

技术与方法 ·

## pBV220 载体中外源基因高效表达的自动化设计

查磊<sup>1,2)</sup>, 应晓敏<sup>1)</sup>, 王立贵<sup>1)</sup>, 曹源<sup>1)</sup>, 骆志刚<sup>2)</sup>, 苑波<sup>2)</sup>\*, 李伍举<sup>1)</sup>\*

<sup>1)</sup> 军事医学科学院基础医学研究所计算生物学中心, 北京 100850;

<sup>2)</sup> 国防科技大学计算机学院, 长沙 410073

**摘要** pBV220 载体是国内科学家构建的原核系统表达载体, 目前仍在广泛应用. 但是, 实现外源基因的高效表达需要综合考虑诸如 RNA 二级结构等多种因素, 极其耗时费力. 为此, 基于我们提出的 pBV220 载体中外源基因高效表达数学模型, 编写了外源基因高效表达的自动化设计软件, 并可定性用于原核系统其它载体中外源基因表达水平分析, 最终为加快实验进程提供帮助.

**关键词** pBV220; 高效表达; 自动化设计

**中图分类号** Q811.4

### Automated Design for High-expression of Exogenous Genes with pBV220 Vector

ZHA Lei<sup>1,2)</sup>, YING Xiao-Min<sup>1)</sup>, WANG Li-Gui<sup>1)</sup>, CAO Yuan<sup>1)</sup>, LUO Zhi-Gang<sup>2)</sup>, YUAN Bo<sup>2)</sup>\*, LI Wu-Ju<sup>1)</sup>\*

<sup>1)</sup> Center of Computational Biology, Institute of Basic Medical Sciences, Academy of Military Medical Sciences, Beijing 100850, China;

<sup>2)</sup> School of Computer, National University of Defense Technology, Changsha 410073, China

**Abstract** pBV220 is a widely used prokaryotic vector constructed in China to expressing exogenous genes. As various factors, such as RNA secondary structure, may influence the yielding of a vector system, it is time- and labor-consuming to achieve high-level expressions of exogenous genes. Here we introduced a program to evaluate the gene expression with a pBV220 vector based on our mathematical model previously reported. The program may be used for automatic vector designs and could be potentially applied to other prokaryotic expression vectors to supply predictive information for molecular biologists.

**Key words** pBV220; high-level expression; automatic design

pBV220 载体是我国预防医学科学院病毒研究所自行构建的大肠杆菌温控表达载体, 目前仍在广泛使用. 为了实现外源基因在此载体中的高效表达, 我们曾运用基于螺旋区随机堆积的 RNA 二级结构预测<sup>[1]</sup>与密码子偏性计算等生物信息学方法, 对人白细胞介素 2 和人白细胞介素 4 等 22 个外源基因表达水平进行了定量分析, 构建了外源基因高效表达数学模型<sup>[2]</sup>. 在随后的多例实验<sup>[3~5]</sup>中, 模型的正确性得到了验证. 但是, 根据设计规则进行外源基因表达水平判别与高效表达基因设计时, 需要手工对多项数据进行分析与计算, 特别是有时计算需要重复多次, 对于不熟悉生物信息学的实验人员来说, 显得颇为繁琐, 并容易出错. 为此, 根据提出的数学模型, 我们曾开发了外源基因高效表达设计软件<sup>[6]</sup>, 但是该软件是基于 DOS 的软件, 目前使用非常不便, 并且仅针对 pBV220 载体. 为了进一步推广该模型, 并加速有关实验过程, 我们重新编写了软件, 目前已

整合到我们开发的辅助分子生物学实验设计的软件系统 BioSun<sup>[7,8]</sup>中. 该功能模块不仅可以对 pBV220 载体外源基因高效表达进行定量评价和自动化设计, 也可以定性分析原核系统中的其它表达载体, 最终为实现原核系统中外源基因高效表达提供帮助.

