Journal of Theoretical Biology 413 (2017) 66–71

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Translational resistivity/conductivity of coding sequences during exponential growth of *Escherichia coli*

Kazuyuki Takai

Department of Materials Sciences and Biotechnology, Graduate School of Science and Engineering, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime 790-8577, Japan

ARTICLE INFO

Keywords:
Codon adaptation index
Electric circuit model
Steady state
Relative codon usage

ABSTRACT

Codon adaptation index (CAI) has been widely used for prediction of expression of recombinant genes in *Escherichia coli* and other organisms. However, CAI has no mechanistic basis that rationalizes its application to estimation of translational efficiency. Here, I propose a model based on which we could consider how codon usage is related to the level of expression during exponential growth of bacteria. In this model, translation of a gene is considered as an analog of electric current, and an analog of electric resistance corresponding to each gene is considered. "Translational resistance" is dependent on the steady-state concentration and the sequence of the mRNA species, and "translational resistivity" is dependent only on the mRNA sequence. The latter is the sum of two parts: one is the resistivity for the elongation reaction (coding sequence resistivity), and the other comes from all of the other steps of the decoding reaction. This electric circuit model clearly shows that some conditions should be met for codon composition of a coding sequence to correlate well with its expression level. On the other hand, I calculated relative frequency of each of the 61 sense codon triplets translated during exponential growth of *E. coli* from a proteomic dataset covering over 2600 proteins. A tentative method for estimating relative coding sequence resistivity based on the data is presented.

1. Introduction

Design of gene sequences is getting more important as the cost for gene synthesis is getting lower in these days. One of the major purposes of gene design is production of proteins from various sources in *Escherichia coli* and other organisms for large-scale preparation and biochemical characterization. It may be a common understanding that bacterial cells have evolved so that selected codon triplets are translated more efficiently than the other synonymous triplets, and that the synonymous codon choice should be optimized for the host cells for maximal production of external proteins (Elena et al., 2014).

The most popular measure of adaptation of coding sequences to the host organism is the codon adaptation index (CAI) (Sharp and Li, 1986, 1987). CAI is calculated as

$$\mathrm{CAI} = \left(\prod_{k=1}^{L} w_k \right)^{1/L},$$

where w_k is a pre-determined parameter for the triplet that appeared as the k-th codon in the coding sequence, and L is the number of codons in the coding sequence excluding ATG and TGG. The w_k is dependent only on the identity of the triplet and is given as the frequency of appearance of the triplet in highly-expressed genes divided by that of the most frequently used synonymous codon. CAI is originally calculated referring to the codon usage data for highly-expressed genes, while other codon usage data can be used. In most of the later variations of CAI, the ATG and TGG codons are taken into account, with L corresponding to the length of the coding sequence. CAI has been used also as a measure of translational efficiency of a coding sequence, probably based on the observation that minor codons generally slow down elongation rates (Sørensen et al., 1989).

The tRNA adaptation index (tAI) is also a popular measure of codon adaptation (dos Reis et al., 2004), in which the w_k parameters for CAI are substituted with a set of parameters that represent relative activities of aminoacyl-tRNA molecules that recognize the triplets. It may be a more straightforward measure of translational efficiency than CAI because it is based on an estimation of aminoacyl-tRNA availability, which is, in other words, cellular activity of decoding of each codon triplet. It was later postulated that the balance between the tRNA availability (supply) and frequency of appearance of each triplet (demand) determines the optimality of each triplet, and the normalized translational efficiency (nTE) parameters for the 61 sense codons were calculated for yeast (Pechmann and Frydmann, 2013).

tAI and the nTE scales have been successfully applied to estimation of translational efficiency of cellular mRNAs in yeast, in particular

E-mail address: takai.kazuyuki.mz@ehime-u.ac.jp.

http://dx.doi.org/10.1016/j.jtbi.2016.11.015
Received 4 July 2016; Received in revised form 11 November 2016; Accepted 17 November 2016
Available online 19 November 2016

(Tuller et al., 2010). On the other hand, CAI has been widely used as an evaluation parameter for codon-optimized artificial genes, though several other measures of fitness of artificial genes have been proposed to better predict recombinant expression levels and used for gene design (Gould et al., 2014). However, it is now evident that CAI and other measures depending only on the codon composition cannot predict the efficiency of recombinant expression by themselves. The major determinant of translational efficiency has been revealed to be the secondary structure potency of the mRNA sequence near the initiation codon (Goodman et al., 2013) at least in *E. coli*, but not the codon choice.

The fact that the mRNA higher order structure is the most important determinant does not directly mean that it is the only reason why CAI cannot predict gene expression levels. The above equation can be rewritten as

$$\ln \mathrm{CAI} = \frac{1}{L}\sum_{k=1}^{L} \ln w_k.$$

Thus, ln CAI is the mean of parameters assigned to the composing triplets. The same is true of tAI. In my opinion, there is no mechanistic reason that rationalizes the use of the mean of the parameters if the indices are measures of translational efficiency. In other words, the relationship between the translational efficiency of a coding sequence and the parameters for the composing triplets is not clear for these indices.

In the present paper, I postulate a model of a quantitative aspect of protein synthesis in exponentially growing bacterial cells. The model clarifies the meaning of an alternative codon adaptation measure calculated through summing up the parameters for the composing codon triplets.

