Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
PERGAMON
Journal of the Franklin Institute 341 (2004) 89–109
Journal of The Franklin Institute
www.elsevier.com/locate/jfranklin

An error-correcting code framework for genetic sequence analysis

Elebeoba E. May[a,*], Mladen A. Vouk[b], Donald L. Bitzer[b], David I. Rosnick[b]

[a] *Sandia National Laboratories, Computational Biology Department, P.O. Box 5800 MS 0310, Albuquerque, NM 87185 0310, USA*

[b] *Computer Science Department, North Carolina State University, Raleigh, NC 27695, USA*

Abstract

A fundamental challenge for engineering communication systems is the problem of transmitting information from the source to the receiver over a noisy channel. This same problem exists in a biological system. How can information required for the proper functioning of a cell, an organism, or a species be transmitted in an error introducing environment? Source codes (compression codes) and channel codes (error-correcting codes) address this problem in engineering communication systems. The ability to extend these information theory concepts to study information transmission in biological systems can contribute to the general understanding of biological communication mechanisms and extend the field of coding theory into the biological domain. In this work, we review and compare existing coding theoretic methods for modeling genetic systems. We introduce a new error-correcting code framework for understanding translation initiation, at the cellular level and present research results for *Escherichia coli* K-12. By studying translation initiation, we hope to gain insight into potential error-correcting aspects of genomic sequences and systems.
Published by Elsevier Ltd. on behalf of The Franklin Institute.

Keywords: Coding theory; Translation initiation; Information theory; Biological coding theory; Error-correcting codes

1. Introduction

The study of the information processing capabilities of living systems was revived in the later part of the 1980s [1–3], due to the increase in genomic data which spurred

*Corresponding author. Tel.: +1-505-844-9933; fax: +1-505-844-5670.
E-mail address: eemay@sandia.gov (E.E. May).

0016-0032/$30.00 Published by Elsevier Ltd. on behalf of The Franklin Institute.
doi:10.1016/j.jfranklin.2003.12.009

a renewed interest in the use of information theory in the study of genomics. Information measures, based on the Shannon entropy [4], have been used in recognition of DNA patterns, classification of genetic sequences, and other computational studies of genetic processes [1,5–18]. Although the focus of current work is on coding theoretic approaches for evaluating genetic systems, results from the field of biological information theory provide a foundation for work in biological coding theory. Informational analysis of genetic sequences has provided significant insight into parallels between the genetic process and information processing systems used in the field of communication engineering [18–21].

The purpose of this work is to explore the application of coding theory to the analysis of biological systems and sequences. Usually, channel codes (also referred to as error-correcting (EC) codes) provide error detection and correction capabilities. Our use of EC codes in this work is in the error detection or error control sense. We are not proposing that genetic systems explicitly correct mutations (errors that are not corrected in the proofreading phase of replication), rather that genetic systems implement error control by incorporating redundancy into the DNA. We review coding theoretic methods for modeling information processing in genetic systems and present an EC code framework for analyzing translation initiation in *Escherichia coli* K-12.

In the section that follows, we give a brief survey of coding theory methods in molecular biology. Section 3 presents an overview of EC codes and reviews coding theoretic frameworks for evaluating genetic systems. EC code models for genetic regulation are presented in Section 4 followed by concluding remarks.

2. Coding theory meets molecular biology

Coding theoretic methods are being used for molecular computing, genetic sequence analysis and synthesis of DNA and oligo-nucleotide chips [16,22–25]. Some researchers have moved beyond qualitative descriptions of genetic communication frameworks to explore EC coding properties of genetic systems using the vast amount of genome data available in public data bases [26–30].

2.1. Coding theoretic methods for genetic sequence analysis and DNA computing

Coding theoretic methods have been used to analyze genetic sequences for various classification purposes. Arques et al. [22] statistically analyzed the results of 12,288 autocorrelation functions of protein coding sequences. Based on the results of the autocorrelation analysis, they identified three sets of circular codes X_0, X_1, X_2 which can be used to distinguish the three possible reading frames in a protein coding sequence. A set of codons X is a circular code, or a code without commas, if the code is able to be read in only one frame without a designated initiation signal [31]. Arques et al. were able to use the three sets of circular codes to retrieve the correct reading frame for a given protein sequence in a 13 base window. They have used

their coding-based model to analyze Kozak's scanning mechanism for eukaryotic translation initiation and other models of translation.

