BIOINFORMATICS APPLICATIONS NOTE Vol. 14 no. 10 1998
Pages 884–885

GeneDn: for high-level expression design of heterologous genes in a prokaryotic system

Li Wu Ju, Lei Hong Xing, Pei Wu Hong and Wu Jia Jin

Laboratory of Bioinformation Engineering, Institute of Basic Medical Sciences, P.O. Box 130 (3), Beijing 100850, China

Received on April 23, 1998; revised on September 10, 1998; accepted on October 13, 1998

Abstract

Results: *Based on the mathematical model of high-level expression of heterologous genes in prokaryotic vector pBV220, we developed a program GeneDn for high-level expression design of natural and synthetic genes.*
Availability: *The program is written in Turbo Pascal 7.0. The source code and related material are available upon request.*
Contact: *wujj@nic.bmi.ac.cn*

Introduction

Due to its many advantages, *Escherichia coli* is still a valuable organism for the high-level production of recombinant proteins. However, in spite of the extensive knowledge of the genetics and molecular biology of *E. coli*, not every gene can be expressed efficiently in this organism (Makrides, 1996). Among the factors regulating gene expression, the secondary structure of translation initiation region, the promotors, the Shine–Dalgarno sequence, the codon usage, and the number of bases between SD sequence and initial codon ATG are widely investigated. However, all these conclusions are qualitative and analyzed separately. In order to investigate the comprehensive effects of these factors, a mathematical model of high-level expression of heterologous genes was constructed (Li and Wu, 1997).

Algorithms and implementation

In order to construct the mathematical model (Li and Wu, 1997), we collected relevant data from 22 heterologous genes expressed in pBV220 vector and classified them into two groups. If the expression level of a heterologous gene is greater than or equal to 20% of the total cellular protein, it is called high-level expression; otherwise, it is called low-level expression. In 22 heterologous genes, there are 13 genes belonging to the high-level expression class and nine genes belonging to the low-level class. Based on this classification, we obtained the following comprehensive equations

$$LESC = -13.64013 + 12.85459 \times X - 0.36121 \times G_5 - 2.03769 \times G_3 \quad (1)$$

$$HESC = -21.82053 + 16.42926 \times X + 0.29896 \times G_5 - 2.84227 \times G_3 \quad (2)$$

where X is a CAI value for the 18bp of 5′ end of the heterologous gene, and G_5 and G_3 are the free energy of the secondary structure of the regions –30 ~ 39 and 30 ~ –39 of 5′ and 3′ ends respectively (5′ region –30 ~ 39 stands for the 30bp in the vector just before initial codon ATG and 39bp in heterologous gene including initial codon ATG. 3′ region 30~ –39 for the 30bp in the heterologous gene including TAA and 39bp in the vector just after terminal codon TAA). If the heterologous gene inserted into the pBV220 vector satisfies the following conditions: the number of bases between the Shine–Dalgarno sequence and the ATG is from 5 to 11, $G_5 \geq -4.0$ Kcal/Mol, $G_3 \leq -11.4$ Kcal/Mol (i.e. the goal is to have a weak secondary structure in the 5′ end and a strong structure in the 3′ end), $LESC < HESC$ and the three codons in 3′ end (including TAA) are codons with high RSCU value (RSCU:relative synonymous codon usage, Sharp and Li, 1987), then high-level expression can be obtained and the expression level is more than 20% of the total cellular protein. The correct rate of backward analysis for selected 22 heterologous genes is 95.5% (i.e. 21/22). In addition, four other heterologous genes expressed in pBV220 vector have also been correctly classified.

Based on the above model and the following data and methods: REBASE (Roberts and Macelis, 1997), the sequence of pBV220 vector (Zhang *et al*., 1990), codon adaptation index (CAI, Sharp and Li, 1987), Turner free energy's rule (Zuker, 1989), the method for ambiguous match (Raghava and Sahni, 1994) and the prediction of RNA secondary structure based on the random stacking of helical regions with discriminate level α=0.01 (Li and Wu, 1996, 1998), the program GeneDn for high-level expression of heterologous genes including natural and synthetic genes was developed which mainly contains the following procedures.

