International Journal of
Molecular Sciences

MDPI

Communication

Presyncodon, a Web Server for Gene Design with the Evolutionary Information of the Expression Hosts

Jian Tian [1,†], Qingbin Li [1,2,†], Xiaoyu Chu [1] and Ningfeng Wu [1,*]

[1] Biotechnology Research Institute, Chinese Academy of Agricultural sciences, Beijing 100081, China; tianjian@caas.cn (J.T.); liqingbin2015@sina.cn (Q.L.); chuxiaoyu@caas.cn (X.C.)
[2] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100081, China
* Correspondence: wuningfeng@caas.cn
† These authors contributed equally to this work.

Received: 15 November 2018; Accepted: 3 December 2018; Published: 4 December 2018

Abstract: In the natural host, most of the synonymous codons of a gene have been evolutionarily selected and related to protein expression and function. However, for the design of a new gene, most of the existing codon optimization tools select the high-frequency-usage codons and neglect the contribution of the low-frequency-usage codons (rare codons) to the expression of the target gene in the host. In this study, we developed the method Presyncodon, available in a web version, to predict the gene code from a protein sequence, using built-in evolutionary information on a specific expression host. The synonymous codon-usage pattern of a peptide was studied from three genomic datasets (*Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae*). Machine-learning models were constructed to predict a selection of synonymous codons (low- or high-frequency-usage codon) in a gene. This method could be easily and efficiently used to design new genes from protein sequences for optimal expression in three expression hosts (*E. coli*, *B. subtilis*, and *S. cerevisiae*). Presyncodon is free to academic and noncommercial users; accessible at http://www.mobioinfor.cn/presyncodon_www/index.html.

Keywords: gene design; presyncodon; expression host; codon optimization; web server

1. Introduction

In most organisms, 61 universal genetic codons encode for 20 standard amino acids, of which 18 are encoded by multiple synonymous codons. In all domains of life, a biased frequency of synonymous codons is observed at the genome level. Many studies have proved that the presence of synonymous codons in the gene coding regions is not inconsequential, and relates to the efficient and accurate translation of the protein [1–3]. Therefore, codon optimization can affect protein expression and function in the heterologous gene expression system [4–7].

Many methods, including JCat [8], Gene Designer [9], OPTIMIZER [10], Gene Composer [11], COStar [12], and COOL [13] have been proposed to design heterologous genes that are expected to be efficiently expressed in the host organism. Based on our experience, we concluded that these methods are prone to select the high-frequency-usage codons and neglect the contribution of the low-frequency-usage codons (rare codons) to the expression of the target gene. However, in the case of some genes, single point synonymous codons can also affect the expression and function of the target protein [4,14–16]; and some rare codons are conserved in the evolution and play an important role to regulate protein folding and protein production [16–18]. Therefore, those methods have an over-reliance on the prediction and usage of codons that are frequently selected in highly-expressed genes. In the natural host, most of the synonymous codons of the gene have been evolutionarily

selected and; therefore, in order to account for all the evolutionary variation, the codon usage pattern should be learned from the natural genes [16,19].

To address the need for heterologous gene design, based on all used codons (the high- or low-frequency-usage codons), a new web server application, Presyncodon, was developed to design the heterologous gene for expression in the three frequently-used recombinant hosts (*Escherichia coli*, *Bacillus subtilis*, and *Saccharomyces cerevisiae*). This big data gene prediction method was used to learn the codon-usage pattern for a peptide as-derived from sequenced genomic data. Machine-learning models were constructed by the random forest classification to predict a selection of synonymous codons (low- or high-frequency-usage codon) for the target gene. Compared with the early version of Pyesyncodon, which could only design the gene to be efficiently expressed in *E. coli* in local [20], this new version could design new genes from protein sequences for optimal expression in three recombinant hosts (*E. coli*, *B. subtilis*, and *S. cerevisiae*) on the web; and the training dataset has been updated with more genomes. Therefore, this method will be easily and efficiently used to design genes for heterologous gene expression in the three popular expression hosts (*E. coli*, *B. subtilis*, and *S. cerevisiae*).

