Sequence analysis Advance Access publication April 10, 2014

Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design

Ju Xin Chin[1], Bevan Kai-Sheng Chung[1] and Dong-Yup Lee[1,2,*]

[1]Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), Singapore 138668 and [2]Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Codon optimization has been widely used for designing synthetic genes to improve their expression in heterologous host organisms. However, most of the existing codon optimization tools consider a single design criterion and/or implement a rather rigid user interface to yield only one optimal sequence, which may not be the best solution. Hence, we have developed Codon Optimization OnLine (COOL), which is the first web tool that provides the multi-objective codon optimization functionality to aid systematic synthetic gene design. COOL supports a simple and flexible interface for customizing various codon optimization parameters such as codon adaptation index, individual codon usage and codon pairing. In addition, users can visualize and compare the optimal synthetic sequences with respect to various fitness measures. User-defined DNA sequences can also be compared against the COOL optimized sequences to show the extent by which the user's sequences can be further improved.

Availability and implementation: COOL is free to academic and non-commercial users and licensed to others for a fee by the National University of Singapore. Accessible at http://bioinfo.bti.a-star.edu.sg/COOL/

Contact: cheld@nus.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online

Received on December 8, 2013; revised on March 2, 2014; accepted on April 7, 2014

1 INTRODUCTION

The advent of synthetic biology ushered in the new era of life science research where synthetic genetic circuits (e.g. artificial biomolecules, usually nucleotides) can be constructed and inserted into the living host to perform novel biological functions such as a metabolic pathway, a specific cell signalling response and/or novel gene regulation (Benner and Sismour, 2005; Khalil and Collins, 2010). However, the desired biological functions can be achieved only if the synthetic genes are properly expressed within the host organism. Therefore, it is highly required to develop a systematic framework that rationally designs the synthetic genes for optimal expression in the recombinant host.

Factors affecting the expression of recombinant genes have been well studied. The codon usage pattern, which relates to relative abundance of tRNA isoacceptors, has been implicated as one of the crucial determinants of heterologous protein expression level (Gustafsson *et al*., 2004). However, experimental studies have shown that the precise factors and their relative contributions are unclear (Welch *et al*., 2009). Computational tools, such as JCat (Grote *et al*., 2005), Synthetic Gene Designer (Wu *et al*., 2006) and OPTIMIZER (Puigbo *et al*., 2007), have been developed to quantify and optimize the codon usage frequency of the coding sequence in terms of the host's codon adaptation index (CAI) or individual codon usage (ICU). Moreover, the optimization of codon pairing, also known as codon context (CC), has been demonstrated to be effective in improving heterologous gene expression (Chung *et al*., 2013; Hatfield and Roth, 2007; Moura *et al*., 2011). Additionally, the presence of hidden stop codons (HSC) also prevents off-frame reading, thereby reducing resource wastage, and maximizing the HSC count might increase gene expression (Seligmann and Pollock, 2004). Thus, several design factors should be simultaneously considered for better synthetic gene design. To address such a need for incorporating multiple design parameters including CAI, ICU, CC and HSC in codon optimization, we have developed a new web server application, Codon Optimization OnLine (COOL), using a multi-objective framework, which has been presented in a recent work (Chung and Lee, 2012).

2 FUNCTIONAL FEATURES

COOL is designed as an adaptable web-based interface that provides an expansive collection of features. Users can fully customize the synthetic gene design process through a step-by-step job submission process that allows them to specify their preferred parameter settings. More notably, COOL supports a wide range of visualization capabilities for comparing the quality of optimized DNA coding sequence with respect to the various design parameters.

2.1 Gene design parameter settings

Similar to existing online applications, COOL can perform the optimization of a coding sequence based on CAI, which was known to correlate well with gene expressivity (Sharp and Li, 1987). Additionally, COOL is the first web server that uses a multi-objective framework that incorporates ICU, CC, CAI, HSC and GC content. For each codon optimization job, users can select any desired combination of design parameters. Notably, COOL enables greater flexibility through the customization of these design parameters, a feature not found in other

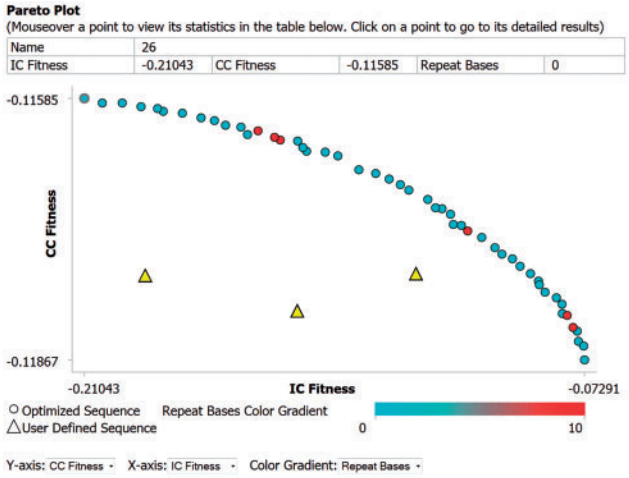
[*]To whom correspondence should be addressed.

codon optimization tools. For a full comparison with other online tools, refer to Supplementary Table S1.

