Review of Application of Coding Theory in Genetic Sequence Analysis

X. H. Wang[1], R. S. H. Istepanian[1], Y. H. Song[2] and E.E. May[3]
[1] School of Computing and Information Systems, Kingston University, Kingston upon Thames, Surrey KT1 1LQ, UK
[2] Dept. of Electronic and Computer Eng., Brunel University, UK
[3] Sandia National Laboratories, Computational Biology Department, P.O. Box 5800, Albuquerque, NM 87185

Abstract- **Applying coding theory in genetic sequence analysis is a new research area for the information theory in biology. It has the potentioal to help the biologist to understand the nature of the living organisms and find the solutions to DNA computations. After an introduction of the biology, this paper reviewed the application of coding theory, especially error ,control theory, in sequence analysis. From the review, the drwabacks of the present applications have been identified and a possible solution to the problem has been given.**

I. INTRODUCTION

The information theory was developed by Shannon to transmit electronic signals[1]. It as introduced into biological world from 1970s by Gatlin[2] and sincethen widely used to analyse the genetic sequences. Coding theory is just a subarea of the information theory which translates the information bits into transmitted data sysmbol. In 1998, May introduced the error control coding in genetic sequence analsys and spikes a new application of infromation theoery in sequcne analysis[3]. In this paper, we are going to review the work of error control theory in sequence analysis and other similar works to try to find out the prospective of this applications, the problem of the application and possible solutions.

The paper is arranged as follows. In section 2, the biological background is given for understanding the basics of the biological worlds. Then the biological world analogs to the communications theory is described in section 4. Section 4 reviews the application of the coding theory in analysing the sequence. In the last sections, the new development of the coding thoery is described and any potentials to use these theory in sequence analysis is discussed.

II. BIOLOGICAL BACKGROUND

For living organisms, its functions are carried out by protein. The process of protein synthesis is illustrated in Fig. 1[4]. For a cell to make a protein, the information from a gene is copied from a strand of DNA into a strand of messenger RNA. Messenger RNA travels out of the nucleus into the cytoplasm, to cell organelles called ribosomes. There, messenger RNA directs the assembly of amino acids that fold into a completed protein molecule.

Protein synthesis starts when the mRNA binds to a small ribosomal subunit near an AUG sequence in the mRNA. The AUG codon is called start codon, since it codes for the first amino acid to be made of the protein. The AUG codon base-pairs with the anticodon of tRNA carry methionine. A large ribosomal subunit binds to the complex, and the reactions of protein synthesis itself can begin.

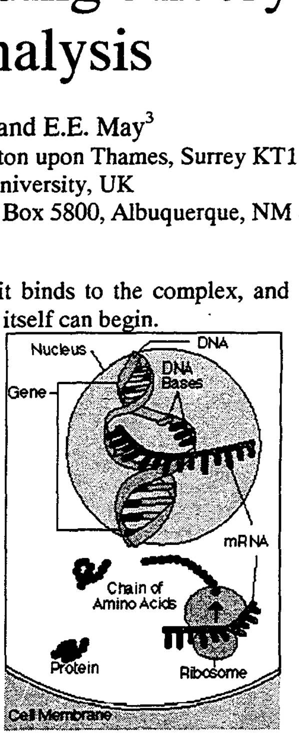

Fig. 1 Process of protein synthesis

Nucleic acids are molecules that dealing the specific structure of the proteins. The backbone of a nucleic acid is made of alternating sugar and phosphate molecules bonded together in a long chain. Each of the sugar groups in the backbone is attached to a third type of molecule called a nucleotide base.

The genetic information is stored in the DNA(molecule deoxyribonucleic acid). There are four different nucleotide bases that occur in DNA: adenine (A), cytosine (C), guanine (G) and thymine (T). These nucleotides bind to the sugar backbone of the molecule. The DNA molecule is actually double-stranded. That is the nucleotide bases of the DNA molecule form complementary pairs: the nucleotides hydrogen bond to another nucleotide base in a strand of DNA opposite to the original. This bonding is specific, and adenine(A) always bonds to thymine(T) (and vice versa) and guanine(G) always bonds to cytosine(C) (and vice versa). This bonding occurs across the molecule leading to a double-stranded system, and the two wrap around each other forming a coil, or helix. Its structure is shown in Fig. 2.

