Journal of Theoretical Biology 296 (2012) 33–40

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

A novel coding method for gene mutation correction during protein translation process

Lei Zhang [a,*], Fengchun Tian [a], Shiyuan Wang [b], Xiao Liu [a]

[a] College of Communication Engineering, Chongqing University, 174 ShaPingBa District, Chongqing, 400044, PR China
[b] School of Electronic and Information Engineering, Southwest University, No. 2 Tiansheng Road, BeiBei District, Chongqing, 400715, PR China

ARTICLE INFO

Article history:
Received 25 May 2011
Received in revised form
25 August 2011
Accepted 30 November 2011
Available online 9 December 2011

Keywords:
Gene expression
Error-correction coding
Shine–Dalgarno
Translation initiation

ABSTRACT

In gene expression, gene mutations often lead to negative effect of protein translation in prokaryotic organisms. With consideration of the influences produced by gene mutation, a novel method based on error-correction coding theory is proposed for modeling and detection of translation initiation in this paper. In the proposed method, combined with a one-dimensional codebook from block coding, a decoding method based on the minimum hamming distance is designed for analysis of translation efficiency. The results show that the proposed method can recognize the biologically significant regions such as Shine–Dalgarno region within the mRNA leader sequences effectively. Also, a global analysis of single base and multiple bases mutations of the Shine–Dalgarno sequences are established. Compared with other published experimental methods for mutation analysis, the translation initiation can not be disturbed by multiple bases mutations using the proposed method, which shows the effectiveness of this method in improving the translation efficiency and its biological relevance for genetic regulatory system.

1. Introduction

The process of protein translation initiation and the formation of the initiation codons determine the efficiency of gene expression at the translation level. The sequences surrounding the start triplet codons act as determinants influencing the gene expression efficiency (Dreyfus, 1988). The Shine–Dalgarno sequence (SD), a few bases upstream of the initiation codon is complementary to the anti-SD sequence near the 3′ end of the 16S rRNA (Shine and Dalgarno, 1974; Faxen et al., 1991). A region located downstream of the start codon has been suggested to influence translation initiation by mRNA–rRNA complementary base pairing (Stenstrom et al., 2001). Even though several other sequences surrounding the start codon are speculated to influence the accuracy of translation initiation, the initiation codons and the SD sequence is generally accepted to be the key determinants during the translation initiation (Prescott and Goringer, 1990). Published results (Hui and De Boer, 1987; Jacob et al., 1987) showed that multiple bases mutations and single base mutations (point mutation) of SD reserved sequences in the 3′-end of the 16S rRNA would be lethal to protein translation. Hui and De Boer (Hui and De Boer, 1987) described the behavior of protein synthesis under mutations in the 3′-end of the 16S rRNA. The mutations were done in positions 2–6 (GGAGG→CCUCC) and positions 3–5 (GAG→UGU) of the SD sequence of a single mRNA species. Jacob et al. (1987) introduced a point mutation in position 5 of the 13 last bases of the 16S rRNA which considered a change of the nucleotide C→U. With knowledge that SD mutation has been thought to be the main reason to influence the efficiency of gene expression at the translation level for that the translation will be advanced or delayed, and even terminated earlier. However, whether the influence produced by mutations can be effectively controlled, has become a key problem. Fortunately, relative research showed that the role of Shine–Dalgarno regions is not the only mechanism to promote translation, and the SD-independent translation initiation mechanism has also been published (Fargo et al., 1998). Also, four different mechanisms such as one prokaryotic mechanism involving Shine–Dalgarno sequence, two eukaryotic mechanisms and one mechanism acting on leaderless transcripts have been concluded. This provides a wide prospect for dealing with mutations effectively.

In the field of mutation correction, more research tends to find all differences between a standard normal ß-globin gene and a suspected abnormal gene and identify any differences as point mutations or frame-shift mutations to list important biochemical effects of the mutations (Sunthornwat et al., 2011). Besides, determination and quantification of point mutation rates by specific nucleotide sequences such as *Escherichia coli* become possible (Kini and Chinnasamy, 2010; Wielgoss et al., 2011). Research also demonstrate that DNA repair and replication would influence the number of mutations per adduct of polycyclic aromatic hydrocarbons in mammalian cells (Lagerqvist et al., 2011). Gore et al. (2011) show that 22 human induced pluripotent

* Corresponding author. Tel.: +86 13629788369.
E-mail address: leizhang@cqu.edu.cn (L. Zhang).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.11.031

stem (hiPS) cell lines reprogrammed using five different methods each contained an average of five protein-coding point mutations in the regions sampled. They declared that hiPS cells acquire genetic modifications in addition to epigenetic modifications. Howden et al. (2011) isolated iPS cells free of transgene sequences from a patient with gyrate atrophy caused by a point mutation in the gene encoding ornithine aminotransferase (OAT) and used homologous recombination to correct the genetic defect.

In correction of gene mutations, error-correction coding theory in communication systems (Sweeney, 1991; Roman, 1995) has been widely used. The study of the information processing capabilities of living systems was revived in the later part of the 1980s, resulting from the increase in genomic data. Information measures, such as entropy, have been used in recognition of DNA patterns, classification of genetic sequences and other computational studies of genetic processes (Schneider, 1997, 1999). And recently, error correction coding has also been introduced for modeling the process of translation initiation of prokaryotic organisms (May, 2004; Dawy et al., 2009; Oliver et al., 2009; May et al., 2004; Bataineh et al., 2011; Mian and Rose, 2011). They found that communication theory could contribute to understanding of myriad of problems in biology and investigations of multicellular biology could also, in turn, lead to advances in communication theory. A detailed exploration of good DNA code design has also been illustrated (Garzon and Deaton, 2004; Rosen, 2006). May (2004) explored the feasibility of using communication theory to model the protein translation initiation mechanism. The mRNA of *E. coli* K-12 was modeled as a noisy (errored) signal and encoded the ribosome as a minimum hamming distance decoder. Dawy et al. (2009) presented the translation initiation modeling and mutational analysis, and showed experimental results for different mutations of the last 13 bases of the 16S rRNA molecule using an average free energy ribosome decoding method. The results of those approaches showed that the different proposed codes were able to distinguish between a set of mRNA sequences that are translated into proteins and a set of mRNA sequences that are not translated. Error correcting codes have also been applied for other processes modeling in genetics (Gupta, 2006). For example, a parity check code interpretation of the nucleotide composition was proposed in Mac Dónail (2006). A related class of problems deals with designing algorithms to detect translation sites in individual sequences using neural networks which use training sequences to develop a model based on their inherent properties (Mori et al., 2007). In fact, mutations and errors presented within the genome of an organism are analogous to an error-producing channel used by an engineering system to transmit information to a receiver. Genetic processes, such as replication may also introduce errors (mutations) into the genome of an organism. Mutations or variations in a genomic sequence can also be caused by external forces such as temperature, radiation and, environment (Rosen, 2006). Research in information based sequence analysis showed that ribosome binding sites evolve to functional requirements rather than perfect sequences. Previous investigations also suggested that it may be possible for us to design an effective error-control coding based algorithm for detecting the leader regions of translated messenger RNA sequences in prokaryotic genomes and it inspires us to explore whether there is an inherent coding structure in DNA.