Stambuk [23,32] also explored circular coding properties of nucleic acid sequences. His approach was based on the combinatorial necklace model which asks: "How many different necklaces of length m can be made from a bead of q given colors." Using $q = [\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}]$ and $q = [\mathrm{R} = \text{Purine}, \mathrm{Y} = \text{Pyrimidine}, N = \mathrm{R} \text{ or } \mathrm{Y}]$, Stambuk applied the necklace model to genetic sequence analysis, enabling the use of coding theory arithmetic in the analysis of the genetic code.

Researchers have applied source or compression coding to genetic sequences [16,33]. In the engineering communication system, source encoding, or data compression, occurs prior to channel (EC) coding. Source encoding removes the redundancy in the information stream to reduce the amount of symbols transmitted over the channel. The compression algorithm assigns the most frequent patterns shorter descriptions and the most infrequent patterns are assigned longer descriptions [34]. Loewenstern [35,36] applies source coding methods to genomic sequences for the purpose of motif identification. Powell et al. implemented compression schemes for finding biologically interesting sites in genomic sequences. Delgrange et al. [37] used data compression methods to locate approximate tandem repeat regions within DNA sequences.

The field of DNA computing was launched when Adleman [39] solved an instance of the Hamiltonian path problem using DNA strands to encode the problem and biological processes (annealing, ligation, etc.) to compute a solution. Researchers have proposed several applications using the DNA computing framework including algorithms for breaking the Data Encryption Standard, DNA encryption methods, and techniques to investigate nature's cellular computing processes [25,38–40]. In DNA computing, the information storage capability of DNA is combined with laboratory techniques that manipulate the DNA to perform computations. A major challenge in DNA computing is the problem of encoding algorithms into the DNA medium. Kari and colleagues apply circular coding methods to the forward encoding problem for DNA computing applications [25]. The forward problem being, how can one encode an algorithm using DNA such that one avoids undesirable folding. Kari et al. use coding theory to define heuristics for constructing codewords for DNA computing applications. The codewords cannot form undesirable bonds with itself or other codewords used or produced during the computational process. Error control and correction in DNA computing is also being investigated [41,42].

2.2. EC coding methods for genomic sequence and system analysis

Application of coding theory to genetic data dates back to the late 1950s [43,44] with the deciphering of the genetic code. Since then, EC coding methods have been applied to genetic sequence analysis and classification, biological chip design, as well as analysis of genetic regulatory processes. Konopka [45] applied the theory of degenerate coding to analyze redundancy and the error control properties of the genetic code (the codon to amino acid mapping). Sengupta and Tompa [24] approach the problem of oligo array design from a combinatorial design framework

and use EC coding methods to increase the fidelity of oligo array construction. Reif and LaBean [46] propose EC coding-based methods for the development of error-correction strands for repairing errors in DNA chips.

Several researchers have moved beyond the qualitative models of biological communication and attempted to determine the existence of EC codes for genomic sequences [20,26,28–30]. Liebovitch et al. [28] and Rosen and Moore [30] both developed techniques to determine the existence of EC code for genomic sequence. Neither found evidence of EC codes for the sequences tested. Given the computational limitations of the study, Liebovitch et al. suggest that a more comprehensive examination would be required. Both methods investigate a subset of linear block codes and neither consider convolutional coding properties nor account for the inherent noise in genomic sequences. Extending beyond specific genomic regions and sequences, MacDonaill [26] develops an EC coding model for nucleic acid sequences in general. He has proposed a four-bit, binary parity check EC code for genetic sequences based on chemical properties of the nucleotide bases. As more researchers explore the EC coding properties of genetic sequences and apply these methods to computational biology and molecular computing problems, the information and coding theoretic properties of genetic systems can be further understood and potentially exploited for bioengineering applications.