The two distinct ends of a DNA sequence are known under the name of the 5' end and the 3' end. DNA sequence provides all the information needed to function, but the actual work of translating the information into a medium that can be used directly by the cell is done by RNA, ribonucleic acid. RNA has a same sugar-phosphate backbone with nucleotide bases

0-7803-7960-8/03/$17.00©2003IEEE.

attached to it as DNA. It contains the bases adenine (A), cytosine (C), guanine (G) and uracil (U) instead of thymine(T) in DNA. Unlike the double-stranded DNA molecule, RNA is a single-stranded molecule.

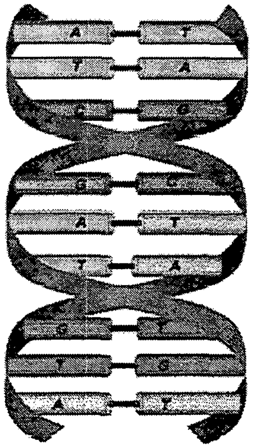

Fig. 2 DNA helix

The RNA has three functions: (a) it serves as the messenger that tells the cell (the ribosomes) what protein to make (messenger RNA; mRNA); (b) it serves as part of the structure of the ribosome, the protein/RNA complex that synthesizes proteins according to the information presented by the mRNA (ribosomal RNA; rRNA); and (c) it functions to bring amino acids (the constituents of the proteins) to the ribosome when a specific amino acid "is called for" by the information on the mRNA to be put in into the protein that is being synthesized; this RNA is called transfer RNA (tRNA).

The messenger RNA (mRNA) serves as an intermediate between DNA and protein. Parts of the DNA are "transcribed" into mRNA which is usually 300-50,000 nucleotides long, and contain the information to translate into a protein of specific sequence.

Over the last couple of years, it has become obvious that the sequence present in DNA does not always dictate literally the sequence of the protein. In a number of instances "RNA editing" has been observed, in which transcripts are chemically modified (for example, some Cs are changed to Us) by enzymes before translation takes place. Thus, the DNA sequence in such cases does not precisely correlate with the sequence of the protein.

III. Finding the Nature of Life — Information Theory

For a general information transmission system, there are three basic elements, namely *transmitter*, *channel* and *receiver*, as depicted in Fig. 3. The transmitter is located at one point, the receiver is located at some other point, and the channel is the physical medium that connects them. The purpose of the transmitter is to convert the message signal into a form suitable for transmission over the channel. However, as the transmitted signal propagates along the channel, it's distorted due to channel imperfections, noise and interfering signal from other sources. Thus the received signal is corrupted. The receiver has the task of operating on the received signal so as to detect, correct the errors and reconstruct a recognizable form of the original message signal for a user.

In biology, the process of conversion of DNA to protein is regarded as one form of information transmission. Information theorist Hubert Yockey depicted this process in his book *Information theory and molecular biology*[5], also shown in Fig. 4.

The information in DNA is transmitted to the information in proteins. DNA is encoded information. Proteins are decoded information. tRNA is the decoder or translator. Noise in the engineering system equals mutation in the biological system. Both systems look much the same. However, in biological world, there is no encoding process. DNA is only decoded. In addition, the genetic code is itself transmitted through the same channel as the message! The genetic code is a coupling of tRNA and one amino acid. Each tRNA/amino acid coupling is catalysed by an 'assignment enzyme'. Those 'assignment enzymes' themselves are encoded in DNA and so are part of the encoded message. And DNA is the only thing that is transmitted (inherited). So the message and the decode instructions are transmitted via the same DNA channel.

The applications of information theory in sequence analysis originated from 1970s when Gatlin introduced it into biology[2]. In the whole 1970s, Gatlin pioneered the work to develop the measures to estimate the complexity of the sequences, such as information, redundancy or divergence. The work was later modified by Sibbald[6]. However, all these attempts didn't succeed in obtaining such a quantitative measure for the sequence. The work in applying information theory in DNA sequence analysis was paused for a period because of the disappointing result.

