# Modeling and Analysis of Gene Expression Mechanisms: A Communication Theory Approach

Zaher Dawy[1,3], Faruck Morcos González[1], Joachim Hagenauer[1], and Jakob C. Mueller[2]

[1]Munich University of Technology, Institute for Communications Engineering (LNT), Arcisstr. 21, 80290 Munich, Germany
[2]National Research Center for Environment and Health (GSF), Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany
[3]American University of Beirut, Department of Electrical and Computer Engineering, Beirut, Lebanon
Email: zaher.dawy@aub.edu.lb, faruckm@mytum.de, hagenauer@tum.de, jakob.mueller@gsf.de

*Abstract*— **The increase in the availability of genetic data in the last years is prompting the efforts to use tools from communications engineering for the understanding of genetic information. Processes in molecular biology can be modeled through the use of these tools. A communication theory based model for the process of translation in gene expression is proposed. The model is based on the assumption that the ribosome decodes the mRNA sequences using the 3' end of the 16SrRNA molecule as a one dimensional codebook. The biological consistency of the model is proven in the detection of the Shine-Dalgarno signal and the initiation codon for translation initiation. Furthermore, implications on the role of the 16SrRNA 3' end in the complete process of prokaryotic translation are presented and discussed. Interestingly, the obtained results lay out the possibility of an interaction of this part of the ribosome in the process of translation termination. Finally, results obtained via the proposed model are compared with published experimental results for different mutations of the rRNA molecule. Total agreement between both sets of results prove the validity of the proposed model. By means of simulated mutations in the last 13 bases of the 16SrRNA, a global analysis of this part of the ribosome in the process of translation is established. This work illustrates the relevance of communication theory based models for genetic regulatory systems.**

## I. Introduction

Yockey and Schneider pointed out the ambivalence between biological interactions in genetic systems and certain aspects in communication systems [1][2][3]. Schneider states that modeling biological systems using concepts from information theory can lead to a better understanding of the accuracy, mechanisms, and evolution of the molecular machines. In order for the living system to survive, Battail and Eigen suggest the necessity of error correcting capabilities in the replication of the DNA [4][5]. Taking all this into consideration, May introduced channel coding models for the process of translation initiation [6][7]. The work of May established some first ideas of modeling the DNA interactions based on coding theory. Additionally, the presented approach paved the way to continue working in the search for new and better models to describe the coding-decoding relations between DNA-mRNA-rRNA molecules.

The main contributions of this work are: i) develop an analogy between information transmission in communications engineering and gene expression, ii) develop and validate a novel biologically-motivated coding model for the process of prokaryotic translation initiation, iii) use this model to gain new insights on the biological interactions between the ribosome and the mRNA, and iv) use this model to test the effect of mutations in the ribosome on protein synthesis. This work is done with the objective of stimulating the interdisciplinary research effort to apply techniques from communications engineering to problems in the area of biology.

In Section II a general theoretical background is presented establishing an analogy between gene expression and communications engineering. Section III presents the construction of a coding model for the process of translation initiation in the *E. coli* organism. In Section IV, the results of implementing this model are presented including analysis and insights. Finally, Section V draws some conclusions.

## II. Analogy

Gene expression is the process in which the information contained in the DNA molecule is transformed into proteins. Proteins are sequences of small molecules called amino acids. These protein products will later be used for different vital processes in the living system. The accuracy of this process is related to the survival of the organism.

Gene expression involves two main stages: *transcription* where the information stored in the DNA is transformed into the messenger RNA (mRNA) and *translation*, where the mRNA molecule serves as an instructive for protein synthesis. When the process of gene expression is analyzed, many similarities with the way engineers transmit digital information come into view. We model gene expression using elements from communication engineering as depicted in Fig. 1.

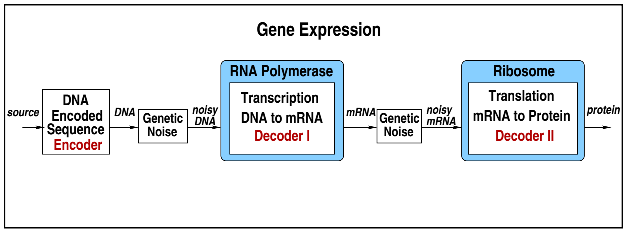

Fig. 1. Communications model for the process of gene expression.

