Available online at www.sciencedirect.com
SCIENCE DIRECT®
Protein Expression and Purification 47 (2006) 441–445
Protein Expression & Purification
www.elsevier.com/locate/yprep

The Synthetic Gene Designer: A flexible web platform to explore sequence manipulation for heterologous expression

Gang Wu *, Nabila Bashir-Bello, Stephen J. Freeland

Department of Biological Sciences, University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

Received 30 September 2005, and in revised form 13 October 2005
Available online 15 November 2005

Abstract

"Codon optimization" is a general approach to improving heterologous expression where genes are moved from their native genomes into alternatives that exhibit different patterns of codon usage. However, despite reports of successful manipulations and the existence of stand-alone codon optimization software packages or commercial services that offer to redesign genes, the scientific community lacks any systematic understanding of what exactly it means to optimize codon usage. Thus we present a bona fide web application, the "Synthetic Gene Designer," which contrasts with existing software by providing a centralized, free, and transparent platform for the broader scientific community to develop knowledge about synthetic gene design. Consistent with this goal, our software is associated with a moderated e-forum that promotes discussion of synthetic gene design and offers technical support. In addition, the Synthetic Gene Designer presents enhanced functionality over existing software options: for example, it enables users to work with non-standard genetic codes, with user-defined patterns of codon usage and an expanded range of methods for codon optimization. The Synthetic Gene Designer, together with on-line tutorials and the forum, is available at http://www.evolvingcode.net/codon/sgd/index.php.

Keywords: Codon optimization; Heterologous expression; Software; Web application

Different patterns of codon usage found in the genomes of different species are widely recognized as a possible cause of low protein yields during heterologous protein expression [1–3]. Therefore, many studies have manipulated codon usage of a coding sequence in an attempt to increase translational efficiency (reviewed in [4]). Some have successfully improved protein expression (e.g. [5–9]) but others have failed (e.g. [10,11]), suggesting that the concept of 'codon optimization' is not a trivial one. Indeed, it is tenuous to draw any detailed or generalized conclusions from the successful studies, since each protein sequence (any protein sequence) corresponds to a huge 'possibility space' of different nucleotide sequences, and current studies have only ever reported tests pertaining to one or a few versions of redesigned genes [4]. In this context, a valuable next-step in developing the science of gene design will come from consolidating, centralizing, and rendering freely accessible knowledge surrounding different re-coding strategies and their relative success or failure.

In particular, there are two widely used codon optimization methods: full optimization (i.e., optimization of every codon [12] or 'one amino acid–one codon' approach [4]) and selective optimization (i.e., only replace so-called "rare" codons [5]). In the light of these two methods, some computer programs have been developed recently to automate the codon optimization [13–16]. However, there is still no clear consensus on: (1) what is an appropriate reference template of codon usage; (2) what is the best algorithm for codon optimization; (3) the extent to which "optimal" codons should be used. For example, the full optimization method has been criticized by Gustafsson et al. for: (1) the potential for translational error because of an imbalanced tRNA pool [17]; (2) introduction of repetitive elements and mRNA secondary structures that might inhibit ribosome processivity; (3) lack of flexibility for introducing restriction sites [4]. Furthermore, use of non-optimal codons at some positions might be required for the correct folding of nascent translated poly-

* Corresponding author. Fax: +1 410 455 3875.
E-mail address: wug1@umbc.edu (G. Wu).

1046-5928/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.pep.2005.10.020

peptides [18]. Indeed, there is no any natural highly expressed gene that uses optimal codons for every amino acid, suggesting that full codon optimization may be more than unnecessary (instead, it may be detrimental) for high protein expression. Thus existing software that uses a fixed model for codon optimization can only partially address current needs [13–16]. To facilitate systematic and coordinated investigation of these questions, we have developed a web application: the "Synthetic Gene Designer," which facilitates flexible redesign of genes for heterologous protein expression.