From 1987, the work was spiked again with the help of great increase in sequence data generated by genome projects and the introduction of various signal processing techniques, such as Fourier transform, autocorrelation, spectral analysis or chaotic dynamics. The most outstanding result was the finding of long-range correlation in DNA sequence[7-11], coding and noncoding sequences classification[12], sequence profiling by chaos-game representation[13- 15], and the latest coding modelling[16-18, 19] of the sequences which will be fully reviewed in next section.

IV. Application of Coding Theory in Sequence Analysis

The application of coding theory in sequence analysis can be classified into two areas. One is using error control coding theory to model the protein translation initiation that is pioneered by May[16-18]. Another is using the coding theory for DNA computation that is contributed by Kari[19] and other researchers[20-22].

While using error channel control theory in modeling the protein translation initiation, May modified the information transmission model described in section 3. The modified model is shown in Fig. 5[16]. In this model, the replication process represents the error introducing channel and the messenger RNA (mRNA) is defined as the output of the communication channel. The genetic decoding process occurs over three levels transcription, translation initiation, and translation elongation plus termination.

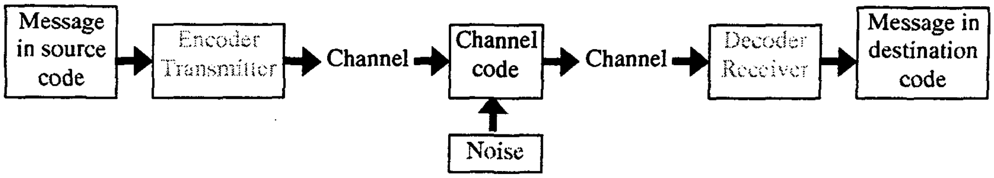

Fig. 3 Electronic information transmission system

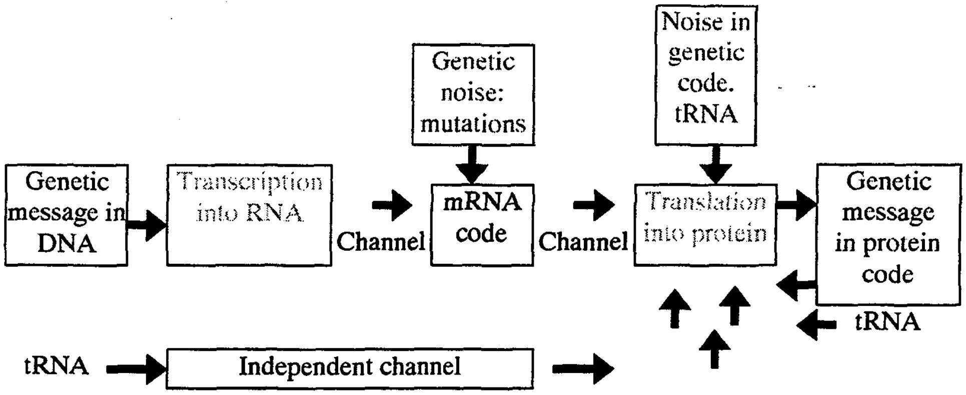

Fig. 4 Biological information transmission system

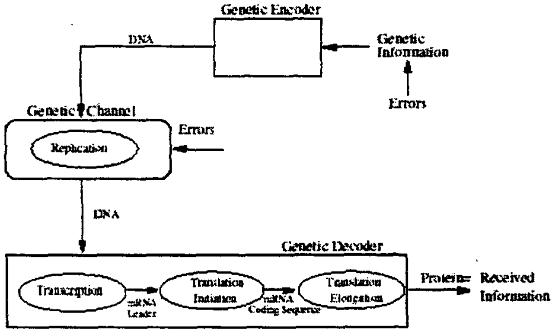

Fig. 5 May's biological information transmission model

Based on this model, the four bases, Adenine(A), Guanine(G), Cytosine(C) and Uracil(U), in mRNA and another base, inosine (I), in transfer RNA are mapped as field of five. That is I=0, A=1, G=2, C=3, and U=4. Therefore, the operation of modulo five is adopted in multiplication and addition.

