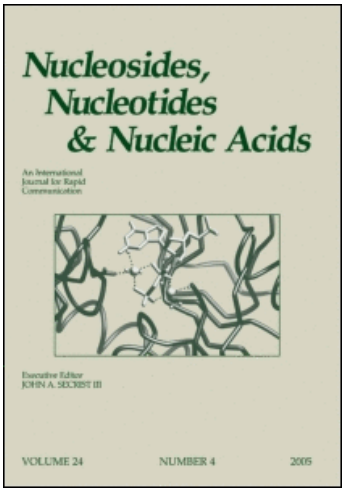

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

Analysis of Similarity/Dissimilarity of DNA Sequences Based on Convolutional Code Model

Xiao Liu[a]; Feng Chun Tian[a]; Shi Yuan Wang[a]
[a] College of Communication Engineering, Chongqing University, Chongqing, China

Online publication date: 24 February 2010

To cite this Article Liu, Xiao , Tian, Feng Chun and Wang, Shi Yuan(2010) 'Analysis of Similarity/Dissimilarity of DNA Sequences Based on Convolutional Code Model', Nucleosides, Nucleotides and Nucleic Acids, 29: 2, 123 — 131
To link to this Article: DOI: 10.1080/15257771003597766
URL: http://dx.doi.org/10.1080/15257771003597766

Nucleosides, Nucleotides and Nucleic Acids, 29:123–131, 2010

ISSN: 1525-7770 print / 1532-2335 online
DOI: 10.1080/15257771003597766

Taylor & Francis
Taylor & Francis Group

ANALYSIS OF SIMILARITY/DISSIMILARITY OF DNA SEQUENCES BASED ON CONVOLUTIONAL CODE MODEL

Xiao Liu, Feng Chun Tian, and Shi Yuan Wang

College of Communication Engineering, Chongqing University, Chongqing, China

□ *Based on the convolutional code model of error-correction coding theory, we propose an approach to characterize and compare DNA sequences with consideration of the effect of codon context. We construct an 8-component vector whose components are the normalized leading eigenvalues of the L/L and M/M matrices associated with the original DNA sequences and the transformed sequences. The utility of our approach is illustrated by the examination of the similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of 11 species, and the efficiency of error-correction coding theory in analysis of similarity/dissimilarity of DNA sequences is represented.*

Keywords Error-correction coding; convolutional code; similarities/dissimilarities analysis

INTRODUCTION

Mathematical analysis of the mass genomic DNA sequence data is still a challenge in biological research. Many researchers have outlined various methods for analyzing the DNA sequences and determining similarities/dissimilarities between them,[1–7] based on different representations of the sequences. For the graphically represented DNA sequences, mathematical invariants have been demonstrated useful for similarity studies of molecules. Euclidean, D/D, L/L, and M/M matrix are used for calculating their leading eigenvalues.[2,8,9]

On the other hand, the rapid increase in genetic data has spurred a renewed interest to use concepts and tools from the field of communication engineering to analyze and understand various processes in the field of molecular biology.[10–16] According to the points Battail proposed,[10,11]

Received 20 October 2009; accepted 16 December 2009.
This work was supported by the Colleges and Universities' Research Foundation for Ph.D. Program of China (20050611022).
Address correspondence to Xiao Liu, College of Communication Engineering, Chongqing University, 174 ShaPingBa District, Chongqing 400044, China. E-mail: liuxiao@cqu.edu.cn

proofreading does not correct errors present in the original genetic message. Only a genetic error correction mechanism can guarantee reliable message regeneration in the presence of errors or mutations due to thermal noise, radioactivity, and cosmic rays. He also proposed the need of the usage of error-correction coding theory in biological field. This idea appeared in some researches. For example, Wang applied error-correction coding theory in microarray data analysis,[12] and May and Ponnala applied error-correction coding theory to analyze the translation initiation.[13,15] However, error-correction coding theory has not been used in similarity study of DNA sequences.