2. Materials and Methods

2.1. Dataset

Three genomic datasets (*E. coli*, *B. subtilis*, and *S. cerevisiae*) were constructed, which contained 353, 62, and 20 genomes, respectively. All selected genomes were the complete genomes downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/) on July 13th, 2016, and their genome accession numbers are shown in Table S1.

In order to train the predicting models, three non-redundant gene datasets (*E. coli*, *B. subtilis*, and *S. cerevisiae*) were also constructed. The software CD-HIT [21] was used to calculate the gene clusters and remove the redundant genes in the cluster with protein sequences exhibiting over 40% identity. For the aim to remove the peculiar genes that might evolve from the horizontal transfer, the typical genes from those gene clusters that contained at least three homologous sequences were selected. The required length of each sequence was over 100 codons. As a result, three gene datasets, covering *B. subtilis*, *E. coli* and *S. cerevisiae*, were constructed with 8091, 11232, and 5905 genes from the total 1461067, 256246, and 107820 genes, respectively.

2.2. Workflow

The general flowchart of the method is shown in Figure 1. Firstly, each gene in the constructed gene database was split into window sizes of five and seven codons. Then a codon selection index (CSI) for each set of genomic data (five and seven residues) was determined, which represented the codon usage distribution for the middle amino acid and the average codon usage for each amino acid in the fragment.

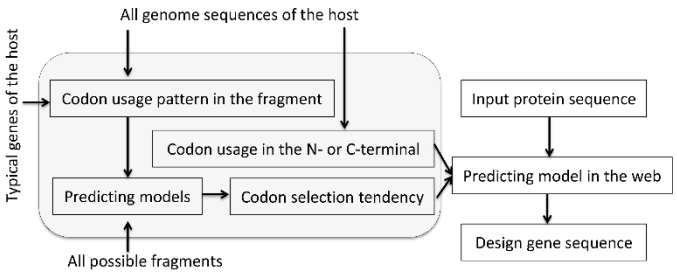

Figure 1. Flowchart summarizing the Presyncodon approach.

The training gene sequences were translated, and were also split into window sizes of five or seven amino acids, and searched against the corresponding CSI files. For each fragment, the matched score (s), expected maximal score (m) of the target fragment, and the matched percent (p, $p = s/m$)

against the CSI file were calculated by the method described in [20]. For a given cut-off level (*c*), if the calculated matched percent of multiple fragments from the CSI file for a fragment was greater than the cut-off level, the coding vector for the middle codon in the fragment was the arithmetic average of those vectors encoding the selected multiple fragments. Here, the training label is the codon for the middle amino acid.

All training labels and input vectors were collected, and the random forest classifier from "R" statistics package (ver.3.4.0) [22,23] was used to train the predicting models with seven cut-off levels (0.7, 0.75, 0.8, 0.85, 0.9, 0.95, and 1), two window sizes (5 and 7 residues), and 18 amino acids (containing multiple synonymous codons). The dimensionality of the input features for each amino acid was the codon number of the amino acid plus the window size. The number of trees of the key parameter of the classifier for random forests was set to 10,000. For each organism, 252 models ($252 = 7 \times 2 \times 18$) were constructed.

In order to increase the speed of target gene design, all possible fragments (a total of 2,880,000 (18×20^4) in case of the five residues' long fragments and of 1152000000 (18×20^5) in case of the seven residues long fragments) were searched against the corresponding CSI files, with four cut-off levels (0.7, 0.8, 0.9, and 1). Input vectors were generated for each fragment and the synonymous codons selection, based on the distribution of the middle residue in the fragment, was predicted for each organism. The results were stored in the PostgreSQL database.

As the training vector only encodes for the middle codon in the fragment, the first and last two codons of a gene could not be predicted by the above machine learning models. The codon usage pattern was generated by measuring the codon-usage bias of the first and last two residues. Therefore, the first and last two codons of a gene were designed as the most frequently used codons at these positions in all genes (Table S2).