May uses two coding methods to model the translation of mRNA to protein, namely block coding and convolutional coding. In the block code model, the genetic encoder is modelled as a (n, k) block code whose output is systematic, zero parity check code. The model employs minimum Hamming distance decoding to verify the block code for translation initiation. An example of (5,2) block coding result is shown in Fig. 6. For the convolutional coding method, a binary table-driven decoding algorithm is employed. A modelling example using convolutional coding is shown in Fig. 7.

The results of the two coding methods share the similarities and the big difference between them also exists. First, both methods can distinguish the translated, including hypothetic translated and translated, and non-translated sequence group. Convolutional method can do this more easily and there is an evident distinct between the hypothetic and translated group. Second, both methods can indicate a region between –15 to –10 for the hypothetic/translated group, which is more evident in block coding model. These regions contain the non-random domain and the Shine-Dalgarno domain, which are the key regions in the translation initiation process.

The difference is for the block coding model, it can indicate the zero position clearly for the translation which correspond to alignment of the codeword with the initiation codon, while in convolutional model, it's difficult to see this. However, for the convonlutional model, some other activities are identified in the hypothetic and translated group. These activities are significant or not, or if they are false indication or not are still unknown. If they are the real activity, then the block coding model lose some information. At the same time, the convolutional model definitely loses some information, for example the reading frame synchronization construct. Then the question is arisen from this? Can we find an optimal model to map all the activities? And then can we identify some unknown activities by modeling the sequence by error control theory?

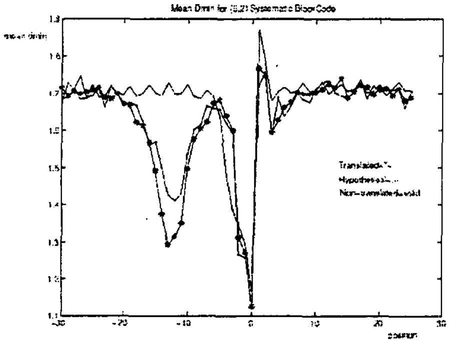

Fig. 6 result of (5,2) block coding model

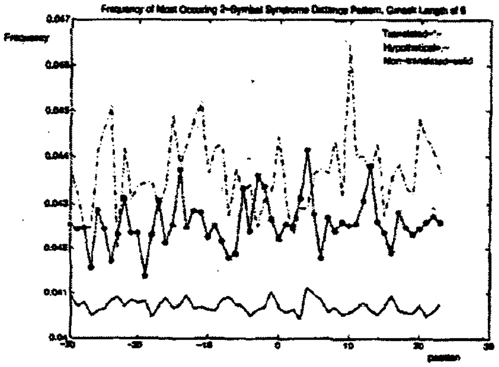

Fig. 7 Result of convolutional modeling of translation initiation

Anyway, three issues are recognized while analyzing the effectiveness of each error-control model for translation initiation: (1) Recognition of regions within the mRNA leader sequence; (2) Distinction between translated and non-translated sequence groups; (3) Indication and recognition of the open reading frame construct. From these points of view, using error channel coding in sequence analysis is a success.

When May modelled the translation initiation using error control coding, the DNA encoding algorithm is unknown and the decoding algorithm is assumed. That arises a question if we can find a suitable encoding algorithm for the DNA sequence. If we can, then the protein synthesis process can be correctly decoded and modelled.

While analysing the DNA computation, a single-stranded DNA is regarded as a codeword and encoded. If the codeword is carefully designed, it will form desired double-stranded structure. That is double helix, also can say the codeword is DNA compliant. Kari studied the relationship of DNA coding and DNA compliance using an involution algorithm[19]. He found out if the codeword is complementarity compliant(c-compliant) or m-compliant for mirror involution, then the codeword is DNA compliant which can form the desirable double strands. That gives us a hint that DNA can be encoded using proper algorithms. Therefore, with the new development of coding technologies, the protein translation initiation or further protein synthesis could be modeled using coding theory.