We also notice that a nucleotide of a DNA sequence is treated as an independent information unit in the traditional methods. However, the functions of the codons in the process of translation imply that a codon could be treated as an information unit. Scientists have discovered the effect of codon context on expression and efficiency of translation of some codons;[17,18] however, the effect of the adjacent nucleotides is not considered enough in existing research model.

In this article, we introduce an approach which employs convolutional code model to study the similarity/dissimilarity among DNA sequences, based on error-correction coding theory. We transform the original DNA sequences by convolutional code model, and then the L/L and M/M matrices are used to calculate the corresponding leading eigenvalues to construct a vector for sequences similarity study. The vector contains the effect/information of the adjacent nucleotides, according to the feature of convolutional code. In one of the used convolutional code model, we set a codon as a basic information unit. The similarities are computed by calculating the Euclidean distance between the end points of the vectors. This provides another method to combine biological information analysis with error-correction coding theory.

In Section 2, we outline the method that we propose. Section 3 presents the similarities/dissimilarities analysis among the coding sequences of the first exon of β-globin gene of the 11 species based on convolutional code model and the comparison with some existing methods, followed by our conclusion in Section 4.

THEORETICAL METHOD

The convolutional code which contains the effect of adjacent symbols is one of the error-correcting coding. It is suitable to be paralleled to the effect of codon context. On receiving an input of k bits, a (n,k,L) convolutional encoder gives an output of n bits that is associated with not only the present k bits input but also the previous $L-1$ groups of k bits inputs, wherein L is

called the constraint length of this convolutional code.[19] This model was used for analyzing translation initiation.[13,15]

The output C of convolutional code is shown as follows:

$$C = [m^2 g^1 + m^1 g^2] \tag{1}$$

where m^2 denotes the present input digits, and m^1 denotes the previous input digits.

The generator matrix is an important component of the convolutional code model. In our study, we winnow three convolutional code models: (2,1,2), (3,2,2), (6,3,2) and design their generator matrix based on the following considerations:

(1) Based on short-range dominance of bases correlation [20], we select 2 as the universal constraint length (i.e., $L = 2$). This denotes we will only concentrate on the effect of the adjacent nucleotides.
(2) Considering a nucleotide as a genetic information unit as usually, we select (2,1,2) convolutional code model. The generator matrix we selected is with one row and two columns:

$$g^1 = g^2 = [0\ 1]_{(1\times 2)}\ .$$

(3) Considering a codon as a genetic information unit as we propose, we select (6,3,2) convolutional code model. According to degeneracy of codons, we know that the first two nucleotides of the codon's synonymous codons are often the same but with the difference at the third nucleotide. So we use 0 to characterize the wobble feature of the third nucleotide of a codon and 1 for the first two nucleotides. The generator matrix with three rows and six columns is designed as

$$g^1 = g^2 = \begin{bmatrix} 1\ 1\ 0\ 1\ 1\ 0 \\ 1\ 1\ 0\ 1\ 1\ 0 \\ 1\ 1\ 0\ 1\ 1\ 0 \end{bmatrix}_{(3\times 6)}\ .$$

And (3,2,2) model is selected as a transition because its output contains three nucleotides (with the same length of a codon). Its generator matrix with two rows and three columns is designed as

$$g^1 = g^2 = \begin{bmatrix} 1\ 1\ 0 \\ 1\ 1\ 0 \end{bmatrix}_{(2\times 3)}\ .$$

First, we use the two-dimensional graphical representation method proposed by Randić et al.[2] to digitize the DNA sequences, and construct the