3. Validation

The performance of the predicting models, obtained from the two fragment window sizes (5 and 7 amino acids) and cut-off level (*c*: 0.7, 0.75, 0.8, 0.85, 0.9, 0.95 and 1), were evaluated by ten-fold cross validation. As shown in Figure 2, the predicting accuracy of models obtained from the window size of seven amino acids was higher than that of the models obtained from the window size of five amino acids. Additionally, the classifier obtained with the larger cut-off level (*c*) achieved higher accuracy than those obtained with smaller cut-off levels. Therefore, the codon-usage tendency for each amino acid could be predicted by only one model, as obtained from the long-window-sized amino acid fragments and characterized by a larger cut-off level (*c*). The first and last two codons were selected statistically (Table S2).

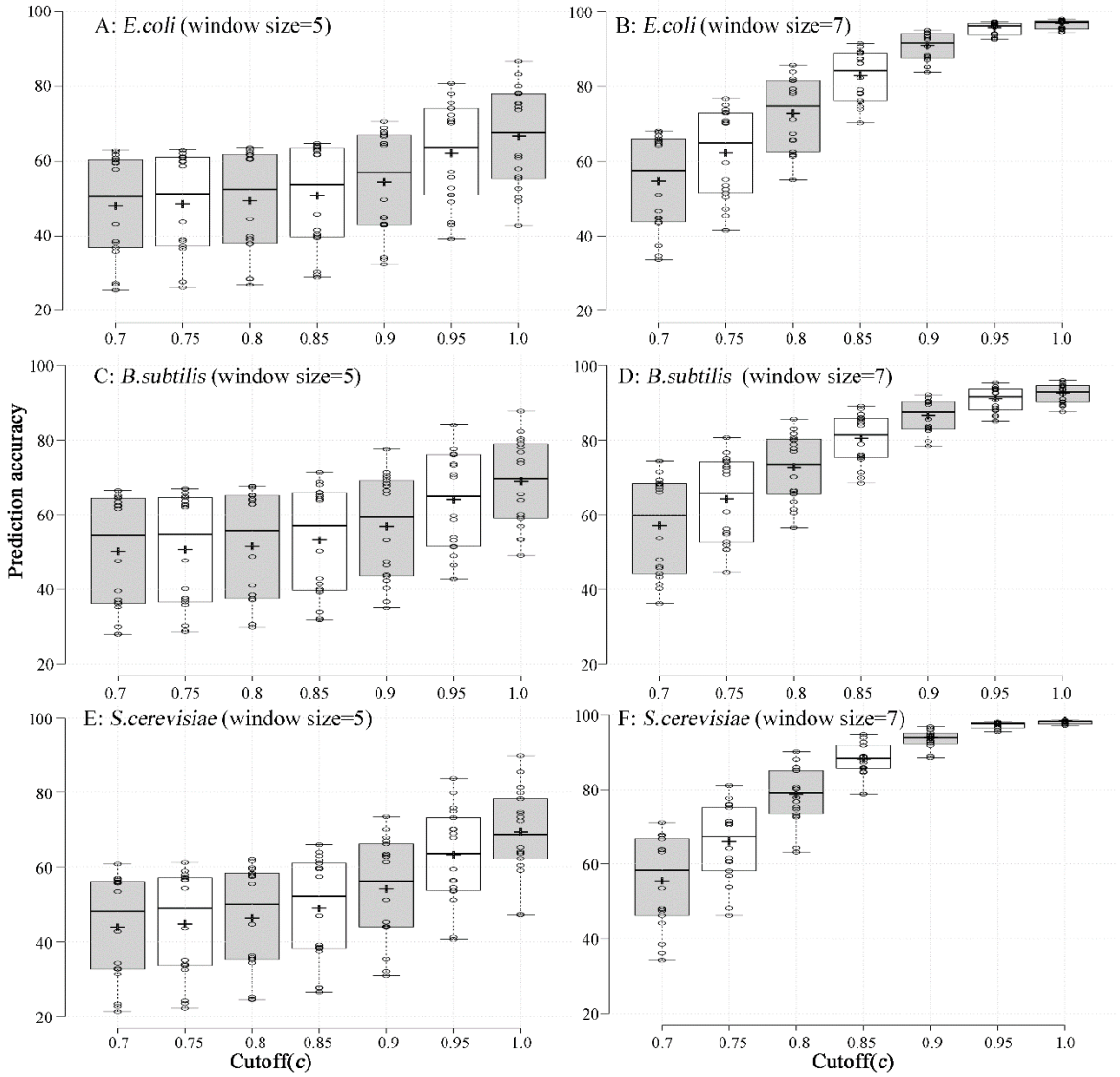

Figure 2. The prediction performance of the 18 classifiers for the 18 amino acids, with different matched cut-offs and window sizes (left: Five amino acids; right: Seven amino acids) in *E. coli*, *B. subtilis*, and *S. cerevisiae*. The *x*-axis is the matched percent and the *y*-axis is the prediction accuracy of the 18 classifiers. Each open circle represents the prediction accuracy with one of the 18 classifiers. The horizontal divisions (from top to bottom) in each box are the upper whisker, 3rd quartile, median, 1st quartile, and lower whisker, respectively. The cross line in each box is the mean prediction accuracy of all 18 classifiers. All of the results were calculated based on a ten-fold cross validation.

4. Implementation

