Methods 102 (2016) 26–35

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Multi-omics data driven analysis establishes reference codon biases for synthetic gene design in microbial and mammalian cells

Kok Siong Ang [a,b,1], Sarantos Kyriakopoulos [c,1], Wei Li [d], Dong-Yup Lee [a,b,c,*]

[a] Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585, Singapore
[b] NUS Synthetic Biology for Clinical and Technological Innovation (SynCTI), Life Sciences Institute, National University of Singapore, 28 Medical Drive, Singapore 117456, Singapore
[c] Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06-01 Centros, Singapore 138668, Singapore
[d] Sangon Biotech (Shanghai) Co., Ltd., 698 Xiangmin Road, Songjiang District, Shanghai 201611, China

ARTICLE INFO

Article history:
Received 22 September 2015
Received in revised form 8 January 2016
Accepted 19 January 2016
Available online 2 February 2016

Keywords:
Reference codon bias
Codon optimization
Multi-omics data
Microbial and mammalian hosts
Synthetic gene design

ABSTRACT

In this study, we analyzed multi-omics data and subsets thereof to establish reference codon usage biases for codon optimization in synthetic gene design. Specifically, publicly available genomic, transcriptomic, proteomic and translatomic data for microbial and mammalian expression hosts, *Escherichia coli*, *Saccharomyces cerevisiae*, *Pichia pastoris* and Chinese hamster ovary (CHO) cells, were compiled to derive their individual codon and codon pair frequencies. Then, host dependent and -omics specific codon biases were generated and compared by principal component analysis and hierarchical clustering. Interestingly, our results indicated the similar codon bias patterns of the highly expressed transcripts, highly abundant proteins, and efficiently translated mRNA in microbial cells, despite the general lack of correlation between mRNA and protein expression levels. However, for CHO cells, the codon bias patterns among various -omics subsets are not distinguishable, forming one cluster. Thus, we further investigated the effect of different input codon biases on codon optimized sequences using the codon context (CC) and individual codon usage (ICU) design parameters, via *in silico* case study on the expression of human IFNγ sequence in CHO cells. The results supported that CC is more robust design parameter than ICU for improved heterologous gene design.

1. Introduction

Efficient expression of heterologous genes is one of the key challenges in industrial biotechnology applications, where the goal is to enhance recombinant protein production. When natural genes are transformed into heterologous hosts, expression levels are typically poor, as the gene had evolved towards efficient expression in its original organism. Thus, in order to overcome this expression bottleneck, it is highly necessary to design synthetic gene effectively, by identifying molecular and regulatory principles related to protein translation and folding [1]. One general strategy for improving recombinant protein expression is codon optimization, as successfully demonstrated in both microbial (*Escherichia coli* [2–4], *Saccharomyces cerevisiae* [5–7], *Pichia pastoris* [8–10]) and mammalian cells (Chinese hamster ovary cells, CHO [11–13]).

Basically, codon optimization considers the degeneracy of the genetic code, *i.e.* that most proteinogenic amino acids (except for methionine and tryptophan) are encoded by two to six codons, and the uneven usage of synonymous codons by different hosts, termed as codon bias. For the codon optimization, codons of the heterologous target sequence to be expressed are replaced with those preferred by the host [14]. As such, it can increase the translation elongation rate of an mRNA molecule, which is one of the potential bottlenecks for recombinant protein expression intracellularly [15]. To this end, researchers have developed several algorithms and tools to optimize the DNA sequence of the recombinant protein [14,16,17], and explored the relevant design parameters (*e.g.* codon adaptation index, CAI; individual codon usage, ICU; codon context, CC [3,13,18]). However, little attention has been paid to the choice of host's reference codon bias that has been commonly computed using all coding sequences from the host's genome, or subsets of genes (Fig. 1). This input bias pattern (or reference table) to the codon optimization, clearly affects output sequences. Thus, it is important to establish appropriate reference bias representing the host's preferred codon usage pattern. In microbial cells, typically the sequences of highly expressed

* Corresponding author at: Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585, Singapore.
E-mail address: cheld@nus.edu.sg (D.-Y. Lee).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2016.01.016
1046-2023/