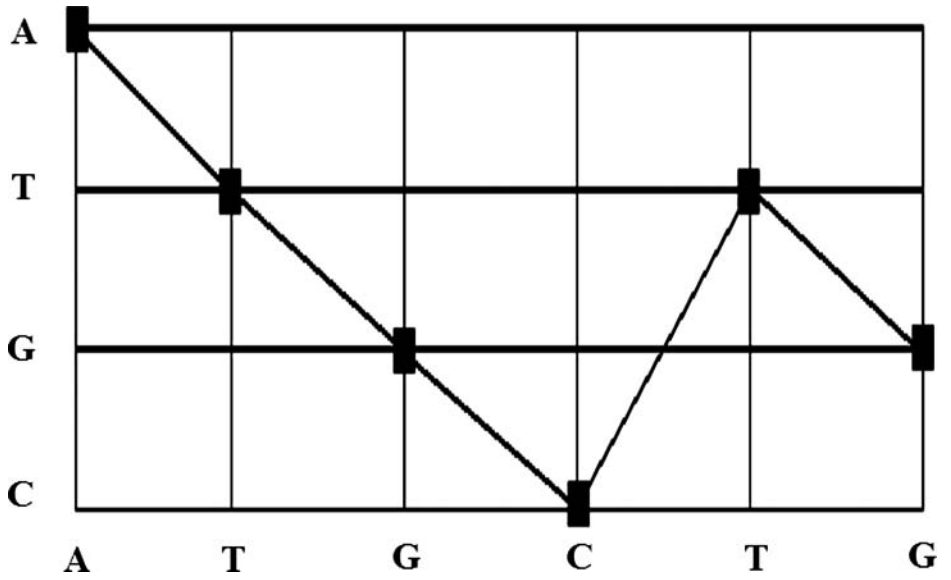

FIGURE 1 The represented sequence segment ATGCTG. The black rectangles denote the bases that make up the sequence.

L/L matrices and M/M matrices of the DNA sequences and calculate the corresponding leading eigenvalues. For example, the first 6 bases of the coding sequence of the first exon of goat β-globin gene (i.e., ATGCTG) are represented as shown in Figure 1. Based on the two-dimensional graphical representation, the L/L matrix is the symmetric matrix whose entries on the main diagonal are defined as zero and off-diagonal entries are defined as a quotient of the Euclidean distance between two vertices of the zigzag curve and the sum of geometrical lengths of edges between the two vertices, and the M/M matrix is the symmetric matrix whose entries on the main diagonal are defined as zero and off-diagonal entries are defined as a quotient of the Euclidean distance between two vertices of the zigzag curve and the number of edges between the two vertices.[2]

Then, we operate the original digitized DNA sequences which are mapped into a set of 4 elements {0, 1, 2, 3} with the convolutional code models. With modulo 4 addition and modulo 4 multiplication, the convolutional code output C is calculated as follows:

$$C = [m^2 g^1 + m^1 g^2] \mod 4 \tag{2}$$

For example, the segment ATGCTG is digitized as 321021. In (2,1,2) convolutional model, its convolutional code output C is 0103010203, according to formula (2). This output is then used to construct its L/L matrix and M/M matrix and the corresponding leading eigenvalues are calculated.

Now, we can construct an 8-components vector for each object to analyze the similarity/dissimilarity among DNA sequences with some of the results of the object (refer to Table 3). The analysis of similarity/dissimilarity among DNA sequences represented by the vectors is based on the assumption that two DNA sequences are similar if the corresponding vectors point to a similar direction in the 8-D space and have similar magnitudes. Here, we calculate the Euclidean distance between these two end-points for measuring the similarity between these two vectors. The smaller the Euclidean distance is, the more similar the two DNA sequences are.

Therefore, we summary the method as follows:

Step 1. Mapping the test DNA sequences into digital sequences.
Step 2. Calculating the leading eigenvalues of the L/L matrix and M/M matrix of the test sequences.
Step 3. Encoding the digitized DNA sequences with the (2,1,2), (3,2,2), (6,3,2) convolutional code models.
Step 4. Calculating the leading eigenvalues of the L/L matrix and M/M matrix of the coded sequences and normalizing them by the lengths of the corresponding data sequences.
Step 5. Constructing an 8-components vector for each object and analyzing the similarity/dissimilarity.