Because changes in metabolic resources can significantly affect gene expression (Dittmar *et al*., 2005; Wohlgemuth *et al*., 2013), the ability to specify CAI, ICU and/or CC usage distribution is accommodated in COOL, where users either select a host from the predefined list, which includes *Escherichia coli*, *Pichia pastoris* and *Saccharomyces cerevisiae*, or fully customize the codon usage rules. When a predefined expression host is chosen, users have the option to select the relevant genes for calculating the reference codon usage patterns that correspond to efficient expression. Thus, this functionality enables users to customize their own list of high-expression genes based on expression data. Furthermore, some motifs, such as restriction enzyme cleavage sites, patented sequences and nucleotide repeats, may not be desirable in the coding sequence, as they pose difficulties in either the preparation of plasmid or the *in vivo* expression of the proteins. Such sequences can be specified for exclusion during job submission.

2.2 Visualization and export of results

When more than one optimization criteria are selected, the multi-objective optimization algorithm generates a number of Pareto-optimal sequences. Although an earlier stand-alone tool EuGene (Gaspar *et al*., 2012) also claimed to perform this function, it did not have the functionality to generate the Pareto-optimal sequences and visualize them on a plot. Therefore, COOL has addressed this limitation by providing a visualization and comparative analysis interface (Fig. 1). The relative positions of the Pareto-optimal sequences on the graph show their fitness levels on the basis of the corresponding optimization criteria. In the case where three or more optimization criteria are selected, the plot displays a third fitness value in the form of the colour of the points using a cyan–red colour gradient. Users can also customize the plot by choosing any optimization criteria to be displayed on the axes and the colour gradient.



Pareto Plot

(Mouseover a point to view its statistics in the table below. Click on a point to go to its detailed results)

Name	26				
IC Fitness	-0.21043	CC Fitness	-0.11585	Repeat Bases	0

Y-axis: CC Fitness · X-axis: IC Fitness · Color Gradient: Repeat Bases ·

Fig. 1. Snapshot of the Pareto plot. In this example, the human insulin sequence was optimized with respect to ICU and CC. User-defined sequences (triangles) can be added and compared against the optimized sequences (circles) on the Pareto front

As a unique feature, user-defined sequence can be mapped into the Pareto plot for comparison (Fig. 1). Thus, the relative positions of the user-defined sequence(s) on the plot can indicate the similarities or differences with the Pareto-optimal sequences so that the users are guided to enhance the sequences further. The Pareto plot is interactive and users can click on any of the optimized sequences to obtain detailed information of the sequence, which includes visualization of the coding sequence and a table of summary statistics showing its fitness values. These results can also be easily exported in the format of PDF files. The optimized sequences may also be exported to a tab-delimited file containing the fitness values for all optimization criteria. This allows users to plot and analyse the sequences using other software such as Microsoft Excel.

3 IMPLEMENTATION

This website was built using the open source LAMP solution stack (Linux, Apache, MySQL, PHP) based on a Model-View-Controller architecture. The source code for the LAMP stack can be downloaded from the COOL website. Although additional interface and visualization features (e.g. sortable tables) were added in JavaScript, the website is still functional even when JavaScript is disabled. The backend gene optimization, which uses a multi-objective genetic algorithm as described in our earlier work (Chung and Lee, 2012), is implemented in Perl and functions independently of the main website, interfacing only with the MySQL database to read inputs and write results. The waiting time for optimizing a 500-amino acid sequence is estimated to be ~20 m. We are currently patenting the backend code and will only release it in binary form, available on the COOL website.

Funding: This work was supported by the National University of Singapore, Biomedical Research Council of A*STAR (Agency for Science, Technology and Research), Singapore, and a grant from the Next-Generation BioGreen 21 Program (SSAC, No. PJ009520), Rural Development Administration, Republic of Korea.

Conflict of Interest: The authors are listed as inventor or co-inventor on a backend code patent, which is pending approval.

REFERENCES

Benner,S.A. and Sismour,A.M. (2005) Synthetic biology. *Nat. Rev. Genet*., **6**, 533–543.

Chung,B.K. and Lee,D.Y. (2012) Computational codon optimization of synthetic gene for protein expression. *BMC Syst. Biol*., **6**, 134.

Chung,B.K. *et al*. (2013) Enhanced expression of codon optimzed interferon gamma in CHO cells. *J. Biotechnol*., **167**, 326–333.

Dittmar,K.A. *et al*. (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep*., **6**, 151–157.

Gaspar,P. *et al*. (2012) EuGene: maximizing synthetic gene design for heterologous expression. *Bioinformatics*, **28**, 2683–2684.

Grote,A. *et al*. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res*., **33**, W526–W531.

Gustafsson,C. *et al*. (2004) Codon bias and heterologous protein expression. *Trends Biotechnol*., **22**, 346–353.

Hatfield,G.W. and Roth,D.A. (2007) Optimizing scaleup yield for protein production: computationally optimized DNA assembly (CODA) and translation engineering. *Biotechnol. Annu. Rev.*, **13**, 27–42.

Khalil,A.S. and Collins,J.J. (2010) Synthetic biology: applications come of age. *Nat. Rev. Genet.*, **11**, 367–379.

Moura,G.R. *et al.* (2011) Species-specific codon context rules unveil non-neutrality effects of synonymous mutations. *PLoS One*, **6**, e26817.

Puigbo,P. *et al.* (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.*, **35**, W126–W131.

Seligmann,H. and Pollock,D.D. (2004) The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.*, **23**, 701–705.

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.

Welch,M. *et al.* (2009) Design parameters to control synthetic gene expression in *Escherichia coli. PLoS One*, **4**, e7002.

Wohlgemuth,S.E. *et al.* (2013) Translational sensitivity of the *Escherichia coli* genome to fluctuating tRNA availability. *Nucleic Acids Res.*, **41**, 8021–8033.

Wu,G. *et al.* (2006) The synthetic gene designer: a flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr. Purif.*, **47**, 441–445.