BBA - Biomembranes 1860 (2018) 2479–2485

Contents lists available at ScienceDirect

BBA - Biomembranes

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

Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons

Alexander L. Cope[a], Robert L. Hettich[a,b], Michael A. Gilchrist[a,c,d,*]

[a] Genome Science and Technology, University of Tennessee, Knoxville, United States of America
[b] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States of America
[c] Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, United States of America
[d] National Institute for Mathematical and Biological Synthesis, Knoxville, TN, United States of America

ARTICLE INFO

Keywords:
Codon usage bias
Signal peptides
Protein secretion
Protein synthesis
Adaptationist

ABSTRACT

The Sec secretion pathway is found across all domains of life. A critical feature of Sec secreted proteins is the signal peptide, a short peptide with distinct physicochemical properties located at the N-terminus of the protein. Previous work indicates signal peptides are biased towards translationally inefficient codons, which is hypothesized to be an adaptation driven by selection to improve the efficacy and efficiency of the protein secretion mechanisms. We investigate codon usage in the signal peptides of *E. coli* using the Codon Adaptation Index (CAI), the tRNA Adaptation Index (tAI), and the ribosomal overhead cost formulation of the stochastic evolutionary model of protein production rates (ROC-SEMPPR). Comparisons between signal peptides and 5′-end of cytoplasmic proteins using CAI and tAI are consistent with a preference for inefficient codons in signal peptides. Simulations reveal these differences are due to amino acid usage and gene expression – we find these differences disappear when accounting for both factors. In contrast, ROC-SEMPPR, a mechanistic population genetics model capable of separating the effects of selection and mutation bias, shows codon usage bias (CUB) of the signal peptides is indistinguishable from the 5′-ends of cytoplasmic proteins. Additionally, we find CUB at the 5′-ends is weaker than later segments of the gene. Results illustrate the value in using models grounded in population genetics to interpret genetic data. We show failure to account for mutation bias and the effects of gene expression on the efficacy of selection against translation inefficiency can lead to a misinterpretation of codon usage patterns.

1. Introduction

A secreted protein can broadly be defined as any protein entering a secretory pathway for transport through a cellular membrane. These proteins serve important cellular functions, including metabolism and antibiotic resistance [1,2]. Secreted proteins also play essential roles in the virulence of pathogenic bacteria [1]. Numerous secretion systems exists and vary between and within taxa [1-3]. Despite the diversity of secretion pathways, the general secretion pathway, also commonly referred to as the Sec pathway, is found across all domains of life [1,4]. In brief, proteins are transported to the SecYEG translocon located in the membrane in a chaperone-dependent (SecA/B and SRP) or chaperone-independent manner [4,5]. All SecA/B-dependent proteins and chaperone-independent, as well as some SRP-dependent proteins, contain a short peptide chain located at the N-terminus of the protein known as the signal peptide [1,4,5]. The signal peptide is an essential component of the Sec pathway, serving as a binding site for the appropriate chaperones and/or helping delay the folding of the protein [4,5]. Although signal peptides do vary in their amino acid sequences, signal peptides have distinct physicochemical properties which biases their amino acid usage [4-6]. A signal peptide generally consists of 3 regions: a positively charged N-terminus, a hydrophobic core, and a polar C-terminus, where the signal peptide is cleaved from the rest of the protein, sometimes referred to as the " mature peptide."

The ability to accurately predict signal peptides is useful for identifying secreted proteins in non-model organisms; this has led to the development of machine learning approaches to predict signal peptides which take advantage of the distinct physicochemical properties of signal peptides, such as SignalP [7]. Although the physicochemical properties of signal peptides are consistent, altering the N-terminus has a range of effects on protein secretion: from a decrease in the number of proteins secreted to no observable effect [8-11]. The variability in the

*abbreviations:*CUB, codon usage bias

* Corresponding author at: 569 Dabney Hall, University of Tennessee, Knoxville, TN 37996-1610, United States of America.
E-mail address: mikeg@utk.edu (M.A. Gilchrist).

https://doi.org/10.1016/j.bbamem.2018.09.010
Received 28 June 2018; Received in revised form 11 September 2018; Accepted 13 September 2018
Available online 19 September 2018

outcomes of neutralizing the N-terminal positive charge led to a search for other mechanisms which also contribute to the efficacy of protein secretion [6,12].

Numerous studies suggests codon usage bias (CUB) – the non-uniform usage of synonymous codons – contributes to effective protein secretion in *E. coli* [13-18]. Power et al. [14] found *E. coli* K12 MG1655 signal peptides are biased for translation inefficient codons, which are predicted to be translated slower than their synonymous counterparts. This is in stark contrast to the rest of the *E. coli* proteome, where *E. coli* is biased towards the most efficient codons [14,19]. Li et al., Liu et al., and Mahlab and Linial [20-22] examined the usage of inefficient codons in signal peptides of *S. coelicolor*, *S. cerevisiae*, and various multicellular eukaryotes and came to similar conclusions when applying codon usage indices such as the Codon Adaptation Index (CAI) [23] and tRNA Adaptation Index (tAI) [24]. Consistent across this work is the interpretation that selection is driving the apparent increase in inefficient codon usage in signal peptides. Furthermore, Zalucki et al. [25] concluded an overabundance of the lysine codon AAA at the second position in the signal peptide promoted efficient translation initiation.