APPLICATION AND COMPARISON

In this section, the proposed approach will be applied to the examination of the similarity among the 11 coding sequences of Table 1.

The leading eigenvalues are listed in Table 2, and they are normalized by the lengths of the data sequences respectively, as listed in Table 3. Then, we construct an 8-component vector for each object with all the entries of the corresponding row in Table 3.

In Table 4, we give the similarity/dissimilarity matrix for the 11 coding sequences of Table 1 based on the Euclidean distances between the end points of the 8-component vectors. We find that the smallest entries in Table 4 are associated with the three kinds of Primates (human, chimpanzee, and gorilla), which denotes strong similarity to each other because of their evolutionary relationship. On the other hand, the largest entries in the similarity/dissimilarity matrix appear in the rows belonging to opossum (the most remote species from the remaining mammals) and gallus (the only nonmammalian representative). On the basis of these findings we conclude that the approach can exhibit the important information of the DNA sequences considered.

Herein, the recently reported representative approaches to examinations of the similarity degree between human and the other 10 species are chosen for comparison, which are listed in Table 5, where in order to make

TABLE 1 The coding sequences of the first exon of β-globin gene of different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGA AAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAA GGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAA GGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAA GGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAA GGTGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAA GGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAA GGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAA AGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

the different approaches comparable, all entries in each row are normalized by setting the similarity value of human–chimpanzee to 1 and scaling the others proportionally. The vectors used for comparison of species in the other approaches of Table 5 are: the vectors having three components based on the symbolic dynamics approach,[5] the 12-component vectors consisting of

TABLE 2 The leading eigenvalues of the L/L matrices and the M/M matrices based on the coding sequences of Table 1

	Leading eigenvalue							
Species	MM	LL	MM212[a]	LL212[b]	MM322[c]	LL322[d]	MM632[e]	LL632[f]
Human	93.375	58.879	184.61	89.194	136.96	78.678	175.59	106.47
Goat	87.347	53.979	172.34	85.027	127.53	79.409	162.70	114.12
Opossum	93.914	50.459	184.68	86.848	136.68	81.469	174.70	128.44
Gallus	93.651	56.070	184.30	90.733	135.96	93.681	174.86	123.93
Lemur	93.409	55.395	184.39	88.844	136.33	89.646	175.94	101.99
Mouse	93.377	57.406	184.59	87.798	136.82	84.534	175.30	113.33
Rabbit	91.232	56.299	180.69	83.907	133.59	82.392	175.19	114.40
Rat	93.381	59.341	184.40	90.165	137.21	79.798	175.36	114.50
Gorilla	94.361	58.530	186.60	89.156	136.96	78.678	181.69	110.65
Bovine	87.245	55.885	172.29	87.013	127.58	79.737	162.66	113.39
Chimpanzee	106.36	66.622	210.63	100.11	155	89.090	205.86	124.33

[a,b,c,d,e,]The number followed the characters MM or LL denotes the convolutional code model we used.

TABLE 3 The normalized leading eigenvalues of the leading eigenvalues of Table 2

Species	Normalized leading eigenvalue							
	MM	LL	MM212[a]	LL212[b]	MM322[c]	LL322[d]	MM632[e]	LL632[f]
Human	1.0149	0.6291	2.0066	0.9586	1.4886	0.8552	1.9086	1.1573
Goat	1.0157	0.6277	2.0039	0.9887	1.4828	0.9234	1.8919	1.3270
Opossum	1.0208	0.5485	2.0074	0.9440	1.4856	0.8855	1.8989	1.3961
Gallus	1.0179	0.6095	2.0033	0.9862	1.4778	1.0183	1.9007	1.3471
Lemur	1.0153	0.6021	2.0043	0.9657	1.4818	0.9744	1.9124	1.1086
Mouse	1.0150	0.6240	2.0064	0.9543	1.4872	0.9189	1.9055	1.2319
Rabbit	1.0137	0.6255	2.0076	0.9323	1.4844	0.9155	1.9465	1.2711
Rat	1.0150	0.6450	2.0044	0.9801	1.4914	0.8674	1.9061	1.2445
Gorilla	1.0146	0.6294	2.0065	0.9587	1.4726	0.8460	1.9537	1.1897
Bovine	1.0145	0.6498	2.0034	1.0118	1.4835	0.9272	1.8914	1.3185
Chimpanzee	1.0129	0.6345	2.0060	0.9534	1.4762	0.8485	1.9606	1.1841