In this model, an unknown source produces the information in the DNA message. Then a process of channel encoding (some sort of serial concatenation) is used to create the structure of bases of the DNA sequence. Once the DNA is

0-7803-8938-7/05/$20.00 (C) 2005 IEEE

released (channel I), it is exposed to noise that may produce mutations in the sequence. The process of transcription done by the macromolecule RNA polymerase is expressed as a decoder. This decoder takes the DNA sequence and decodes it to produce the mRNA. The argument that the RNA polymerase is actually decoding and not encoding is supported by the fact that the mRNA sequence is shorter, thus, some redundancy is removed. The resulting mRNA (also called mature mRNA) contains only the exons or protein coding regions (message) whereas the introns (redundancy) are sliced. At this point the mRNA molecule is again exposed to thermal noise and radiations, especially when it travels outside the nuclear membrane (channel II) in eukaryotic organisms. Once the mRNA reaches the ribosome, a second decoding process takes place. Subsequently, the ribosome will take the mRNA sequence to start the protein synthesis. The protein output of our model is the final recovered message.

This work focuses on the modeling of translation in gene expression. It is important to mention that the mechanism of translation is different for prokaryotes and eukaryotes [8]. We are interested in how translation initiates in prokaryotes, more specifically in the organism *E. coli*. Translation involves the chaining of amino acids which form proteins. In order to do this, the ribosome binds to the messenger RNA to create a closed complex. The ribosome is able to "scan" the mRNA in the search for sequences that contain a sign to start translation.

## III. Coding Model for *E. coli* Translation

In translation initiation , the ribosome binds to what is called the leader region in the mRNA sequence. The leader region is composed by the bases upstream of the initiation codon. The initiation codon, typically AUG, marks the start of a coding region that is the part of the mRNA that will code for a protein. A typical structure of a mRNA sequence is shown in Fig. 2.

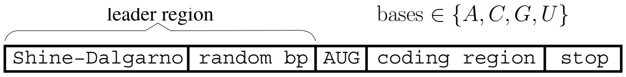

Fig. 2. Structure of mRNA sequence (bp stands for base pairs).

The ribosome recognition of the signals in the leader region, the initiation codon as well as the creation of the initiation complex, are modeled using elements from communications engineering as shown in Fig. 3.

The noisy mRNA is the input of the system. In the leader region decoding, the ribosome decodes the mRNA leader region in order to detect a signal that will enable the ribosome to start translation. If it does not find the signal then the ribosome keeps scanning the mRNA. Once the signal is found, the ribosome will try to start translation. When the synchronization signal (AUG) is found, the process of elongation starts where the ribosome uses the genetic code to start demapping (demodulating) the triplets (codons) in the mRNA sequence into a chain of amino acids. In the last stage of the model, the ribosome recognizes one of the stop codons (UAA, UAG, UGA) and then the protein production complex is liberated from the mRNA resulting in the production of the protein. In the sequel, we concentrate on how the ribosome "decodes" the leader region of the mRNA in order to get the signals to start translation.

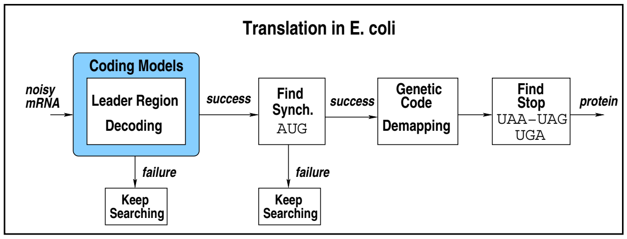

Fig. 3. Communications model for the process of translation initiation.

### A. Model Construction

In order to start translation, the last 13 bases of the molecule 16SrRNA inside the small subunit of the ribosome interact with the leader region of the mRNA [9]. This interaction is based on the hydrogen bonding between nucleotide bases in the genetic sequences. This principle allows the bonding of cytosine (C) with guanine (G) and adenine (A) with thymine (T) (or uracil (U) in case of RNA). This chemical interaction permits the recognition of signals between DNA and RNA molecules. As in other translation initiation models [7], the coding model presented in this work is based on the structure of the 16SrRNA and more specifically on its 3' end.

A common assumption made in biology is that the ribosome recognition of the initiation signal in the leader region is achieved when the so called Shine-Dalgarno sequence is found [10]. A consensus sequence of the Shine-Dalgarno (SD) is AGGAGG. It is obtained by calculating the most frequent nucleotide for each position of aligned mRNA sequences. In the case of the SD sequence, the base A in the first position is the base with the highest frequency among all aligned sequences. The approach of assigning the recognition of binding sites just to one sequence is in general wrong as it discards all the variability of the sequences and introduces "hard" decisions that incur a loss of information [11]. In order to improve the accuracy of the model, a more flexible approach for the translation initiation mechanism is needed. The approach taken in this work is the use of error correction coding theory.