Zalucki et al. [6] hypothesized an adaptive role for inefficient codons in the protein secretion process in which the combination of efficient translation initiation and inefficient translation reduced the distance between sequential ribosomes along the mRNA, leading to more efficient recycling of the necessary chaperones. Other explanations for the observed increase in inefficient codons include the inability of *E. coli* SRP to induce a translational pause following signal peptide recognition [6,26] and slowing down the co-translational folding of the protein, as a folded protein cannot be translocated through the SecYEG translocon [12,14-16]. If signal peptides have a different CUB relative to the rest of the genome, then codon-level information could be incorporated into signal peptide prediction tools.

In contrast Liu et al. [21] found no significant differences in the ribosome densities between the signal peptides and the 5′ -ends of nonsecretory genes in various eukaryotes. Ribosome densities are expected to be higher in signal peptides relative to the 5′ -end of nonsecretory genes if selection is acting to increase translation inefficiency in the signal peptide. Additionally, while both Liu et al. [21] and Mahlab and Linial [22] examined codon usage in relation to secretion in *H. sapiens* using a metric based on tAI, only Mahlab and Linial [22] found results consistent with increased frequencies of inefficient codons in signal peptides. From a population genetics perspective, it is surprising statistically significant results were obtained in a mammal, which usually have little adaptive CUB due to their lower effective population sizes [27,28]. More recently, Samant et al. [29] found codon optimization of a signal peptide improved localization of the protein to the periplasm of *E. coli*, seemingly contradicting a general role for inefficient codon usage in signal peptides. A potential reason for these contradictions is the previous analyses of signal peptide codon usage by [14,20-22] did not adequately account for the effects of mutation bias and drift in shaping codon usage [30-35].

We re-examined CUB in signal peptides of *E. coli* using CAI, tAI, and ROC-SEMPPR - a population genetics model which accounts for selection, mutation bias, and gene expression - to determine if selection on codon usage in signal peptides differs from the 5′ -ends of genes. Although we find significant differences in codon usage using CAI and tAI, we present evidence these differences are due to signal peptide-specific amino acid biases and differences in the gene expression distributions of genes with and without signal peptides. When comparing signal peptides and the 5′ -ends of genes not containing a signal peptide with ROC-SEMPPR, we find signal peptide codon usage is consistent with the 5′ -ends. We find selection on codon usage favors the efficient codons, but the strength of selection is weaker at the 5′ -ends, corroborating previous analyses [14,31,32,36,37].

Our work demonstrates the value of analyzing CUB from a formal population genetics framework, as well as highlights potential limitations with using more common metrics such as CAI for analyzing codon usage on relatively small regions of the genome. Failure to account for variation in the strength of selection due to variation in gene expression can lead to conflating mutation bias with selection, resulting in a misinterpretation of observed codon usage patterns. Our work also illustrates the importance of considering non-adaptive forces in shaping biological phenomenon before invoking adaptive explanations [38]. We believe this is particularly important in the modern genomic-age when the combination of large datasets, misinterpretation of p-values, and an inherent bias towards adaptationist interpretations can lead to the proliferation of over-interpreted hypotheses within the biological community.

2. Materials and methods

2.1. Signal peptide prediction

Signal peptides were predicted using SignalP 4.1 [7] using both the default cutoff D-score of 0.51 and a more conservative D-score of 0.75. In brief, SignalP consists of two neural networks, one for determining the amino acid sequence similarity to signal peptides and the other for identifying the most likely cleavage site. The results of both neural networks are combined into one value, called the D-score, which ranges between 0 and 1. Setting the cutoff D-score closer to 1 results in a lower false positive rate. A set of confirmed signal peptides for *E. coli* K12 MG1655 was taken from The Signal Peptide Website (http://www.signalpeptide.de/). All analyses in the main text will focus on the set of signal peptides with $D \geq 0.51$ as this set provides us with the most data; analyses of the $D > 0.75$ and set of confirmed signal peptides give similar results (see Supplementary material).

2.2. ROC-SEMPPR

Given a set of protein-coding genes, ROC-SEMPPR employs a Markov Chain Monte Carlo (MCMC) to estimate codon specific parameters for mutation bias ΔM and pausing times $\Delta\eta$ for each codon within a synonymous codon family. In previous work, $\Delta\eta$ was scaled relative to the most efficient codon, which had $\Delta\eta$ and ΔM values fixed at 0. To avoid the choice of reference codon affecting our comparisons of CUB between regions, all $\Delta\eta$ values in this paper are re-scaled by the mean such that these values are centered around 0 for each amino acid. The $\Delta\eta$ values reflect the strength and direction of selection against translation inefficiency in a set of protein-coding regions (e.g. the signal peptides). A region with stronger selection against translation inefficiency will have higher $\Delta\eta$ values on average than a region with weaker selection. Similarly, a region which favors translation inefficiency would be expected to have $\Delta\eta$ values which negatively correlate with a region which favors translation efficiency.

ROC-SEMPPR also estimates an average protein production rate ϕ for each gene. It is important to note ROC-SEMPPR is structured such that the average value of ϕ across the genome is 1. This choice of scaling means the pausing times $\Delta\eta$ represent the average strength of selection relative to genetic drift for or against a given codon. We find ROC-SEMPPR estimated ϕ values correlate well with empirical measurements of protein production rates for *E. coli* (see Supplementary Methods: Assessing ROC-SEMPPR Model Adequacy and Figs. S1–S2). If changes in synonymous codon usage alter the efficiency at which a protein is translated, then such a change will have the largest impact on the energetic costs of proteins with high production rates, making ϕ a more appropriate gene expression metric than say, mRNA abundance or protein abundance. Thus, we use protein production rates ϕ as our metric of gene expression. For more details on ROC-SEMPPR, see [33]. Analysis of CUB with ROC-SEMPPR was performed using AnaCoDa [39].