[a,b,c,d,e,f]The number followed the characters MM or LL denotes the convolutional code model we used.

the normalized leading eigenvalues of the L/L matrices based on the four-line representation,[9] the 15-component vectors of the average bandwidth based on a three-dimensional representation,[3] the 4-component vectors with their components being the geometrical of a four-dimensional representation,[4] and the vectors having eight components consisting of the normalized leading eigenvalue of the related matrix.[21] It can be therefore seen from Table 5 that there exists an overall qualitative agreement among similarities based on different approaches for species comparison, despite some variations among them. The presented approach based on the convolutional code model of error-correction theory is shown effectively for application to analysis of similarity/dissimilarity of different species.

TABLE 4 The similarity/dissimilarity matrix for the 11 species of Table 1 based on the Euclidean distances between the end points of the 8-component vectors of the average leading eigenvalues of Table 3

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.1862	0.2546	0.2529	0.1320	0.0984	0.1370	0.0922	0.0585	0.1864	0.0606
Goat		0	0.1207	0.0993	0.2278	0.1023	0.0968	0.1029	0.1722	0.0334	0.1791
Opossum		0	0	0.1600	0.3068	0.1843	0.1578	0.1845	0.2326	0.1508	0.2403
Gallus				0	0.2438	0.1566	0.1473	0.1866	0.2418	0.1074	0.2464
Lemur					0	0.1378	0.1809	0.1792	0.1602	0.2261	0.1585
Mouse						0	0.0611	0.0627	0.0983	0.1084	0.1025
Rabbit							0	0.0859	0.1111	0.1111	0.1134
Rat								0	0.0823	0.1018	0.0896
Gorilla									0	0.1744	0.0125
Bovine										0	0.1810
Chimpanzee											0

TABLE 5 Similarity/dissimilarity comparison of the first exon of β-globin genes between human and other species based on (A) the Euclidean distances between the 8-component vectors from Table 4, (B) the Euclidean distances between the 3-component vectors, (C) the Euclidean distances between the 12-component vectors, (D) the Euclidean distances between the 15-component vectors, (E) the Euclidean distances between 4-component vectors, and (F) the Euclidean distances between 8-component vectors

Species	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
A[a]	3.0726	4.2013	4.1733	2.1782	1.6238	2.2607	1.5215	0.9653	3.0759	1
B[b]	2.4885	6.2061	6.6412	3.145	2.9237	2.5725	3.145	1.1145	1.9771	1
C[c]	3.5882	8.7059	6.4118	5.1176	4.8824	2.4706	2.5294	1.2353	4.9412	1
D[d]	1.6626	2	0.9264	2.0583	0.9831	1.0383	1.0184	0.8359	1.4755	1
E[e]	2.1165	3.8786	4.2864	2.5874	1.2573	2.0728	1.9757	1.068	1.9806	1
F[f]	1.3999	1.227	1.4444	1.1893	0.9961	0.9571	1.3725	0.99	1.3454	1

[a] A from Table 4 of this work.
[b] B from Table 3 of [5].
[c] C from Table III of [9].
[d] D from Table VII of [3].
[e] E from Table IV of [4].
[f] F from Table III of [21].

CONCLUSION

We present an approach for analyzing DNA sequences and a similarity measure method between DNA sequences, with the application of the convolutional code model of error-correction coding theory. According to the comparison of the performance of our method and other existing methods, our method presents its effectiveness. In this article, we try to reflect the feature of the effect of codon context by the convolutional code model, however, the only proof we provide is the validity of the usage of the (6,3,2) model in which a codon is used as an information unit. More evidence should be provided.