In genome analysis, the representation of a DNA sequence is the first stage. Until now, many different kinds of representations of nucleic acids sequences which are long polymers of four nucleotide bases: adenine (A), cytosine (C), guanine (G) and ,thymidine (T, DNA) or uracil (U, mRNA), have been outlined for different applications such as analysis of similarity and dissimilarity of DNA sequences. One advantage of graphical representation is that they allow visual inspection of DNA data, helping to recognize major differences among similar DNA sequences (Randic et al., 2009). A 2D graphical representation proposed by Nandy involves an arbitrary assignment of the four bases to four directions of Cartesian coordinate axes which is accompanied by some loss of visual information associated with crossing and overlapping of the curve by itself (Nandy, 1994). More recently, it has been shown that graphical representations of a DNA sequence could lead to numerical characterizations of the sequence (Randic et al., 2003). The novel 2D graphical representation proposed by Randic in which the four bases are assigned to four symmetric non-equivalent horizontal lines has overcome the limitations of the Nandy's method. However, the published representations ignored the element of the bases mutations in gene expression. For knowledge of representation of genes, we refer to the published review about graphical representation of proteins in which a number of graphical representations of DNA and the secondary structure of RNA have been described in detail (Randic et al., 2011). For degrading the effect of mutations, an effective 2D graphical representation is proposed in this paper based on the properties of triplet of codons which can maintain the stability of biological system with gene mutations.

In this study, we consider the prokaryotic mechanism and the proposed error-correction model based on minimum hamming distance by using a one-dimensional codebook under the outlined 2D graphical representation proved that multiple bases mutations of SD sequences in the process of translation could be controlled to some extent. We focus on the computational recognition of the biological relevance for genetic regulatory system. This paper is organized as follows. Section 2 illustrates an effective 2D graphical representation of DNA based on the principle of triplet codons which also contain all the biological characteristics of DNA sequences. In Section 3, considering the mechanism of translation, a block coding model is constructed for generating a codebook, and a minimum hamming distance decoding method is presented for analysis of the translation initiation. And the multiple bases mutations and single base mutations are introduced to Shine–Dalgarno sequences to analyze the efficiency of gene expression in Section 4. Moreover, the newly proposed method is also compared with existing methods for detecting a better performance of the presented method. The conclusions of this paper are given in Section 5.

2. Improved 2D graphical representation

In a DNA primary sequence, according to biological characteristics of DNA sequences, the four DNA bases A, C, G and, T can be divided into three types: purine (R=A, G) and pyrimidine (Y=C, T); amino (M=A, C) and, keto (K=G, T); weak H-bonds (W=A, T) and, strong H-bonds (S=G, C).

Considering the initiation signal ATG and the principle of triplet codons, we have divided them into three types in which 8 combinations are included: (I) {RRR, RRY, YRR, YRY, YYY, YYR, RYY, RYR}, (II) {WWW, WWS, SWW, SWS, SSS, SSW, WSS, WSW} and, (III) {KKK, KKM, MKK, MKM, MMM, MMK, KMM, KMK}.

In detail, we denote a DNA primary sequence by $S=\{s_1, s_2, \ldots, s_N\}$ (N is the number of bases). With consideration of the triplet codon as a unit, we define a mapping set Ψ shown as follows

$$\Psi(S)=\{\Psi(s_1s_2s_3),\Psi(s_2s_3s_4),\ldots,\Psi(s_is_{i+1}s_{i+2}),\quad i=1,2,\ldots,N-2\} \tag{1}$$

Then the mapping set Ψ is designed in details as follows.

$$(\mathrm{I}):\ \Psi_1(s_is_{i+1}s_{i+2})=\begin{cases}(i,0), & if\ \ s_is_{i+1}s_{i+2}\in RRR\\ (i,1), & if\ \ s_is_{i+1}s_{i+2}\in RRY\\ (i,2), & if\ \ s_is_{i+1}s_{i+2}\in YRR\\ (i,3), & if\ \ s_is_{i+1}s_{i+2}\in YRY\\ (i,4), & if\ \ s_is_{i+1}s_{i+2}\in YYY\\ (i,5), & if\ \ s_is_{i+1}s_{i+2}\in YYR\\ (i,6), & if\ \ s_is_{i+1}s_{i+2}\in RYY\\ (i,7), & if\ \ s_is_{i+1}s_{i+2}\in RYR\end{cases} \tag{2}$$

$$(\mathrm{II}):\Psi_2(s_is_{i+1}s_{i+2})=\begin{cases}(i,0), & if\ \ s_is_{i+1}s_{i+2}\in WSW\\ (i,1), & if\ \ s_is_{i+1}s_{i+2}\in WSS\\ (i,2), & if\ \ s_is_{i+1}s_{i+2}\in SSW\\ (i,3), & if\ \ s_is_{i+1}s_{i+2}\in SSS\\ (i,4), & if\ \ s_is_{i+1}s_{i+2}\in SWW\\ (i,5), & if\ \ s_is_{i+1}s_{i+2}\in SWS\\ (i,6), & if\ \ s_is_{i+1}s_{i+2}\in WWW\\ (i,7), & if\ \ s_is_{i+1}s_{i+2}\in WWS\end{cases} \tag{3}$$

$$(\mathrm{III}):\Psi_3(s_is_{i+1}s_{i+2})=\begin{cases}(i,0), & if\ \ s_is_{i+1}s_{i+2}\in KKK\\ (i,1), & if\ \ s_is_{i+1}s_{i+2}\in KKM\\ (i,2), & if\ \ s_is_{i+1}s_{i+2}\in MKK\\ (i,3), & if\ \ s_is_{i+1}s_{i+2}\in MKM\\ (i,4), & if\ \ s_is_{i+1}s_{i+2}\in MMM\\ (i,5), & if\ \ s_is_{i+1}s_{i+2}\in MMK\\ (i,6), & if\ \ s_is_{i+1}s_{i+2}\in KMM\\ (i,7), & if\ \ s_is_{i+1}s_{i+2}\in KMK\end{cases} \tag{4}$$