### 1 设计原理

为了构建模型<sup>[2]</sup>, 我们首先从文献上收集了 pBV220 载体中 22 个外源基因表达数据. 根据外源

收稿日期: 2008-09-08; 接受日期: 2008-12-09

国家自然科学基金资助项目 (No. 30470411)

\* 联系人 Tel: 010-66931324;

E-mail: liwj@nic.bmi.ac.cn; yuan.33@osu.edu

Received: September 8, 2008; Accepted: December 9, 2008

Supported by National Natural Science Foundation of China (No. 30470411)

\* Corresponding author Tel: 010-66931324;

E-mail: liwj@nic.bmi.ac.cn; yuan.33@osu.edu

基因表达水平,将 22 个外源基因分为两组,如果外源基因的表达量占细菌总蛋白 20% 以上,则称为高表达基因,这样的基因共有 13 例;否则划分为低表达基因,共有 9 例。然后,运用基于螺旋区随机堆积的 RNA 二级结构预测方法<sup>[1]</sup>和密码子偏性计算等工具,研究了 pBV220 载体中外源基因高效表达的定量条件,得到了如下的定量设计规则。

### 1.1 外源基因 5 端与 3 端局部自由能对表达水平的影响

通过对重组质粒中外源基因 5 端与 3 端多个片段分析表明,5 端 - 30 ~ +39 区域以及 3 端 +30 ~ -39 区域对表达水平具有重要影响,并运用距离判别分析方法构建了低水平表达函数 LES (low expression based on secondary structure) 和高水平表达函数 HES (high expression based on secondary structure)。其中,  $G_5$  为 5 端 - 30 ~ +39 区域的自由能,  $G_3$  为 3 端 +30 ~ -39 区域的自由能。对一个新的外源基因来说,首先计算 5 端与 3 端二级结构自由能  $G_5$  与  $G_3$ , 然后计算 LES 和 HES 的值,如果  $LES < HES$ , 判断外源基因为高表达;如果  $LES > HES$ , 判断外源基因为低表达。

$$LES = -10.80356 - 0.47319 \times G_5 - 1.86489 \times G_3$$

$$HES = -17.19699 + 0.15585 \times G_5 - 2.62142 \times G_3$$

### 1.2 5 端与 3 端自由能范围对表达水平的影响

进一步分析得出,为了获得外源基因的高效表达,上述区域自由能不仅要满足  $HES > LES$ , 通过判别分析研究,  $G_5$  和  $G_3$  还要求尽可能满足下列条件:

$$G_5 < -16.7 \text{ kJ/mol}$$

$$G_3 < -72.01 \text{ kJ/mol} \quad G_3 < -47.61 \text{ kJ/mol}$$

### 1.3 局部密码子偏性对表达水平的影响

通过分析密码子偏性对外源基因表达高低的影响,发现外源基因 3 端包含 TAA 在内的 3 个密码子对表达水平影响最大。因此,要求外源基因 3 端的 3 个密码子为大肠杆菌的优势密码。另外分析表明,5 端密码子对表达水平无显著影响。

## 2 程序的设计

基于上述设计规则,我们编写了 pBV220 载体外源基因高效表达评价与自动设计软件。

### 2.1 数据的选取

首先,选取与外源基因高低表达密切相关的 5 端与 3 端的两个区间,5 端区间的选择范围为 -30 ~ +39 区域(即 ATG 之前 30 bp 和编码区 39 bp 构成的片段)和 3 端 +30 ~ -39 区域(即终止密码子

TAA 前 30 bp 与 TAA 后 39 bp 组成的片段),如 Fig. 1 所示,这样,我们就得到了两条长度为 69 nt 的片段。

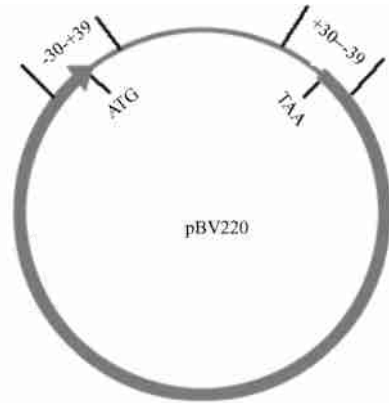


Fig. 1 Demonstration to extract two sequence fragments

When an exogenous gene was inserted into the multicloning sites in pBV220 vector, two sequence fragments, -30 ~ +39 around the translation initiation codon ATG and +30 ~ -39 around the translation termination codon TAA, were extracted. Then the secondary structure of the two sequences was predicted using the prediction methods based on helical regions random stacking presented previously by Li *et al.*<sup>[1]</sup>. Finally, the free energy was used to determine whether the inserted gene could be highly expressed or not.

