathione reductase-encoding cDNA to expression in *Escherichia coli*. *Gene*, 1990, 96(2): 271 – 276

[6] Sharp PM, Devine KM. Codon usage and gene expression level in dictyostelium discoideum: highly expressed genes do prefer optimal codons. *Nucleic Acids Res*, 1989, **17**(13): 5029 – 5039

[7] Kurland C, Gallant J. Errors of heterologous protein expression. *Curr Opin Biotechnol*, 1996, **7**(5): 489 – 493

[8] Zhang SH(**张思河**), Xing JL(**邢金良**), Yao XY(**姚西英**) *et al*. Non-fused expression of HAb18GEF by reducing stability of translational initiation region in mRNA. *Chinese Journal of Biotechnology* (**生物工程学报**), 2004, **20**(2): 175 – 180

[9] Zhou Z, Schnake P, Xiao L *et al*. Enhanced expression of a recombinant malaria candidate vaccine in *Escherichia coli* by codon optimization. *Protein Expression and Purification*, 2004, **34**(1): 87 – 94

[10] Li YX(**李云祥**). Delphi Programming Foundation Course. Beijing: Publishing House of Electronics Industry(**电子工业出版社**), 2004

[11] Sagiya Y, Yamagata H, Udaka S. Direct high-level secretion into the culture medium of tuna growth hormone in biologically active from by *Bacillus brevis*. *Appl Microbiol Biotechnol*, 1994, **42**(2 – 3): 358 – 363