For example, we take the Watson–Crick complement of the thirteen bases sequence as an example. The sequence of the 3′ end of the 16S rRNA is shown by

rRNA	A	U	U	C	C	U	C	C	A	C	U	A	G
mRNA	U	A	A	G	G	A	G	G	U	G	A	U	C
Position	1	2	3	4	5	6	7	8	9	10	11	12	13

For the rRNA sequence above, the mapping coordinates of (I) are (1,6), (2,4), (3,4), (4,4), (5,4), (6,4), (7,5), (8,3), (9,6), (10,5), (11,2), the mapping coordindates of (II) are (1,6), (2,7),(3,1), (4,2), (5,5), (6,1), (7,2), (8,5), (9,0), (10,4), (11,7), and the mapping coordinates of (III) are (1,2), (2,1), (3,6), (4,5), (5,3), (6,6), (7,4), (8,4), (9,5), (10,3), (11,7). Similarly, the mRNA sequence can also been illustrated using the designed mapping rules (I), (II) and, (III). In this paper, we only care about the average condition of three biological properties not only one characteristic.

The 2D diagrams of these two sequences rRNA and mRNA are illustrated in Fig. 1 using one 2D coordinate plane. The vertical axis denotes the mean value of rules (I), (II) and (III). It can be observed from this figure that crossing and overlapping of the curve by itself are impossible. As we have known that it′s more appropriate to use 2D graphical representation for the reason of degeneracy when the occurrence frequency of the four DNA bases

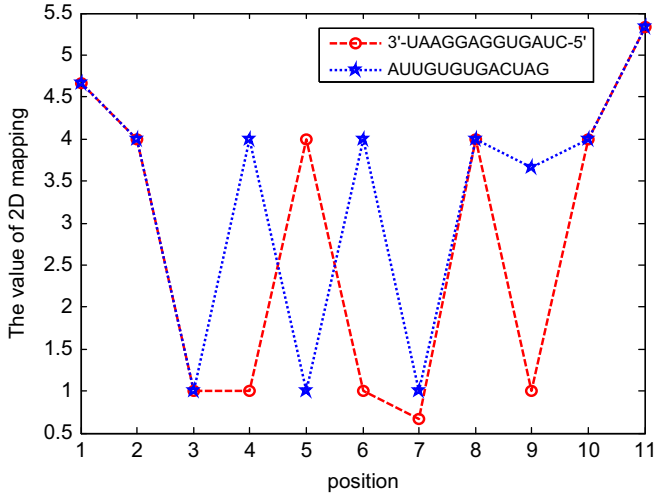

Fig. 1. The mean line representation of the last 13 bases of the 16S rRNA and its complementary sequence under the (I), (II) and (III) mapping rule.

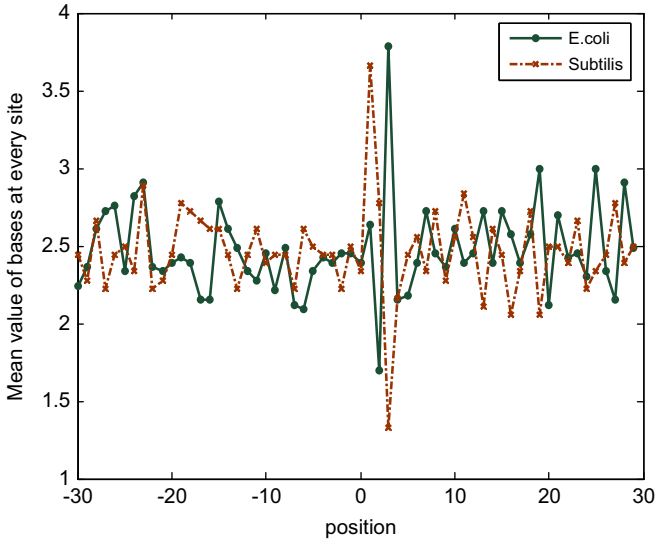

Fig. 2. Curves of the mean bases value in each site of *E. coli* and *B. subtilis* sequences by setting A=1, C=2, G=3, T=4.

A, C, G and T is equal or balanced in a sequence. For detecting the smoothness of DNA sequence strings, we define the mathematic expression for the average values of bases of every site as follows

$$\text{mean base}_i=\frac{\sum_{j=1}^{m}\text{Base value}_{j,i}}{m},\ \ i=1,\ldots,N-2 \tag{5}$$

where m denotes the total number of analyzed sequences ($m=3$, in this paper); i denotes the position of bases N denotes the length of one sequence; $Basevalue_{j,i}$ suggests the defined value of the ith site of the jth sequence. The *meanbase* can reflect the biological properties more clearly using the designed three expression formula (2)–(4). So, it would be more effective for mutation analysis. For visualization, by setting A=1, C=2, G=3, T=4, the mean values of DNA bases corresponding to every site of *Escherichia coli* and *B. subtilis* are shown in Fig. 2 using (5). The zero position is the first base of the initiation codon (usually AUG). It shows that the four DNA bases occur as an equivalent possibility basically. Figs. 1 and 2 provide enough arguments to determine the feasibility of the given DNA representation.

3. Material and methods

In this section, consider the DNA representation, a block code model is presented for analysis of translation initiation. Block codes are referred to be (n, k) codes. And one codebook is constructed through a generation matrix **G** for the purpose of error correction during the decoding process. We select the minimum hamming distance decoding method which has been found in May (2004). However, there is much room for further improvement. A codebook is built by the coding mechanism combined with the mechanism of protein formation instead of choosing the last 13 bases of 16S rRNA, and different sequence corresponds to different codebook. Note that, the determination of codebook in this paper is based on the generation matrix **G** which captures the corresponding biological implications, we also aim at providing an effective and interesting computational methodology.