### 2.2 自由能的计算

采用我们自己提出的基于螺旋区随机堆积的 RNA 二级结构预测算法计算外源基因 5 端与 3 端自由能<sup>[1]</sup>。在预测过程中,通过螺旋区随机堆积生成一系列的二级结构,找出其中出现频率较大的结构,并取该结构的自由能作为后续计算的能量值。

### 2.3 优势密码子的挑选与密码子偏性计算

针对 *E. coli* 表达系统,挑选出了相应的优势密码子,并且采用自己编写的密码子偏性分析模块计算其密码子偏性。

### 2.4 序列改造

如果外源基因 5 端和 3 端序列经过评估,不能满足高效表达的理论要求,则根据同义密码子替换原则,将外源基因 5 端和 3 端序列进行密码子替换。显然,在密码子同义替换的同时,相应的局部 RNA 二级结构和自由能均发生变化,该过程重复进行,直至最后找出满足外源基因高效表达条件的序列为止。

## 3 程序的使用

程序界面如图 2 所示,分为 4 部分:Sequence、Evaluation、Design 和 Settings,分别用于输入序列、评估结果显示、设计结果显示以及参数设置。

### 3.1 参数设置

能量参数设置:为了提高预测精确度,程序默认设置为 3 端能量介于 - 72.01 kJ/mol G3 - 52.30 kJ/mol,5 端能量大于 - 16.7 kJ/mol,可以根据实际需要修改.

结果数目设置:用于设置输出多少个设计结果.

超时设置:对于某些外源基因,可能较难设计出高效表达序列,因此,设置该选项用于超时退出.

### 3.2 评估功能

程序提供了对现有序列的评估功能,以检验是否能满足高效表达的要求.首先,我们将 5 端 - 30 ~

+ 39 区域这一片段(即 ATG 之前 30 bp 和编码区 39 bp 构成的片段)放入 Fig. 2 中标示有 5' End 的文本框中,同理,将 3 端 + 30 ~ - 39 区域这一片段(即终止密码子 TAA 前 30 bp 与 TAA 后 39 bp 组成的片段)放入标示有 3' End 的文本框中,点击 Evaluation 按钮后,程序会自动检测序列长度是否符合要求,是否包含了 4 种碱基 A、T、G、C 以外的字符以及相应位置是否是 ATG 与 TAA,如果有错误,会给出相应提示.若无错误,程序会按照预设参数,逐条检验各项条件,并给出评估结果.



Fig. 2 The interface for evaluation and design for exogenous genes To evaluate whether a exogenous gene could be highly expressed or not, the sequence fragments from exogenous gene 5' end and 3' end were firstly input into the sequence window. Then, the evaluation results would be displayed in the evaluation window by clicking the evaluation button. Obviously, the example given here did not meet the condition for high-level expression. Thus, to obtain the high-level expression of this gene, the coding region of the first 39 nucleotides and the last 30 nucleotides of the gene would be modified by the replacement of synonymous codons. For each modification, the related sequences were evaluated by the condition provided by the Settings panel. The design results would be displayed in the design window

### 3.3 设计功能

对于设计功能来说,前序准备工作与评估功能

类似,即将 5 端与 3 端片段放入相应的文本框中.对于评估结果不满足高效表达条件的序列,可以点击

Design 按钮,对其进行改造,程序会依据设置的结果数目,将设计结果显示于文本框中.

### 3.4 使用范例

假定某个外源基因插入 pBV220 载体后,其 5 端 -30 ~ +39 区域的序列为 AAGCATTGGTTAAAAAT-TAAGGAGGAATTCATGTCATCA TCCCATCCCATCTT-CCACAGGGGCGAATTC,其 3 端 +30 ~ -39 区域的序列为 TGTGTGCTCAGCAGGAA GCCTGTGAGATA-AAGGATCCGTCGACCTGCA GCCAAGCTTGGCTGTTTTGGTT,该外源基因是否适合高效表达的评价和设计步骤如下:

(1) 将上述两序列复制粘贴至 Sequence 一栏中标示有 5 End 及 3 End 的文本框中.

(2) 对 setting 一栏中的相关参数进行设置,这里,我们采用默认参数.

(3) 首先,我们先对其进行评估,看是否满足高效表达的要求. 点击 Evaluation 按钮后,在 Evaluation 一栏中会从 5 端能量、3 端能量、密码子偏性等方面给出评估结果,对于本示例来说,是不满足高效表达要求的.