### B. 16SrRNA Based Codebook

The primary assumption of our model is that the ribosome has a method to verify if a mRNA sequence contains the initiation signals: Shine-Dalgarno signal and initiation codon signal. This method is based in the existence of an embedded codebook in the ribosome molecular structure. We propose that the last 13 bases of the 16SrRNA conformation are constructed in such a way that they create a one dimensional codebook. The ribosome uses this highly conserved sequence (among prokaryotes) in the 16SrRNA molecule to do the comparisons needed to detect the relevant signals. We want to use a codebook to detect translation signals (we care for error detection capability and not error correction). To build the codebook, a codeword length $N = 5$ is assumed and the codewords are obtained by taking a sliding window through the Watson-Crick complement of the sequence of 13 bases. The length $N = 5$ is selected after testing the algorithm with different codeword lengths as it has shown good performance in various scenarios. Having $N = 5$ and a length of 13 bases, the codebook is obtained via a sliding window operation which results in 9 codewords. The complement sequence is shown below:

```
5′ U A A G G A G G U G A U C ... 3′
```

The codewords are taken from the complement because those are the words that will be found in the mRNA molecule. The codebook constructed for this model is shown in Table I.

TABLE I

16SrRNA BASED CODEBOOK.

| $C_l$ | Codeword |
|---|---|
| $C_1$ | U A A G G |
| $C_2$ | A A G G A |
| $C_3$ | A G G A G |
| $C_4$ | G G A G G |
| $C_5$ | G A G G U |
| $C_6$ | A G G U G |
| $C_7$ | G G U G A |
| $C_8$ | G U G A U |
| $C_9$ | U G A U C |

A sliding window is applied on the received noisy mRNA sequence to select sub-sequences of length $N$ and compare them with all codewords in the codebook. The codeword that results in the minimum distance metric is selected and the metric value is saved. Biologically, the ribosome achieves this by means of the complementary principle. The energetics involved in the rRNA-mRNA interaction tell the ribosome when a signal is detected and, thus, when the start of the process of translation should take place. In our model, the method of *free energy* doublets presented in [12] is adopted to calculate a free energy distance metric in kcal/mol instead of minimum distance (see Table II). The proposed modeling of ribosome decoding is summarized in Algorithm 1. The codeword yielding the minimum free energy, i.e. most complementary, will be the valid codeword. The minimum energies are stored and plotted in order to show the performance of the algorithm.

### C. Tested Sequences

In order to test our model, sequences of the complete genome of the prokaryotic bacteria *E. coli* strain MG1655 were obtained. These sequences are available in the National Center for Biotechnology Information [13]. Coding regions (CDS) are annotated and its respective protein translation is described.

TABLE II

ENERGY TABLE (KCAL/MOL) [12].

| Free Energy Doublets | | | |
|---|---|---|---|
| AA -0.9 | AG -2.3 | GA -2.3 | GG -2.9 |
| AU -0.9 | AC -1.8 | GU -2.1 | GC -3.4 |
| UA -1.1 | UG -2.1 | CA -1.8 | CG -3.4 |
| UU -0.9 | UC -1.7 | CU -1.7 | CC -2.9 |

**Algorithmus 1** Free Energy Ribosome Decoding

*Given:* Codebook $\mathcal{C}$ with $L$ codewords of length $N$ and a subsequence $S$ of length $N$ from received noisy mRNA sequence. *Notation:* $c_n^l$ is the $n$th symbol of codeword $l$, $s_n$ is the $n$th symbol of $S$, $E_l$ is the free energy when codeword $l$ is used ($E_l$ initialized to 0, $0 \leq l \leq L$), and $E_{a,b}$ is the energy dissipated on binding with the nucleotide doublets $ab$ (see Table II, e.g. the energy dissipated by binding with GC is $-3.4$ kcal/mol).

```
Minimum Free Energy
  for $l = 1 \dots L$ do
    for $n = 1 \dots N-1$ do
      if $c_n^l c_{n+1}^l$ are complementary with $s_n s_{n+1}$ then
        codeword energy $E_l = E_l + E_{c_n^l, c_{n+1}^l}$
      else
        $E_l = E_l$
      end if
    end for
    if $E_l$ is less than $E_{l-1}$ then
      valid codeword $C_l$
      minimum codeword free energy $E_{\min} = E_l$
    end if
  end for
```