3.1. The mechanism of protein formation

The formation of protein generally takes place in two stages, namely, transcription and translation. During transcription, the genes in the DNA sequence are used as templates to form the

pre-messenger RNA (pre-mRNA). The pre-mRNA is a polymer formed from four characteristic strings: A, C, G and U. Then, the exons in the pre-mRNA are spliced together to form a polymer of only coding regions known as the mRNA. The mRNA along with the transfer RNA (tRNA) controls the formation of protein. Translation is often regulated by base pairing of the last 13 bases of 16S rRNA. The base pairing is disrupted as ribosome advance through the upstream cistron, thus activating the start site in the downstream. The structure of communication model for the process of gene expression has been given in Dawy et al. (2009).

3.2. Block code model and minimum distance decoder

Since DNA can be denoted by a finite, symbolic sequence, it is natural to extend coding theory to DNA sequence analysis. The mathematical theory of coding is performed by using a set of discrete source symbols based on a finite field (May, 2004). Block codes are referred to be (n,k) codes. A codeword is the output of the block encoder for a given input data block. There are several ways to produce codeword from a k-symbol information sequence. The information symbols are then followed by $(n–k)$ parity symbols. The value of the $(n–k)$ parity symbols is determined by the selected encoding method. In binary codes, each symbol is a bit and can be represented as 0 or 1. A codeword is generated for every k-symbol information sequence. The codebook is the set of all codewords generated by the encoder. A decoder provides a strategy for selecting the transmitted codeword for a given received sequence. There are various decoding methods. In this paper, we select the decoding method based on the minimum hamming distance for the reason that it has corresponding biological properties associated with the energy between mRNA and ribosome. It can help us find out the translation initiation site because of the difference of binding energy between DNA and ribosome for different positions of DNA sequence.

Also, much research has been done by May et al. to study *E.coli* translation initiation sequences by using block coding model in which mRNA is viewed as a noisy encoded signal, and the ribosome is regarded as the decoder. May et al. showed that the block model is effective in recognizing the ribosomal binding site.

3.3. Design of an underlying (n, k) code

With consideration of selecting the triplet codons as units of 2D graphical representation, the number 3 can be regarded as the length of information symbol to produce codeword. We can speculate that the number of parity symbols should be associated with the number 3 which also corresponds to the period-3 characteristic of DNA sequences. Therefore, the possible values of n can be chosen to be a multiple of 3.

For testing the ability of translation initiation recognition, the approximate simulation curves with different values of n are illustrated in Fig. 3 by using the proposed method below with different generating matrix **G**. The horizontal axis denotes the codeword length n. The vertical axis shows the translation initiation position with different value of n, where the translation initiation position should be 0. We can see from Fig. 3 that the codeword length $n=6$ can perform the best recognition of the information about translation initiation and SD regions. Hence, we may speculate about the implied (6, 3) code mechanism existing in DNA. Then, more important left is the design of specific code polynomial. However, Garzon et al. have illustrated a detailed exploration of DNA code design (Garzon and Deaton, 2004). The template method which is a more systematic method using binary strands as templates in a first step to select the poligos to be used (by interpreting 0 as a/t and 1 as c/g); a good error-correcting code from information theory is used in a second

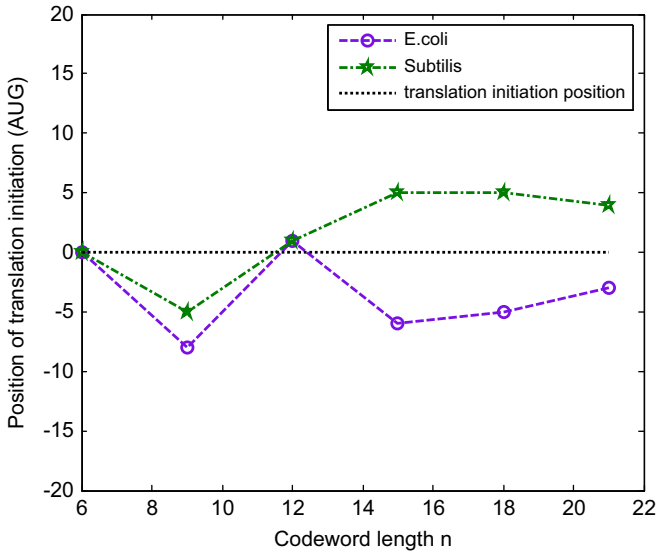

Fig. 3. The test results with different codeword length n, the horizontal axis denotes the codeword length n, and the vertical axis shows the translation initiation position with different n.

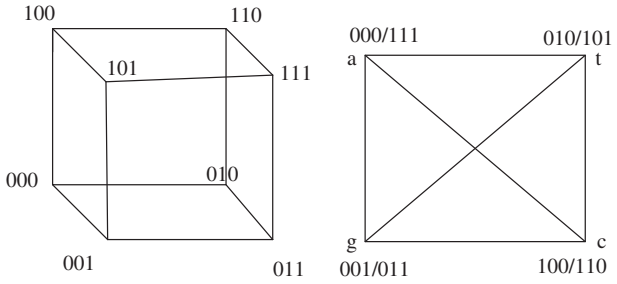

Fig. 4. The quality of the codes obtained using given quantification of the Gibbs energy in the thermodynamic model coming from Garzon and Deaton (2004).

step to disambiguate into actual DNA strands (0 in a codeword at the corresponding place is an a or c, while 1 is read as t or g) (Garzon and Deaton, 2004).

An evolutionary method to develop good codeword sets are discussed which denotes the quality of the codes and best seed codes are constructed in Fig. 4. We can observe that base a is viewed as 000/111, t is 010/101, g is 001/011 and c is100/110. According to theory of Z-curve (Zhang and Zhang, 1994), base pairs combinations A–C, A–T, A–G represent the three types of distribution of the bases along the DNA sequences. For convenience of discussion, we consider the case of a is 111, t is 010, g is 001 and c is 100. Therefore, the parity polynomial can be shown as follows

$$\begin{cases} c_1+c_2+c_3+c_4=0 \\ c_1+c_2+c_3+c_5=0 \\ c_1+c_2+c_3+c_6=0 \end{cases}. \tag{6}$$

Decompose the parity polynomial, the parity matrix is therefore

$$\mathbf{H}=\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}_{3\times 6} \sim \begin{pmatrix} a & c \\ a & t \\ a & g \end{pmatrix}. \tag{7}$$