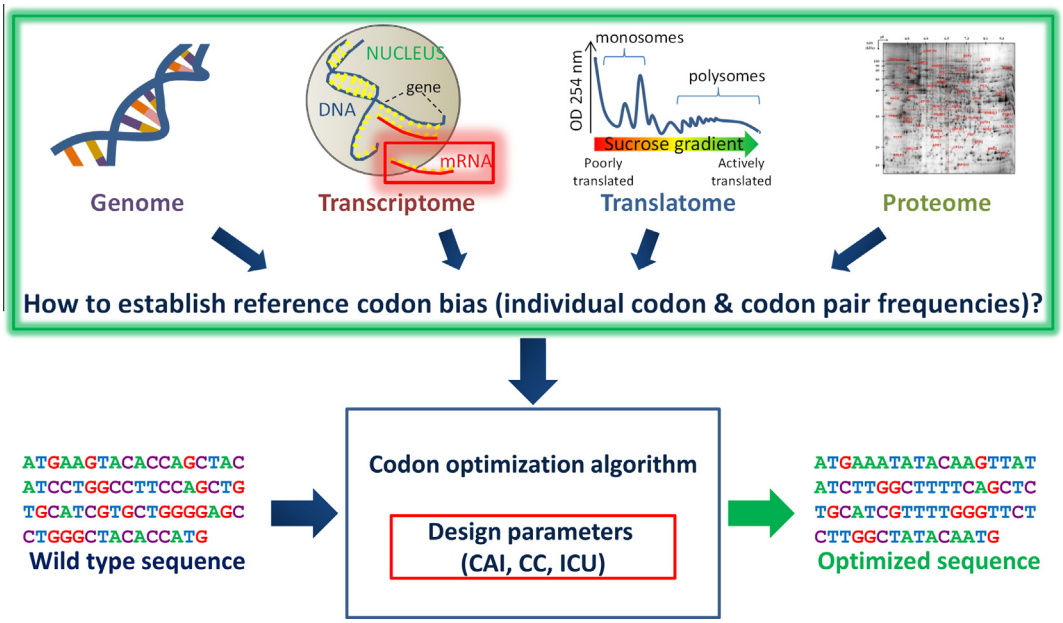

Fig. 1. The general codon optimization strategy. Inputs to the codon optimization include the heterologous gene (left) to be expressed in a certain host and the codon usage reference of the host (top). Selecting a design parameter (bottom) is also an essential part of the algorithmic process. The output (right) is an optimized sequence, where the codons of the heterologous target sequence to be expressed are replaced with those preferred by the host. In the upper part of the figure, the multi-omics datasets (genome, transcriptome, translatome, proteome) used in this study to calculate relevant codon usage biases are also depicted. **CAI**, codon adaptation index; **CC**, codon context; **ICU**, individual codon usage.

genes (as identified in the transcriptome) or the genome are used to calculate the codon usage reference of the host (see Supplementary file 1 for a comprehensive list of studies). Similarly, in mammalian cells and specifically the workhorse for biopharmaceutical glycoprotein production, CHO cells [19], the transcriptome data have been exploited mainly due to the data abundancy [13]. However, the need to cross-validate these results with other -omics datasets recently available with the enhancement of more sophisticated experimental and analytical techniques (e.g. proteome and translatome), is apparent due to the weak correlation between gene transcripts and protein abundance, which has been consistently reported [20].

In this study, we assessed multi-omics datasets and subsets thereof for various host cells in order to identify the most appropriate codon bias reference for codon optimization in synthetic gene design. Initially, we collected the genome, transcriptome, translatome and proteome datasets and extracted relevant subsets from four expression hosts: *E. coli*, *S. cerevisiae*, *P. pastoris* and Chinese hamster ovary (CHO) cells (Fig. 2). Subsequently, the resultant codon biases generated from the various -omics datasets were compared and contrasted using principal component analysis and hierarchical clustering. Finally, we investigated the effect of the codon bias inputs and design parameters (ICU and CC) on the codon optimized sequences via a case study with the human interferon-gamma (IFN-γ) sequence in CHO cells.

2. Material and methods

2.1. Multi-omics data sets for generating reference codon biases

We collected multi-omics datasets for various microbial and mammalian host cells, *E. coli*, *S. cerevisiae*, *P. pastoris*, and Chinese hamster ovary (CHO) cells to calculate the codon bias patterns of the respective gene subsets. They include the genome, transcriptome, translatome and proteome providing various snapshots of translational efficiency and control, as well as global gene expression levels. The information on the datasets is summarized in Table 1. Note that translatome data for *P. pastoris* is not included due to the difficulty to process the raw data available in [21]. The gene lists of the -omics dataset were used to extract the corresponding coding sequences to calculate their respective codon biases. The genes in each available -omics dataset were also divided into subsets of top 10% expression and 10% lowest expression to calculate their respective codon biases. For the CHO translatome (*CHO transl. hi*), the list of translationally efficient CHO genes was obtained directly from [22].

2.2. Calculating reference codon biases from multi-omics datasets

The individual codon and codon pair frequency tables for expression hosts were computed from each -omics dataset provided in Table 1. The individual codon distribution is a vector of occurrence frequencies for all 64 codons. The frequency value p^k for each codon is defined as:

$$p^k = \frac{\theta_C^k}{\theta_A^{j(k)}} \forall k \in \{1, 2, \ldots, 64\} \quad (1)$$

where θ_C^k denotes the total number of occurrences of codon k, and $\theta_A^{j(k)}$ denotes the total number of occurrences of the amino acid encoded by codon k. As each amino acid is encoded by at least one codon, the value of p^k falls in the range of $0 \leqslant p^k \leqslant 1$. In the similar manner, the codon pair distribution covers all 3904 possible codon pairs ($61 \times 64 = 3904$, as the first codon in each pair cannot be a stop codon). The frequency value q^k for each codon pair is defined as:

$$q^k = \frac{\theta_{CC}^k}{\theta_{AA}^{j(k)}} \forall k \in \{1, 2, \ldots, 3904\} \quad (2)$$