Algorithms

The Synthetic Gene Designer is written in PHP, JavaScript, and Perl. It has been extensively tested with various web browsers (e.g., IE5.0, Firefox 1.0 or Netscape 4.7 and higher versions) under different operating systems (Windows, Macintosh OS, and Linux). Given a gene of interest and a target genome in which it is to be expressed, the Synthetic Gene Designer comprises three major phases for gene design (Fig. 1). Detailed description of the computation methods is located at http://www.evolvingcode.net/codon/sgd/methods.php.

The functional core of our software allows flexible modifications of codon usage according to the two, currently popular methods ("full optimization" and "selective optimization") and a third, novel model that we describe here for the first time: "probabilistic optimization." This third model is designed to allow researchers to explore re-coding strategies that lie between the two extremes of "full optimization" and "selective optimization" (=rare codon replacement): under "probabilistic optimization," synonymous codons are used in the re-coded gene proportional to their observed frequency in a user-provided reference set of genes. The precise relationship between a codon's frequency in reference data and its frequency in the re-coded gene is controlled by a scaling factor ('s,' or "optimality factor"). For a codon family which has n synonymous codons, the frequency of each codon in the reference template is X_i, $i = 1, \ldots, n$ (if X_i equals to 0, it will be arbitrarily replaced with 0.5 as suggested in Sharp and Li [19]). We defined the relative frequency of codon i (F_i) in the reference template as

$$F_i = X_i \Big/ \sum_{i=1}^{n} X_i. \quad (1)$$

The upper boundary of the probability of using codon i (UB_i) is defined as

$$\mathrm{UB}_i = \left(\sum_{j=0}^{i} F_j \right)^s, \quad (2)$$

where s is a non-negative number and $\mathrm{UB}_0 = 0$. The probability of using codon i (P_i) is

$$P_i = \mathrm{UB}_i - \mathrm{UB}_{i-1}. \quad (3)$$

During the gene redesign, the program scans through the gene of interest codon by codon, and then it chooses a synonymous codon according to P_i calculated as Eq. (3). In short, each codon is used proportional to its frequency in the reference data set, weighted by the scaling factor.

Once a gene has been re-coded, our software measures the extent to which codon usage has been optimized using the well-established Codon Adaptation Index (CAI[1] [19]): previous analysis reveals that this metric correlates better with gene expression level than other codon bias measurements [20].

Results and discussion

Effects of the optimality factor on the choice of synonymous codons

The probability of selecting each synonymous codon during codon optimization is affected by the optimality factor (Fig. 2). This scaling factor serves as a convenient tuner to control the overall optimality of codon usage for the re-coded genes. For example, three special types of codon adjustments can be realized by simply changing s value:

(1) When s equals to 0 (e.g., $s = 0$ in Fig. 2), only the most frequently used codon ("optimal codon") will be used. This setting, then, amounts to "full optimization."
(2) When s equals to a large number (e.g., $s = 64$ in (Fig. 2), the least frequently used codons will be selected most often for every amino acid. We call this process as "anti-optimization."
(3) When codon numbers in the reference template are equal and s equals 1, each codon is equally weighted in a codon family. Thus, codons in the generated sequence are randomly selected.

Though results of gene re-coding for options 2 and 3 have not been reported to date, investigating the effect of various degrees of anti-optimization, and of randomized patterns (controls, in effect), will be crucial to developing a robust theoretical model of codon-mediated translation.

In recognizing how far current science lies from a comprehensive understanding of the interaction between codon usage and heterologous expression, we have enhanced the Synthetic Gene Designer with further flexibility that will allow users to investigate new ideas. In particular, the software allows manual editing for selective optimization and automatic avoidance of unwanted patterns such as high G/C repeats and restriction sites.

Unique and important features of the Synthetic Gene Design

Beyond our new, generalized algorithm for codon optimization, the Synthetic Gene Designer offers further

[1] *Abbreviations used:* CAI, Codon Adaptation Index.