Hence, the standard generation matrix derived from **H** can be shown by (Schneider, 1999)

$$\mathbf{G}=\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}_{3\times 6} \tag{8}$$

From (7) and (8), we observe that our generation matrix **G** owns the relevant biological implication. In addition, we also provide the possibility of coding theory in biological research from the angle of computational methodology. With the generation matrix **G**, our codebook is then easy to obtain using the formula (Schneider, 1999)

$$\mathbf{C} = m*\mathbf{G} \tag{9}$$

where m denotes the information bits, and **C** is the corresponding codeword. The detailed production of our codebook is introduced in (14) and (15) of Section 3.4

3.4. Analysis of minimum hamming distance decoder

In decoding, a minimum hamming distance decoder is designed to verify the proposed block coding method. Even though the minimum hamming distance method has been found by May et al., there is room for improvement. The studied sequences are composed of thirty bases of the mRNA leader sequence preceding the initiation signal, and twenty-seven bases from the coding region following the initiation signal. Thus, the sequences we studied should satisfy the format shown as follows

$$[b_{-30}b_{-29}\cdots b_{-1}AUG\ \ b_{+3}\cdots b_{+29}] \tag{10}$$

The A of the initiation codon AUG is at the zero position in the sequences. Bases numbered -30 to -1 are part of the leader region of mRNA, and bases numbered $+3$ to $+29$ are part of the coding region. The ribosome covers about thirty bases of the mRNA, therefore, a sixty bases sequence could be sufficient in investigating the region during the translation process. Consider DNA representation (I), (II) and (III), three bases are grouped beginning from the first base in the left obeying the rules depicted as follows

$$r_{-30} = [b_{-30}b_{-29}b_{-28}],\quad r_{-29} = [b_{-29}b_{-28}b_{-27}],\ \ ...,$$
$$r_{+27} = [b_{+27}b_{+28}b_{+29}] \tag{11}$$

For convenience of analysis based on communication theory, we have performed the binary conversion by using (12) shown as follows

$$0\to\begin{pmatrix}0\\0\\0\end{pmatrix}, 1\to\begin{pmatrix}0\\0\\1\end{pmatrix}, 2\to\begin{pmatrix}0\\1\\0\end{pmatrix}, 3\to\begin{pmatrix}0\\1\\1\end{pmatrix}, 4\to\begin{pmatrix}1\\0\\0\end{pmatrix},$$
$$5\to\begin{pmatrix}1\\0\\1\end{pmatrix}, 6\to\begin{pmatrix}1\\1\\0\end{pmatrix}, 7\to\begin{pmatrix}1\\1\\1\end{pmatrix} \tag{12}$$

Then a DNA sequence corresponds to a binary matrix defined as **T** (which is a size of $3\times(N\text{-}2)$) given by

$$\mathbf{T} = [t_{-30}, t_{-29}, \ldots, t_{+27}] \tag{13}$$

In our method of decoding, one sequence corresponding to a codebook which is defined as follows

$$\mathbf{C}_j = \begin{pmatrix} c_{-30}\\ c_{-29}\\ \vdots\\ c_{+27}\end{pmatrix},\quad j = 1,2,\ldots,num \tag{14}$$

where j denotes the jth sequence, num is the number of total sequences and the element c_i can be obtained by using the expression shown as follows

$$c_i = t_i'*\mathbf{G}, i = -30,-29,\ldots,+27 \tag{15}$$

where t_i' is the transpose of t_i. The received sequence should be an n-element subset of the analysis sequence. For instance, the first two received sequences are shown as follows

$$r_{-30} = [t'_{-30}, t'_{-29}] \tag{16}$$

$$r_{-29} = [t'_{-29}, t'_{-28}] \tag{17}$$

In the sense that hamming distance is used to determine how close the received sequence is to the codeword in the codebook, the minimum distance $d_{\min}$ of a received sequence is defined as follow

$$d_{\min} = \min\{(r_p, C)\}, p = -30,-29,\ldots,+26 \tag{18}$$

where p represents the relative position to the translation initiation, and ***C*** denotes a codebook of one sequence generated by using (15). This distance is used to evaluate how well the proposed coding method captures the biological aspects of the translation process.

The decoder stores the minimum distance for each sequence group in matrices of the form as follow

$$\mathbf{D} = \begin{pmatrix} d\min^1_{-30} & d\min^1_{-29} & \cdots & d\min^1_{numValid}\\ \vdots & \vdots & \vdots & \vdots\\ d\min^n_{-30} & d\min^n_{-29} & \cdots & d\min^n_{numValid}\end{pmatrix} \tag{19}$$

where n is the total number of the analyzed sequences, and *numValid* is the last valid comparison position on the sequence. In this work, the *numValid*=26.

In May et al., averaging is a standard signal processing technique used to enhance a signal in the presence of noise, therefore, we calculate the average value by using (20) illustrated as follows

$$\mathbf{D}_{average}(p) = \frac{1}{n}\sum_{k=1}^{n} d\min^k(p),\quad p = -30,-29,...,numValid \tag{20}$$

And smaller distance values mean the stronger hydrogen bond formations between the 16S rRNA and the mRNA.

Based on the analysis above, we summarize the proposed method for evaluating translation efficiency of DNA as the following steps

Step 1. Obtain the test sequences of *Escherichia coli* and *B.subtilis* according to (10), and map the test DNA sequences into 2D-numeric sequences.
Step 2. Convert the numeric sequences from step1 into binary sequence matrix **T** using (12).
Step 3. Form the codebook of every sequence by using (15) combined with the generation matrix **G**.
Step 4. Obtain the minimum hamming distance matrix **D** through repetitive comparisons by using (18).
Step 5. Calculate the average value of every column of **D** as the output.

4. Results and comparisons

The discussed method above is applied to *Escherichia coli* and *B.subtilis* for explaining the biological relevance of genetic regulatory system in this section. Furthermore, the published single base and multiple bases mutations are introduced to the SD sequences for investigating the translation efficiency applying the proposed method. And in the following, we compare the experimental results obtained by multiple bases mutations of SD sequences with two previous methods for proving the validity of the presented method.