## IV. RESULTS, ANALYSIS, AND INSIGHTS

The 16SrRNA codebook based model uses principles from error correcting codes to explain the behavior of the biological regulatory systems. The assumptions made for the construction of the model consider the biological molecular interactions as we are modeling a biological process and not a purely computational process. In this section, we prove the validity of the proposed model and we demonstrate its usefulness in pointing out interesting and new biological insights related to the process of translation in gene expression.

### A. Translation Signals

In the search for biological completeness, the proposed model is tested for the whole process of translation to seek insights of how the ribosome performs during the rest of the stages of translation: elongation and termination. By doing this, important realizations are obtained on the behavior of this gene expression mechanism. For the analysis, 1568 sequences of coding regions greater or equal than 500 base pairs are taken from the *E.coli* genome. In addition, sequences of the same length taken arbitrarily (non-coding regions) from the genome are used for comparison. The algorithm is applied to these sequences and average results are plotted in Fig. 4.

The x-axis represents the position in the selected and aligned sequences. For presentation purposes, the positions of the initiation and termination codons for all coding sequences are fixed at 101 and 398, respectively, thus, only 294 nucleotides from the coding region are kept while the others are removed. The y-axis represents the calculated average free energy measure in kcal/mol for each position in the mRNA. The results found are significant. The Shine-Dalgarno and the initiation codon signals are identified in the translated sequences while they are

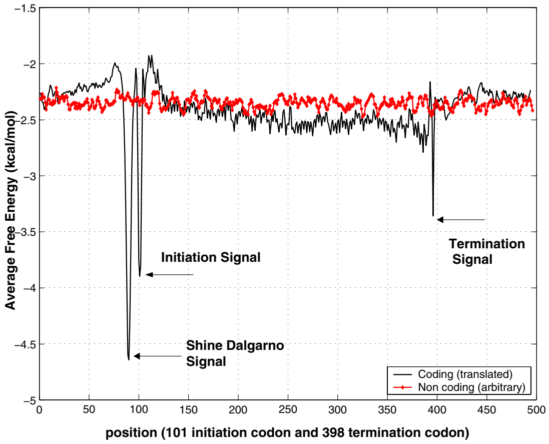

Fig. 4. Detected translation signals.

not in the arbitrary sequences, just as is expected from a model for translation initiation. Furthermore, in the coding region part of the translated sequences, one can see a mean energy performance which demonstrates a steady state in the process of elongation. Finally, a remarkable fact is that the model was additionally able to recognize the presence of the termination codon. As a result, the plot presents an overview of the complete process of translation. First a detected signal tells the ribosome to slow down because a coding region is approaching (this was also realized in [8]). Second, a synchronization flag is detected as shown in the initiation signal. Once the translation complex has been constructed, the process of elongation starts and a steady state takes place in the coding region. Finally, a third detected region informs the ribosome that the protein (message) has been synthesized (decoded).

### B. *Role of 16SrRNA: Evolutionary Hypothesis*

The translation signals found led us to the hypothesis that the last 13 bases sequence of the 16SrRNA have a broader role in the process of translation, i.e. not only initiation but also elongation and termination. Although there have been some suggestions that the 16SrRNA is involved in the process of termination [14], we have no reference that pointed out the involvement of the 3' end of the 16SrRNA. Hence, this is a novel finding that extends the significance of this structure. It is believed that the earliest form of information unit in molecular biology was the RNA and not the DNA [15]. Furthermore, the characteristics of translation initiation in eukaryotes where no SD signal is used [8] indicate that the one simple ancestral recognition system which controls several steps in the translation process was replaced by several more sophisticated systems in eukaryotes.

As a result, an evolutionary point of view on the role of the 3' end of the 16SrRNA in all the stages of translation is proposed. Initially, it was responsible for detecting the SD signal, detecting the initiation signal, monitoring elongation, and finally detecting the termination signal. Later, with evolution, its role became less relevant with the introduction of initiation and release factors that work with the ribosome to translate the mRNA. Afterwards, the SD signal was not needed any more, for eukaryotes, and the use of the last 13 bases to solely terminate translation was replaced by the use of a release factor.