In conclusion, this method presents the efficiency of error-correction coding theory in analysis of similarity/dissimilarity of DNA sequences. It may provide motivation for further study on error-correction coding theory and other communication engineering principle in biological field, and help us understand the biological systems further.

REFERENCES

1. Rosen, G.L. Examining coding structure and redundancy in DNA. *IEEE Eng. Med. Biol. Mag.* **2006**, 25, 62–68.
2. Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **2003**, 368, 1–6.
3. Liao, B.; Wang, T.M. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. *Chem. Phys. Lett.* **2004**, 388, 195–200.
4. Liao, B.; Tan, M.S.; Ding, K.Q. A 4D representation of DNA sequences and its application. *Chem. Phys. Lett.* **2005**, 402, 380–383.

5. Wang, S.Y.; Tian, F.C.; Feng, W.J.; Liu, X. Applications of representation method for DNA sequences based on symbolic dynamics. *J. Mol. Struc.: THEOCHEM.* **2009**, 909, 33–42.
6. Randić, M. Another look at the chaos-game representation of DNA. *Chem. Phys. Lett.* **2008**, 456, 84–88.
7. Randić, M.; Vračko, M.; Novič, M.; Plavšić, D. Spectrum-Like graphical representation of RNA secondary structure. *Int. J. Quantum Chem.* **2009**, 109, 2982–2995.
8. Randić, M.; Kleiner, Alexander F.; DeAlba, L.M. Distance/distance matrices. *J. Chem. Inf. Comp. Sci.* **1994**, 34, 277–286.
9. Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **2003**, 371, 202–207.
10. Battail, G. Does information theory explain biological evolution? *Europhys. Lett.* **1997**, 40, 343–348.
11. Battail, G. An engineer's view on genetic information and biological evolution. *Biosystems.* **2004**, 76, 279–290.
12. Wang, X.H.; Istepanian, R.S.H.; Song, Y.H.; May, E.E. Review of application of coding theory in genetic sequence analysis, enterprise networking and computing in healthcare industry, 2003, Healthcom 2003. Proceedings of the 5th International Workshop.
13. May, E.E.; Vouk, M.A.; Bitzer, D.L.; Rosnick, D.I. An error-correcting code framework for genetic sequence analysis. *Journal of the Franklin. Institute* **2004**, 341, 89–109.
14. May, E.E.; Vouk, M.A.; Bitzer, D.L. Classification of Escherichia coli K-12 ribosome binding sites. *IEEE Eng. Med. Biol. Mag.* **2006**, 25, 90–97.
15. Ponnala, L.; Bitzer, D.L.; Vouk, M.A. On finding convolutional code generators for translation initiation of Escherichia coli K-12, Engineering in Medicine and Biology Society 2003, Proceedings of the 25th Annual International Conference of the IEEE, 3854–3857.
16. Dawy, Z.; Morcos, F.; Weindl, J.; Mueller, J.C. Translation initiation modeling and mutational analysis based on the 3′-end of the Escherichia coli 16S rRNA sequence. *Biosystems.* **2009**, 96, 58–64.
17. Yarus, M.; Folley, L.S. Sense codons are found in specific contexts. *J. Mol. Biol.* **1985**, 182, 529–540.
18. Shpaer, E.G. Constraints on codon context in Escherichia coli genes.Their possible role in modulating the effiency of translation. *J. Mol. Biol.* **1986**, 188, 555–564.
19. Proakis, J.G. *Digital Communications: Third Edition*, McGraw-Hill, Inc., New York, 1995.
20. Luo, L.F.; Lee, W.J.; Jia, L.J.; Ji, F.M.; Tsai, L. Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev. E.* **1998**, 58, 861–871.
21. Liu, Y.Z.; Wang, T.M. Related matrices of DNA primary sequences based on triplets of nucleic acid bases. *Chem. Phys. Lett.* **2006**, 417, 173–178.