4.1. Application to prokaryotic organisms

The simulations are designed to verify the proposed method by using the newest *Escherichia coli* (13000 bp) and *B.subtilis* (12060 bp)

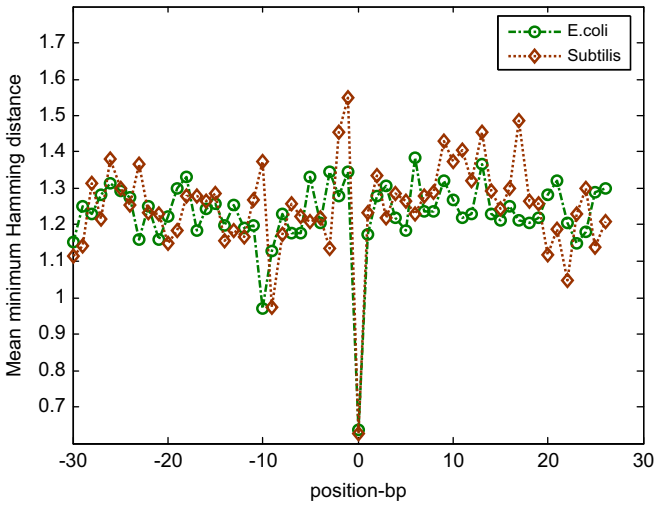

Fig. 5. The protein translation process with original sequences of *E. coli* and *Subtilis*.

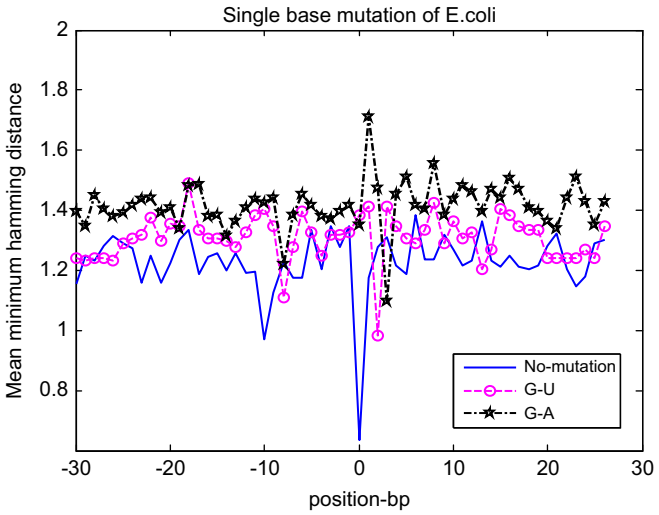

Fig. 6. The results with Jacob mutations G→U and G→A of *E. coli* in the translation process.

translated sequences downloaded from the GENEBANK database on the internet site "www.ncbi.nlm.nih.gov".

The protein translation process curves of *Escherichia coli* and *B.subtilis* are described in Fig. 5, where the horizontal axis is the position relative to the first base A of the initiation codon. And zero on the horizontal axis corresponds to the first base A of the initiation codon. The vertical axis shows the mean minimum hamming distance. From Fig. 5, we can obtain three kinds of information: (1) Recognition of biologically significant regions within the mRNA leader sequence. (2) Indication and recognition of the open reading frame. (3) Distinction between *Escherichia coli* and *B.subtilis.*

In detail, two minimum distances peak occur between the 0 and −10 region within the mRNA leader sequence. The first peak position is zero (translation initiation AUG), and the second is just the ribosomal binding site (SD sequence region) between −5 and −10 region. It's corresponding to the biological characteristics that smaller distance values indicate stronger hydrogen bond formations between the 16S rRNA and the mRNA.

4.2. Single base mutations

A single base change (point mutation) in the complementary sequence of 3′ end of 16S rRNA has previously been shown to have a dramatic effect on protein synthesis in *E. coli (*Jacob et al., 1987*)*. According to the degeneracy of codons, it has an advantage of decreasing the influence of mutations which promotes the stabilization of mutated species. And the influences of single base mutations still confused our biologist and medical scientists.

The proposed error-correcting encoding model has successfully recognized the existence of single base mutation. A single base change in 16S rRNA (C→G) has previously been shown to have a dramatic effect on protein synthesis in *E. coli* (Prescott and Goringer, 1990).The SD sequence is a region of rich purine. Thus, we select the single base mutations (G→U, position 5) and (G→A, position 6) as examples to exploit our method further.

Fig. 6 illustrates the translation process of *E.coli* under the single base mutations G→U and G→A of SD reserved sequence. It shows the incomplete loss of translation initiation signal and the SD region. It can be inferred from the plot that the levels of protein production will be reduced but not completely stopped. It's also consistent with the published experimental results (Jacob et al., 1987).

Table 1
Three possible groups of gene mutations.

Types	Mutations
Purine-Pyrimidine	A→G, G→A, C→T, T→C
Amino-Keto	A→C, C→A, G→T, T→G
Strong-weak H bond	A→T, T→A, G→C, C→G

4.3. Multiple bases mutations

Based on the four bases A, T, G and C, we classify the gene mutation into three possible groups shown as Table 1 according to the chemical information: R/Y, S/W and M/K. Two published mutation types (position 4–8) of SD sequence at the 3′ end of the 16S rRNA are shown as follows:

$$\begin{cases} I : GGAGG \rightarrow GUGUG(mutation) \\ II : GGAGG \rightarrow CCUCC(mutation) \end{cases} \quad (21)$$

Theoretically, the results of both mutations were lethal for the organism in the sense that the production of protein stopped (Hui and De Boer, 1987). However, the results are surprised when these published mutations are tested by using the proposed method.

The mutated curves of the mean hamming distance of *E. coli* and *B. subtilis* are illustrated as Figs. 7 and 8 respectively. From Figs. 7 and 8, we can observe that mutations have little or no influence on translation initiation at both start sites. It just proves the feasibility of our method in recognition of translation initiation without considering multiple bases mutations. Also, it demonstrates that multiple bases mutations of SD sequence can not inhibit the natural operation of translation initiation. It is consistent with theory that triplet codons play a good role in making the whole genome more stable in biological system.

For comparison, we have analyzed two researchers' methods: (1) an evolutionary block encoding model based on minimum hamming distance decoder by using the last thirteen bases of 16S rRNA (May, 2004); (2) a minimum free energy method by using the same parity bases as May (Dawy et al., 2009).