At first, GeneDn reads the protein sequence into the computer memory in text file or SWISS-PROT name. The output of the GeneDn is an exact DNA sequence retranslated from the protein sequence which is ensured to accommodate as many unique restriction sites for 'modular mutagenesis' and codons with high RSCU value as possible. Then, two restriction endonucleases not found in the exact DNA sequence from the multiclone sites of pBV220 vector are selected as

the 5′ and 3′ end endonuclease sites of heterologous gene and a recombinant plasmid was constructed. The secondary structures of the regions –30 ~ 39 and 30 ~ –39 of 5′ and 3′ ends are considered and their free energy is indicated by G_5 and G_3. At last, the codon adaptation index X of 18bp of the 5′ end is calculated. If G_5, G_3 and X satisfy the condition of high-level expression of heterologous gene, the process of gene design is completed. Otherwise, automatic and manual methods including the replacement of synonymous codons or point mutation are used to adjust the secondary structure of 5′ and 3′ ends.

In order to demonstrate the ability of the program GeneDn, the design of Human CD28 (SWISS-PROT CODE: CD28_HUMAN) is computed. Here we specially introduce the design of high level expression of ricin A chain.

When the ricin A chain was inserted into the pBV220 vector directly between EcoRI and BamHI sites, we observed no expressed protein on SDS-PAGE compared to the marker. When we analyzed the 5′ and 3′ RNA secondary structure with GeneDn, we obtained the data (G_5=–4.4 Kcal/Mol, G_3=–11.2 Kcal/Mol and X=0.2638) which is consistent with the low-level expression. To obtain high-level expression, we adjusted the secondary structure of the 5′ and 3′ ends of the ricin A chain by designing two PCR primers based on the model and obtained the data (G_5=–0.3 Kcal/Mol, G_3=–13.0 Kcal/Mol and X=0.2638). After a simple calculation, we obtain *LESC*<*HESC*. Therefore according to the model, the ricin A chain can be highly expressed with expression level more than 20% of total cellular protein. In fact, we have verified this and the expression level is 23% of the total cellular protein (Pei *et al*., 1998).

References

Li,W.J. and Wu,J.J. (1996) Prediction of RNA secondary structure based on helical regions random stacking. *Acta Biophysa Sinica*, **12,** 213–218.

Li,W.J. and Wu,J.J. (1997) Quantitative analysis on expression level of foreign gene in pBV220 vector. *Chin. J. Virol*., **13,** 126–133.

Li,W.J. and Wu,J.J. (1998) Prediction of RNA secondary structure based on helical regions distribution. *Bioinformatics*, **14,** 700–706.

Makrides,S.C. (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev*., **60,** 512–538.

Pei,W.H., Sheng,B.F. and Li,W.J. (1998) Computer-aided design in high-expression of recombinant ricin-A chain in *E. coli*. *J. Cell. Molec. Immunol*., **14,** 33–36.

Raghava,G.P.S. and Sahni (1994) GMAP: a multi-purpose computer program to aid synthetic gene design, cassette mutagenesis and the introduction of potential restriction sites into DNA sequences. *BioTechniques*, **16,** 1116–1123.

Roberts,R.J. and Macelis,D. (1997) REBASE-restriction enzymes and methylases. *Nucleic Acids Res*., **25,** 248–262.

Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index — a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*., **15,** 1281–1295.

Zhang,Z.Q., Yao,L.H. and Hou,Y.D. (1990) Construction and application of a high level expression vector containing P_RP_L promotor. *Chin. J. Virol*., **6,** 111–116.

Zuker,M. (1989) Computer prediction of RNA structure. *Meth. Enzymol*., **180,** 262–288.