2. Theory

2.1. Concentration of every component is assumed to be constant during exponential growth

I assume here that, during exponential growth of bacteria, the volume V [m^3] of the system, which is the inner part of bacteria including cell wall, grows exponentially along time, t [s], according to $\frac{dV}{dt} = \alpha V$, where α [s^{-1}] is a constant inversely proportional to the doubling time. The cellular chemical reactions are in the steady state, so that concentration of every component is constant. The steady-state concentration of the i-th protein, C_i, and that of the mRNA species that codes for the i-th protein, m_i, are constant. For simplicity, I do not take into account the fact that some mRNA species code for more than one proteins. The amount of the i-th protein in the system, n_i [mol], increases exponentially over time because it is proportional to V, that is, $n_i = C_iV$.

2.2. Translational current and resistance

The rate of synthesis of the i-th protein, v_i [mol s^{-1}], minus the sum of the rates for degradation and flow out of the bacterial body, v'_i, equals to the accumulation rate of the i-th protein, which means $v_i - v_i' = \frac{dn_i}{dt}$. v_i and v'_i are proportional to V. v_i divided by V is a constant and represents the rate of translation per volume. This is termed here "translational current" and symbolized by I_i. v'_i over V is also a constant and symbolized here by I'_i. Therefore, $I_i = \frac{1}{V}\frac{dn_i}{dt} + I_i'$ [mol m^{-3} s^{-1}] is obtained.

I consider here an analog of electric potential or electromotive force, E, that drives translation. The dimension associated with E is not known. The unit of E is tentatively symbolized by "U" in this text. It is assumed here that E is independent of i. "Translational conductance", G_i [mol m^{-3} s^{-1} U^{-1}], is now defined by $G_i = \frac{I_i}{E}$, and "translational resistance", R_i [mol^{-1} m^3 s U], by $R_i = \frac{1}{G_i} = \frac{E}{I_i}$ (Fig. 1a).

R_i is the sum of the contributions of the elementary reactions that proceed successively during one round of translation of the i-th mRNA species, as the elementary reactions are modeled by resistance units tandemly connected together on a branch of the electric circuit. Here, I divide R_i into contributions of two pieces of resistance units. One corresponds to decoding of all of the sense codons and is symbolized by R_{CSi}, where "CS" symbolizes "coding sequence". The other corresponds to all of the other elementary reactions, including initiation, termination, *etc.*, and is symbolized by R_{0i}. Therefore, $R_i = R_{0i} + R_{CSi}$ (Fig. 1b).

In this view, an mRNA species is modeled as a resistance unit connected in parallel with the other ones to a direct current power source (Fig. 1b). The flow of synthesis of a protein per volume is modeled as electric current in a branch of the electric circuit. Thus, the relative rate of production of different protein species is proportional to the relative conductance of the mRNA species. The model may be termed "the electric circuit model" hereafter.

2.3. Translational conductivity and resistivity

G_i is considered to be proportional to m_i. Thus, G_i divided by m_i should be dependent only on the sequence of the mRNA and is termed here "translational conductivity". Translational conductivity of the i-th mRNA is represented here by g times g_i, so that g_i is dimensionless and g does not depend on i: $gg_i = \frac{G_i}{m_i}$ [s^{-1} U^{-1}]. The reciprocal of gg_i is "translational resistivity" and is r times r_i: $rr_i = \frac{m_i}{G_i} = m_iR_i$ [s U], again r_i being dimensionless and r independent of i. g_i and r_i are termed relative translational conductivity and resistivity, respectively.

By multiplying the equation $R_i = R_{0i} + R_{CSi}$ (Section 2.2) by m_i/r, $r_i = r_{0i} + r_{CSi}$ is obtained, where r_{0i} and r_{CSi} come from R_{0i} and R_{CSi}, respectively. In the theory, "coding sequence" contains all of the sense (elongation) codons to be decoded by A-site aminoacyl-tRNAs, but not the initiation and termination codons. r_{CSi} depends only on the i-th coding sequence (CSi), while r_{0i} may depend on the CSi sequence in addition to the other parts of the mRNA sequence.

r_{CSi} is further separated into the contributions of the codons that constitute CSi. Thus, $r_{CSi} = \sum_{j=1}^{L_i} r_{ij}$, where L_i is the number of codons in CSi, and j represents the position of each codon. Relative conductivity of CSi could also be defined by $g_{CSi} = \frac{1}{gr}\frac{1}{r_{CSi}}$. r_{ij} is, in general, a function of i and j, while, if it depends only on the identity of the codon triplet, r_{CSi} could be estimated by summing up pre-determined parameters assigned to the 61 kinds of sense codon triplets. This can be formulated as $r_{CSi} = \sum_p^{61} q_{ip} r_p$, where p specifies one of the 61 sense codon triplets, and q_{ip} is the number of codon triplet p within CSi. The "61" over the sigma symbol means that it is the sum for all of the 61 different sense codon triplets. r_p may be termed here relative resistivity parameter for triplet p, or relative codon resistivity parameter.

2.4. A tentative method for parameter determination

One of the major problems associated with the electric circuit model is how r_p for each codon could be reasonably determined. In the present paper, I tentatively define a set of r_p parameters in the following sections, based on a codon usage dataset for *E. coli*, just as the w_k parameters for calculation of CAI were determined from codon usage data. However, I attempt here to use a more refined dataset.

During exponential growth, increase of the amount of the i-th protein should be proportional to the amount of the i-th protein: $\frac{dn_i}{dt} = \alpha n_i$. It is expected that the amount of the proteins removed from the system by degradation or by secretion is much smaller than that of the proteins remaining within the system, which means $I_i \approx \frac{\alpha}{V} n_i = \alpha C_i$. Therefore, if relative abundance of each of the protein species composing the cells during exponential growth could be measured somehow, relative frequency of decoding of each of the 61 sense codon triplets (relative codon usage: RCU) could be calculated referring to the gene sequences.