May's method based on the last 13 bases of 16S rRNA is illustrated in Fig. 9. And Dawy's minimum free energy based method is shown in Fig. 10. We can see from Figs. 9 and 10 that the translation initiation site is at the position +1(May) and

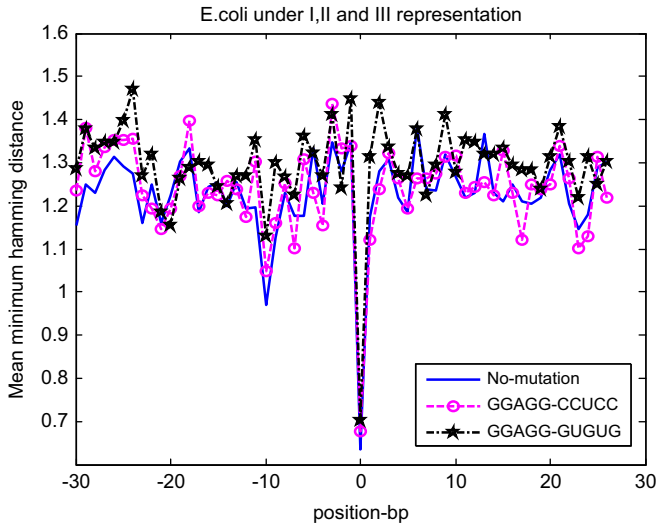

Fig. 7. The results of *E. coli* with Hui and De Boer mutations in the translation process.

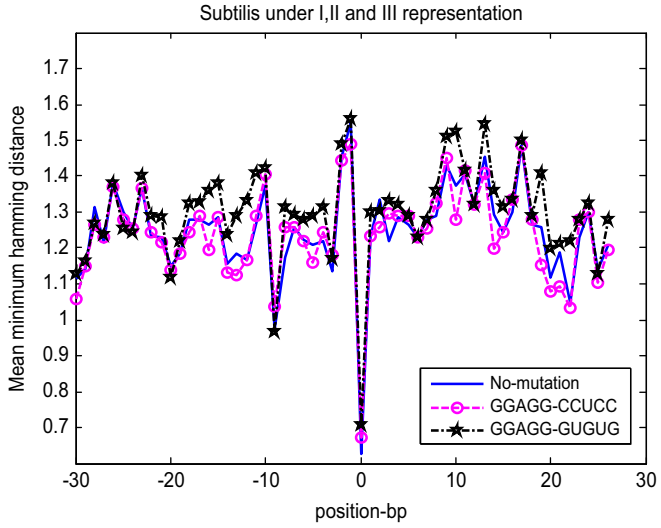

Fig. 8. The results of *Subtilis* with Hui and De Boer mutations in the translation process.

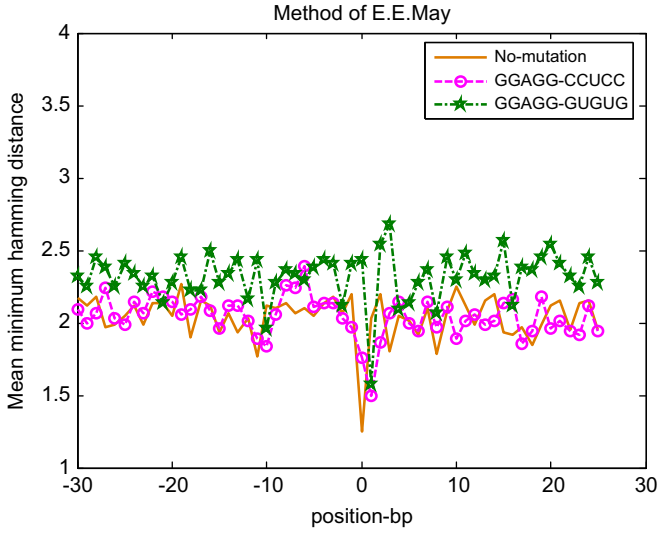

Fig. 9. The results with Hui and De Boer mutations under the method of E.E. May.

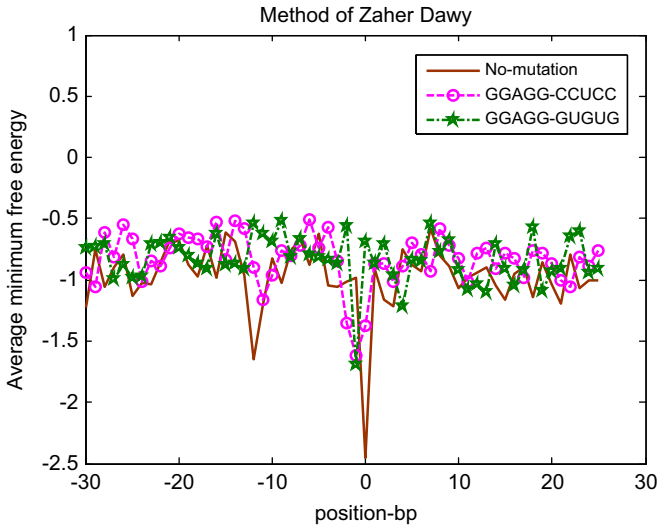

Fig. 10. The results with Hui and De Boer mutations under the method of Zaher Dawy.

-1(Dawy) which is not in phase with the non-mutated sequences at the translation initiation position. The SD region signal (position -11) and the translation initiation signal (position 0) have been lost partly so that we can't recognize them correctly. The results of May and Dawy show that multiple bases mutation has a strong negative influence on the recognition of the SD signal and translation initiation signal.

The significant differences between the two methods and our work mainly contain two aspects: (1)The construction of model and the selection of codeword. The codeword in our work is based on a generation matrix **G** (determined by experiments) based on communication theory; (2) The DNA representation of our work is combined with the function of triplet codons resisting mutations in the biological system.

5. Conclusions

The proposed method for analysis of translation efficiency based on coding theory can accurately obtain key information such as the recognition of biologically significant regions within the mRNA leader sequences, the precise location of translation initiation and the recognition of the open reading frames in the protein translation process. However, the advantage of the proposed method is that it allows the existence of multiple bases mutations of SD sequence which has played a crucial role in the process of translation. Seemingly, the results of multiple bases mutation based on the proposed method is violated with the published investigations, however, our model may be an effective method to scientists engaged in medication research and biological technology which is also associated with our previous motives of research. Although single base mutation can't be corrected in the process of translation, it further proved the biological relevance and validity of the proposed method for its congruence with published investigations. Therefore, we can speculate about the importance of the presented method in the improvement of protein translation efficiency and its wide prospect in drug design.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (CDJXS10160001) and the Central

University Postgraduate' Science and Innovation Funds of China (CDJXS10161114) and the authors are grateful to anonymous referees for their corrections, suggestions and improvements.