### C. *Mutations*

Experimental results obtained by mutating regions of the 3' end of the 16SrRNA are compared with results obtained by incorporating these mutations in the 16SrRNA based codebook of our model. Jacob introduced a point mutation in the the 5th position of the 16SrRNA [16]. Specifically, the 5th position in the arrangement illustrated below:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Base | 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 |

This point mutation consisted in a change of the nucleotide C → U in the ribosome small subunit. This is equivalent to make a mutation from G → A in the complement sequence shown above. The result of this mutation was a reduction in the level of protein synthesis. Another published record of the behavior of the protein synthesis under mutations in the 3' end of the 16SrRNA, was done by Hui and De Boer [17]. In this experiment, the mutations were done in positions 4 to 8 (GGAGG → CCUCC) and positions 5 to 7 (GAG → UGU). The results of both mutations were lethal for the organism in the sense that the production of proteins stopped.

These published mutations are tested using our model. First the mutations as specified in [16] and [17] are performed in the 13 bases. For each case, the codebook is constructed based on the mutated sequence. The resulting "mutated" decoder is used in the algorithm and the response of the system is observed. Fig. 5 shows how the recognition of the Shine-Dalgarno signal is affected for the Jacob mutation. It can be inferred from the plot that the levels of protein production will be reduced but not completely stopped. After introducing the mutations as in [17], the results showed a complete loss of the SD signal. Hence, it can be inferred that the translation will never take place. This is illustrated in Fig. 6. Note that results obtained by mutations in the 16SrRNA also apply to scenarios with mutations in the mRNA at corresponding positions.

These results are completely consistent with the published experimental results. This demonstrates the relevance of our model, its biological accuracy, and its flexibility to incorporate and study structural changes. Moreover, a laboratory work that usually takes months was simplified through the introduction of mutations to our model.

To exploit our model further, we have introduced point mutations in all positions of the last 13 bases of the 16SrRNA molecule in order to study their influence on the process of translation. The obtained results are summarized in Table III by quantizing into 5 levels the influence of these mutations on each of the translation signals (SD, initiation, stop). The levels are: – represents no influence in the recognition of the signal, ⇓ represents a strong negative influence, ↓ a weak influence, ↑ a weak positive influence, and ⇑ a strong positive influence.

For example, results show how a mutation in position 5 has a strong negative influence in the recognition of the SD signal,

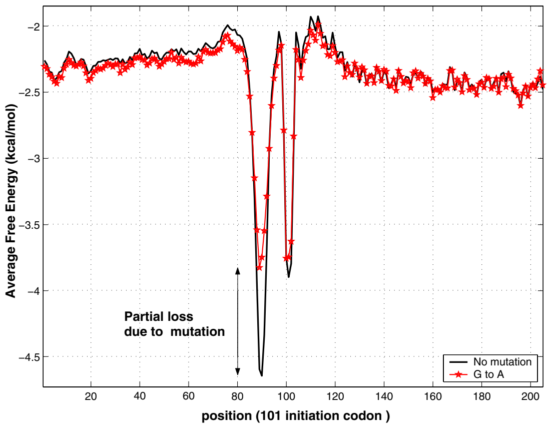

Fig. 5. Results with Jacob mutation.

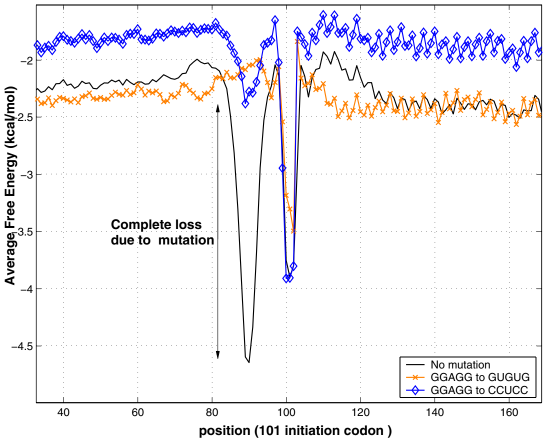

Fig. 6. Results with Hui and De Boer mutations.

just as found in the Jacob investigation. Inspecting the results more carefully, several remarkable and "new" findings can be observed. Some of these are: i) A mutation in position 8 has no influence in the detection of the translation signals, probably the reason is that the role of this nucleotide is to introduce spacing at the moment of decoding the mRNA sequence. ii) A mutation at position 6 has nearly the same influence as a mutation at position 5. iii) A mutation at position 9 affects the recognition of the initiation codon even if it does not affect the SD signal. This could lead to a wrong initiation of translation or a "frame shift". iv) Exactly the central part of the 13 bases (bases 4-8) which influences the SD is missing in eukaryotes. The rest of the sequence that involves AUG and stop codon recognition are still there.