The software Presyncodon is designed as an adaptable, web-based interface that could be easily used by scientists. This website was built using Linux (Centos ver. 6.5), Apache (ver. 2.2), PostgreSQL (ver. 8.4.20), and Perl (ver. 5.10.1). The input of the user is the target protein sequence and the only external parameter required is the selection of the target expression host (Figure 3). The waiting time for optimizing a 100-amino acid sequence is estimated to be two minutes. Therefore, the method could be easily used to design synthetic genes for heterologous gene expression in biotechnology. Based on this method, we have successfully designed the genes of GFP [20], mApple [20], laccase, penicillin-binding protein, alpha-1,4 glucan phosphorylase L-1 isozyme, pirin-like protein, and cadmium-binding proteins from maize to be efficiently expressed in *E. coli*. Now, this version of Presyncodon could be used to design the heterologous genes for expression in the three frequently-used recombinant hosts

(*E. coli*, *B. subtilis*, and *S. cerevisiae*). In the next step, we will develop this optimizing method for more expression systems. Therefore, this method could be easily used to design synthetic genes for heterologous gene expression in biotechnology.

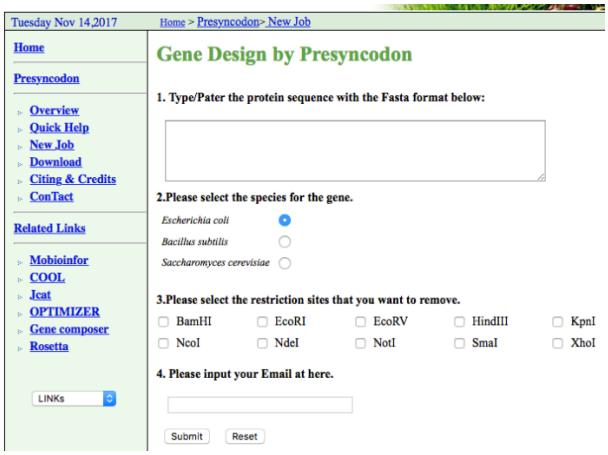

Figure 3. Screenshot of the web version of Presyncodon.

Supplementary Materials: Supplementary materials can be accessed at: http://www.mdpi.com/1422-0067/19/12/3872/s1.

Author Contributions: J.T. and Q.L. developed the server, performed the analyses, and drafted the paper. J.T. set up the server. J.T., X.C., and N.W. contributed to the data analysis and paper writing. All authors read and approved the final manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China (NSFC, Grant no. 31770124).