V. CONCLUSION

From the above review, we can say that applying coding theory in genetic sequences can indicate some important activities in protein translation, classify the translated and non-nontralated sequences and can also indicate some unknown activities. This would help us to explore the insight of the protein synthesis. However, owing to the unkown coding model, the selected decoding model couldn't clearly indicate all these activities in a signle model. A proper decoding model is needed for all these functions. We can also see from the DNA computation that the DNA could be coded as a codeword and makes itself a DNA compliant. If we could combine the two methods together to coding the DNA sequence with a proper algorithm and then decoding the protein synthesis, there would be possible to model the process. That would help to identify the important regions of the sequences or some mutations within the living organisms.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, 1948, pp. 379-423 and 623-656

[2] L. L. Gatlin, *Information Theory and the Living System.* Columbia University Press, New York, 1972

[3] E. E. May, "Comparative analysis of information based models for initiating protein translation in escherichia coli K-12", MS thesis, NCSU, December, 1998

[4] National Cancer Institute, http://press2.nci.nih.gov/sciencebehind/genetesting/genetesting07.htm

[5] Hubert Yockey, *Information Theory and Molecular Biology*, Cambridge University Press, 1992

[6] P.R., Sibbald, S. Banerjee, and J. Maze, "Calculating higher order DNA sequence information measures. *J. Theor. Biol.* Vol.136, 1989, pp.475-483.

[7] W Li, K Kaneko, "Long-range correlation and partial 1/f spectrum in a non-coding DNA sequence", *Europhysics Letters*, Vol.17, No.7, 1992, pp.655-660

[8] W Li and K Kaneko, "DNA correlations (scientific correspondence)", *Nature*, Vol.360, 1992, pp.635-636

[9] W Li, TG Marr and K Kaneko, "Understanding long-range correlations in DNA sequences", *Physica D* (Nonlinearity), Vol.75,1994,pp.392-416

[10] C-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simon, and H.E. Stanley, "Long-range correlations in nucleotide sequences", *Nature*, Vol.356, 1992, pp.168-170

[11] R. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", *Physical Review Letters*, Vol.68, No.25, 1992, 3805-3808

[12] C. Cosmi, V. Cuomo, M. Ragosta, and M. Macchiato, "Characterization of nucleotidic sequences using maximum entropy techniques", *J. Theor. Biol*, Vol.147, 1990, pp.423-432

[13] J.M. Jeffrey, "Chaos game representation of gene structure", *Nucl Acid Res* Vol.18, 1990, pp.2163-2170

[14] J.L. Oliver, P. Bernaola-Galván, J. Guerrero-García and R. Román-Roldán, "Entropic profiles of DNA sequences through chaos-game-derived images", *J. Theor. Biol.* Vol. 160, 1993, pp.457-470

[15] R. Roman-Roldan, P. Bernaola-Galvan and J. L. Oliver, "Application of information theory to DNA sequence analysis: a review", *Pattern Recognition*, Vol. 29, No.7 1996, pp1187-1194

[16] E. May, M. A. Vouk, D. L. Bitzer, D. I. Rosnick, "A coding theory framework for genetic sequence analysis", *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, North Carolina, USA, October 11-13, 2002

[17] E.E. May, MA Vouk, DL Bitzer, DI Rosnick, "The ribosome as a table-driven convolutional decoder for the Escherichia coli K-12 translation initiation system" Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Societ, Piscataway, NJ, USA 2000, pp.2466-9

[18] EE May, Vouk MA, Bitzer DL, Rosnick DI. "Coding model for translation in E. coli K-12" Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society, Piscataway, NJ, USA, 1999, pp.1178

[19] L.Kari, R.Kitto, G.Thierrin, "Codes, involutions and DNA encoding", Workshop on Coding Theory, London, Ontario, July, 2000

[20] Didier G. Arques and Christian J. Michel, "A code in the protein coding genes", *J BioSystems*, vol. 44, pp. 107-134, 1997

[21] Nikola Stambuk, "On circular coding properties of gene and protein sequences", *Croatica Chemica ACTA*, vol.72, no. 4, pp. 999-1008, 1999

[22] Nikola Stambuk, "On the genetic origin of complementary protein coding", *Croatica Chemica ACTA*, vol. 71, no. 3, pp. 573-589, 1998