References

Bataineh, M.A., Huang, L., Alonso, M., Menhart, N., Atkin, G.E., 2011. Analysis of gene translation using a communications theory approach. Advanced Computational Biology 680, 387–397.

Dreyfus, M., 1988. What constitutes the signal for the initiation of protein synthesis on *Escherichia coli* mRNAs? Journal of Molecular Biology 204, 79–94.

Dawy, Z., Morcos, F., Weindl, J.W., Mueller, J.C., 2009. Translation initiation modeling and mutational analysis based on the 3′ end of the *Escherichia coli* 16S rRNA sequence. BioSystems 96, 58–64.

Faxen, M., Plumbridge, J., Lsaksson, L.A., 1991. Codon choice and potential complementarity between mRNA downstream of the initiation codon and bases 1471 –1480 in 16S ribosomal RNA affects expression of glnS. Nucleic Acids Research 19, 5247–5251.

Fargo, D.C., Zhang, M.G., Boynton, J.E., 1998. Shine–Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. Molecular Genetics 257, 271–282.

Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Bourget, J.A., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., Lee, J.H., Loh, Y.H., Manos, P.D., Montserrat, N., Panopoulos, A.D., Ruiz, S., Wilbert, M.L., Yu, J., Kirkness, E.F., Belmonte, J.C.I., Rossi, D.J., Thomson, J.A., Eggan, K., Daley, G.Q., Goldstein, L.S.B., Zhang, K., 2011. Somatic coding mutations in human induced pluripotent stem cells. Nature 471, 63–67.

Garzon, M.H., Deaton, R.J., 2004. Codeword design and information encoding in DNA ensembles. Natural Computing 3, 253–292.

Gupta, M.K., 2006. The quest for error correction in biology. IEEE Engineering in Medicine and Biology Magazine 25, 46–53.

Hui, A., De Boer, H.A., 1987. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. Proceedings of the National Academy of Sciences 84, 4762–4766.

Howden, S.E., Gore, A., Li, Z., Fung, H.L., Nisler, B.S., Nie, J., Chen, G., Mclntosh, B.E., Gulbranson, D.R., Diol, N.R., Taapken, S.M., Vereide, D.T., Montgomery, K.D., Zhang, K., Gamm, D.M., Thomson, J.A., 2011. Genetic correction and analysis of induced pluripotent stem cells from a patient with gyrate atrophy. Proceedings of the National Academy of Sciences of the United States of America 19, 6537–6542.

Jacob, W.F., Santer, M., Dahlberg, A.E., 1987. A single base change in the Shine Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. Proceedings of the National Academy of Sciences 84, 4757–4761.

Kini, R.M., Chinnasamy, A., 2010. Nucleotide sequence determines the accelerated rate of point mutations. Toxicon 56, 295–304.

Lagerqvist, A., Hakansson, D., Lundin, C., Prochazka, G., Dreij, K., Segerback, D., Jernstrom, B., Tornqvist, M., Frank, H., Seidel, A., Erixon, K., Jenssen, D., 2011. DNA repair and replication influence the number of mutations per adduct of polycyclic aromatic hydrocarbons in mammalian cells. DNA Repair 10, 877–886.

May, E.E., 2004. Coding theory based models for protein translation initiation in prokaryotic organisms. BioSystem 76, 249–260.

May, E.E., Mladen, A.V., Donald, L.B., David, I.R., 2004. An error-correcting code framework for genetic sequence analysis. Journal of the Franklin Institute 341, 89–109.

Mian, I.S., Rose, C., 2011. Communication theory and multicellular biology. Integrative Biology 3, 350–367.

Mac Dónail, D.A., 2006. Digital parity and the composition of the nucleotide alphabet. IEEE Engineering in Medicine and Biology Magazine 25, 54–61.

Mori, K., Saito, R., Kikuchi, S., Tomita, M., 2007. Inferring rule of *Escherichia coli* translational efficiency using an artificial neural network. BioSystems 90, 414–420.

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes. Current Science 66, 309–314.

Oliver, H., Mariam, B., Jurgen, B., Beatrix, S., Jorg, S., 2009. A novel mechanism for translation initiation operates in haloarchaea. Molecular Microbiology 71, 1451–1463.

Prescott, C.D., Goringer, H.U., 1990. A single mutation in 16S rRNA that affects mRNA binding and translation-termination. Nucleic Acids Research 18, 5381–5386.

Roman, J., 1995. The Theory of Error-Correcting Codes. Springer-Verlag, Berlin.

Rosen, G.L., 2006. Examining coding structure and redundancy in DNA. Medical Biology 6, 0739–5175.

Randic, M., Vracko, M., Novic, M., Plavsic, D., 2009. Spectrum-like graphical representation of RNA secondary structure. Internatinal Journal of Quantum Chemistry 109, 2982–2995.

Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters 368, 1–6.

Randic, M., Zupan, J., Balaban, A.T., Vikic-Topic, D., Plavsic, D., 2011. Graphical representation of proteins. Chemical Reviews 111, 790–862.

Shine, J., Dalgarno, L., 1974. The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementary to non-sense triplets and ribosome binding sites. Proceedings of the National Academy of Sciences 71, 1342–1346.

Stenstrom, C.M., Jin, H., Major, L.L., Tate, W.P., 2001. Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. Gene 263, 273–284.

Sunthornwat, R., Moore, E.J., Temtanapat, Y., 2011. Detection and classifying mutations in genetic code with an application to ß-thalassaemia. ScienceAsia 37, 51–61.

Sweeney, P., 1991. Error Control Coding: an Introduction. Prentice Hall, New York, NY.

Schneider, T.D., 1997. Information content of individual genetic sequences. Journal of Theoretical Biology 189, 427–441.

Schneider, T.D., 1999. Measuring molecular information. Journal of Theoretical Biology 201, 87–92.

Wielgoss, S., Barrick, J.E., Tenaillon, O., Cruveiller, S., Woon Ming, B.C., Médigue, C., Lenski, R.E., Schneider, D., 2011. Mutation rate inferred from synonymous substitutions in a long term evolution experiment with *Escherichia coli*. Genes, Genomes, Genetics 1, 183–186.

Zhang, R., Zhang, C.T., 1994. Z curve, an intuitive tool for visualizing and analyzing the DNA sequences. Journal of Biomolecular Structure and Dynamics 11, 767–782.