TABLE III

MUTATIONS IN 16SrRNA.

| position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SD** | – | – | – | ↓ | ⇓ | ⇓ | ↓ | – | – | – | – | – | – |
| **Initiation** | – | – | ⇓ | ↑ | ↓ | ↓ | ↓ | – | ⇓ | ↓ | ↓ | ↓ | ↑ |
| **Stop** | ↓ | ⇓ | ↓ | – | – | – | – | – | ↓ | ⇓ | ⇓ | – | ↑ |

## V. CONCLUSIONS

A coding model for the process of translation initiation in gene expression was developed using concepts from communications theory. The model helped to discover and hypothesize about other biological interactions between the ribosome and the mRNA sequences, in addition to the ones already known. This is the case of the implication of the 16SrRNA 3' end in the complete process of translation and not only in initiation. Evolutionary hypothesis were drawn to try to understand the results obtained. The model investigated facilitated the testing of mutations in the ribosome molecular structure. Results showed total agreement with some published investigations on mutations which certifies the correctness of the model. Moreover, the proposed algorithm allows the testing of various combinations of mutations without the need for time and cost consuming laboratory experimentation. The analysis of the results made possible by this model can serve as a way to introduce new lines of biological research. In practice, these results can lead to better recognition of signals in translation, hence, enhancing *in vitro* translation systems in genetic engineering.

## REFERENCES

[1] H. Yockey, *Information theory and molecular biology*. Cambridge: Cambridge University Press, 1992.

[2] T. Schneider, "Theory of molecular machines I. Channel capacity of molecular machines," *Journal Theoretical Biology*, vol. 148, pp. 83–123, 1991.

[3] T. Schneider, "Theory of molecular machines II. Energy dissipation from molecular machines," *Journal Theoretical Biology*, vol. 148, pp. 125–137, 1991.

[4] G. Battail, "An engineer's view on genetic information and biological evolution," *Submitted to Elsevier Science*, July 2003.

[5] M. Eigen, "The origin of genetic information: viruses as models," *Gene*, vol. 135, pp. 37–47, 1993.

[6] E. May, *Analysis of Coding Theory Bases Models for Initiating Protein Translation in Prokaryotic Organisms*. PhD thesis, North Carolina State University, Raleigh, January 2002.

[7] E. May, M. Vouk, D. Bitzer, and D. Rosnick, "Coding model for translation in E.coli K-12," *Proceedings of The First Joint BMES/EMBS Conference*, October 1999.

[8] M. Kozak, "Initiation of translation in prokaryotes and eukaryotes," *Gene*, vol. 234, pp. 187–208, 1999.

[9] J. Steitz and K. Jakes, "How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in E. coli," *Proc. Natl. Acad. Sci.*, vol. 72, pp. 4734–4738, 1975.

[10] J. Shine and L. Dalgarno, "The 3' terminal sequence of *Escherichia coli* 16S ribosomal RNA:complementarity to nonsense triplets and ribosome binding sites," *Proc. Natl. Acad. Sci.*, vol. 71, pp. 1342–1346, 1974.

[11] T. Schneider, "Consensus sequence Zen," *Applied Bioinformatics*, vol. 3, pp. 111–119, 2002.

[12] D. Rosnick, *Free Energy Periodicity and Memory Model for Genetic Coding*. PhD thesis, North Carolina State University, Raleigh, 2001.

[13] "NCBI: National Center for Biotechnology Information." http://www.ncbi.nlm.nih.gov/.

[14] H. Goringer, K. Hujazif, E. Murgolat, and A. Dahlberg, "Mutations in 16S rRNA that affect UGA (stop codon)-directed translation termination," *Proc. Natl. Acad. Sci.*, vol. 88, pp. 6603–6607, August 1991.

[15] J. Watson, *DNA: the secret of life*. London: Arrow Books, 2004.

[16] W. Jacob et al., "A single base change in the Shine Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins," *Proc. Natl. Acad. Sci.*, vol. 84, pp. 4757–4761, 1987.

[17] A. Hui and H. D. Boer, "Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli," *Proc. Natl. Acad. Sci.*, vol. 84, pp. 4762–4766, 1987.