(4) 对于本示例来说,评估结果表明不满足高效表达要求,因此,可以点击 Design 按钮对其进行改造. 点击 Design 按钮后,等待一段时间,在 Design 一栏中就会显示自动设计后的结果,即 5 和 3 端改造后的序列. 依照此结果来对序列进行改造,就有可能提高表达水平.

## 4 总结

本文介绍了我们开发的 pBV220 载体中外源基因高效表达评价与自动化设计系统,其核心是 pBV220 载体中外源基因高效表达数学模型,结合我们最近提出的酵母系统 pPIC9 载体中外源基因高效表达数学模型<sup>[9]</sup>,我们认为,实现外源基因高效表达的定性条件是外源基因的 5 端要求有不稳定的二级结构,而外源基因 3 端需要有稳定的二级结构. 因此,这里介绍的软件不仅可以定量用于 pBV220 载体,也可以定性用于原核系统中的其它表达载体,最终为实验人员提供了一个快速有力的工具,使得凭

实验经验的外源基因表达设计成为一个理性化的过程,为加快实验进程提供了有力帮助.

## 参考文献 (References)

- [1] 李伍举,吴加金. 基于螺旋区随机堆积的 RNA 二级结构预测[J]. 生物物理学报(Li Wu-Ju, Wu Jia-Jin. Prediction of RNA secondary structure based on random stacking of helical regions[J]. Acta Biophys Sin), 1996, 12(2): 213-218
- [2] 李伍举,吴加金. pBV220 载体中外源基因表达水平定量分析[J]. 病毒学报(Li Wu-Ju, Wu Jia-Jin. Quantitative analysis on expression level of foreign gene in pBV220 vector[J]. Chin J Virol), 1997, 2(13): 126-133
- [3] 裴武红,沈倍奋,李伍举,等. 计算机辅助设计使重组 Ricin-A 链在 E. coli 中的高效表达[J]. 细胞与分子免疫学杂志(Pei Wu-Hong, Sheng Bei-Fen, Li Wu-Ju. Computer-aided design in high-expression of recombinant ricin-A chain in E. coli [J]. J Cell Mol Immunol), 1998, 14(1): 33-36
- [4] 裴武红,胡美茹,李伍举,等. 人 FKBP12 基因克隆及其表达产物的生物活性[J]. 中国生物化学与分子生物学报(Pei Wu-Hong, Hu Mei-Ru, Li Wu-Ju, et al. The gene cloning and bioactivity of the expression product of the human FKBP12 [J]. Chin J Biochem Mol Biol), 2000, 16(3): 322-325
- [5] 刘淑红,孙长凯,张玉梅,等. 人 NR1 靶片段的原核表达[J]. 军事医学科学院院刊(Liu Shu-Hong, Sun Chang-Kai, Zhang Yu-Mei, et al. Expression of the cDNA encoding a target fragment of human NT1 protein in E. coli [J]. Bull Acad Mil Med Sci), 2002, 26(3): 188-190
- [6] Li Wu-Ju, Lei Hong-Xing, Pei Wu-Hong, et al. GeneDn: for high-level expression design of heterologous genes in a prokaryotic system [J]. Bioinformatics, 1998, 14(10): 884-885
- [7] 李伍举,应晓敏. BioSun: 计算机辅助分子生物学实验设计的软件系统 [J]. 军事医学科学院院刊(Li Wu-Ju, Ying Xiao-Min. BioSun: a software system for computer-aided design for molecular biology experiments [J]. Bull Acad Mil Med Sci), 2004, 10(5): 401-404
- [8] 查磊,应晓敏,曹源,等. BioSun2.0: 一个综合性的辅助分子生物学实验设计软件 [J]. 军事医学科学院院刊(Cha Lei, Ying Xiao-Min, Cao Yuan, et al. BioSun2.0: a comprehensive software for molecular biology experiments [J]. Bull Acad Mil Med Sci), 2006, 30(5): 461-464
- [9] Wu Bingli, Cha Lei, Du Zepeng, et al. Construction of mathematical model for high-level expression of foreign genes in pPIC9 vector and its verification [J]. Biochem Biophys Res Commun, 2007, 354(2): 498-504