Acknowledgments: The numerical calculations in this study were performed with a supercomputing system on the high-performance computing platform of the Biotechnology Research Institute, Chinese Academy of Agricultural Sciences.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Cannarozzi, G.; Schraudolph, N.N.; Faty, M.; von Rohr, P.; Friberg, M.T.; Roth, A.C.; Gonnet, P.; Gonnet, G.; Barral, Y. A role for codon order in translation dynamics. *Cell* **2010**, *141*, 355–367. [CrossRef] [PubMed]
2. Gamble, C.E.; Brule, C.E.; Dean, K.M.; Fields, S.; Grayhack, E.J. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell* **2016**, *166*, 679–690. [CrossRef] [PubMed]
3. Brandis, G.; Hughes, D. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLoS Genet.* **2016**, *12*, e1005926. [CrossRef] [PubMed]
4. Brule, C.E.; Grayhack, E.J. Synonymous Codons: Choose Wisely for Expression. *Trends genet.* **2017**, *33*, 283–297. [CrossRef] [PubMed]
5. Boel, G.; Letso, R.; Neely, H.; Price, W.N.; Wong, K.H.; Su, M.; Luff, J.D.; Valecha, M.; Everett, J.K.; Acton, T.B.; et al. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* **2016**, *529*, 358–363. [CrossRef] [PubMed]
6. Goodman, D.B.; Church, G.M.; Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **2013**, *342*, 475–479. [CrossRef] [PubMed]
7. Yu, C.H.; Dang, Y.; Zhou, Z.; Wu, C.; Zhao, F.; Sachs, M.S.; Liu, Y. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* **2015**, *59*, 744–754. [CrossRef]
8. Grote, A.; Hiller, K.; Scheer, M.; Munch, R.; Nortemann, B.; Hempel, D.C.; Jahn, D. JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.* **2005**, *33*, W526–W531. [CrossRef]

9. Villalobos, A.; Ness, J.E.; Gustafsson, C.; Minshull, J.; Govindarajan, S. Gene Designer: A synthetic biology tool for constructing artificial DNA segments. *BMC Bioinform.* **2006**, *7*, 285. [CrossRef]
10. Puigbo, P.; Guzman, E.; Romeu, A.; Garcia-Vallve, S. Optimizer: A web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.* **2007**, *35*, W126–W131. [CrossRef]
11. Lorimer, D.; Raymond, A.; Walchli, J.; Mixon, M.; Barrow, A.; Wallace, E.; Grice, R.; Burgin, A.; Stewart, L. Gene composer: Database software for protein construct design, codon engineering, and gene synthesis. *BMC Biotechnol.* **2009**, *9*, 36. [CrossRef] [PubMed]
12. Liu, X.; Deng, R.; Wang, J.; Wang, X. COStar: A D-star Lite-based dynamic search algorithm for codon optimization. *J. Theor. Biol.* **2014**, *344*, 19–30. [CrossRef] [PubMed]
13. Chin, J.X.; Chung, B.K.; Lee, D.Y. Codon Optimization OnLine (COOL): A web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics* **2014**, *30*, 2210–2212. [CrossRef] [PubMed]
14. Zhou, M.; Guo, J.; Cha, J.; Chae, M.; Chen, S.; Barral, J.M.; Sachs, M.S.; Liu, Y. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **2013**, *495*, 111–115. [CrossRef] [PubMed]
15. Blazej, P.; Mackiewicz, D.; Wnetrzak, M.; Mackiewicz, P. The Impact of Selection at the Amino Acid Level on the Usage of Synonymous Codons. *G3* **2017**, *7*, 967–981. [CrossRef] [PubMed]
16. Napolitano, M.G.; Landon, M.; Gregg, C.J.; Lajoie, M.J.; Govindarajan, L.; Mosberg, J.A.; Kuznetsov, G.; Goodman, D.B.; Vargas-Rodriguez, O.; Isaacs, F.J.; et al. Emergent rules for codon choice elucidated by editing rare arginine codons in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E5588–E5597. [CrossRef] [PubMed]
17. Chaney, J.L.; Steele, A.; Carmichael, R.; Rodriguez, A.; Specht, A.T.; Ngo, K.; Li, J.; Emrich, S.; Clark, P.L. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput. Biol.* **2017**, *13*, e1005531. [CrossRef] [PubMed]
18. Jacobs, W.M.; Shakhnovich, E.I. Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 11434–11439. [CrossRef] [PubMed]
19. Zhou, Z.; Dang, Y.; Zhou, M.; Li, L.; Yu, C.H.; Fu, J.; Chen, S.; Liu, Y. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E6117–E6125. [CrossRef] [PubMed]
20. Tian, J.; Yan, Y.; Yue, Q.; Liu, X.; Chu, X.; Wu, N.; Fan, Y. Predicting synonymous codon usage and optimizing the heterologous gene for expression in *E. coli*. *Sci. Rep.* **2017**, *7*, 9926. [CrossRef]
21. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
22. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
23. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]