3. Coding theory and genetic information transmission

3.1. Overview of coding theory

The need for coding theory and its techniques stems from the need for error control mechanisms in a communication system. The system in Fig. 1 illustrates how EC coding is incorporated into a typical communication system [47]. In an engineering communication system, digitized information is encoded by the channel

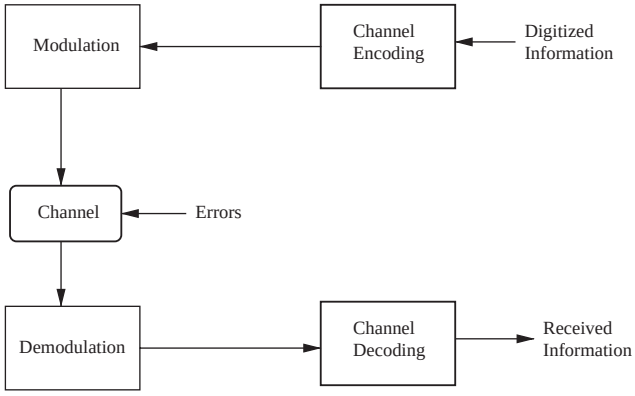

Fig. 1. Communication system that incorporates coding.

encoder and prepared for transmission (modulation). The encoded stream is transmitted through a potentially noisy channel where the sequence can be corrupted in a random fashion. The output of the channel, the received message, is prepared for decoding (demodulation) and then decoded by the channel decoder [47,48]. The decoding process involves removal and possibly correction of errors introduced during transmission. The decoding mechanism can only cope with errors that do not exceed the code's error correction capability.

The encoder processes the digitized information frame by frame. An input frame consists of a fixed number, k, of information symbols that are presented to the encoder. The output frame, the frame to be transmitted, consists of n (also fixed) output symbols, where n is larger than k. Since the number of output symbols is greater than the number of input symbols, redundancy has been introduced [47]. The coding rate

$$R = k/n \tag{1}$$

is the ratio of the number of input symbols in a frame to the number of output symbols in a frame. The lower the coding rate, the greater the degree of redundancy [47,49]. The encoder combines the input symbols and introduces additional symbols based on a deterministic algorithm. This results in a mapping of input frames into a set of output frames known as codewords. The type of output produced is determined by the number of input frames used in the encoding process. Block coding uses only the current input frame. Convolutional coding uses the current frame plus m previous input frames [47,48]. EC codes are referred to as (n,k) codes or (n,k,m) codes in the case of convolutional codes.

The communication channel is the medium through which the information is transmitted to the receiver. The channel can corrupt the transmitted message through attenuation, distortion, interference, and addition of noise. Channels can be characterized as memoryless, symmetric, additive white Gaussian noise (AWGN), bursty, or as compound channels. Channel characteristics determine the type of EC encoding method used in the engineering communication system [47].

The decoder receives a series of frames that, given an errorless transmitted sequence, should be composed only of codewords. If the received sequence has been corrupted during transmission, there will be sequences which do not map uniquely to any codewords. This is used to detect the presence of errors. Decoding algorithms are then used to determine the original codeword and correct the error. When the error rate exceeds the error correction capacity of the code, two things can occur. The decoder may be able to detect the error but may not be able to find a unique solution and thus correct the error or, the decoder may not detect the error because the corruption has mapped one legal codeword into another legal codeword. The method of decoding is dependent on the method of encoding.

The decoding of received bit streams is fairly straightforward when the channel encoding algorithms are efficient and known. What if the encoding scheme is unknown or part of the data is missing? How would one design a viable decoder for the received transmission? Communication engineers may not frequently encounter this situation, but for computational biology this is the immediate challenge and

barrier to understanding the vast amount of sequence data produced by genome sequencing projects. To determine the algorithm used by living systems to transmit vital genetic information, several researchers have explored the parallel between the flow of genetic information in biological systems and the flow of information in engineering communication systems [1,19,20,29,50].

3.2. The need for EC coding in living systems

Battail argues similar to Eigen that for Dawkins' model of evolution to be tractable, error-correction coding must be present in the genetic replication process [21,50–52]. According to Battail [50], proof-reading, a result of the error avoidance mechanism suggested by genome replication literature, 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.

Battail further asserts that the need for error protection becomes obvious when one considers that the number of errors in a k-symbol message that has been replicated r times is comparable to the number of errors in an unreplicated rk-symbol message. For a given error rate, an organism undergoes numerous replication cycles. Hence, for a message to remain reliable within an organism's life cycle and sufficiently viable to ensure the survival of a species the message must have strong error protection. Battail points out that if there exists a minimum Hamming distance d between codewords, then almost errorless communication is possible if and only if the following holds:

$$pn < d/2, \tag{2}$$

where p is the error probability for the channel, n is the length of the codewords, and the Hamming distance is the number of positions in which two codewords differ. Eq. (2) is a restatement of the well-known EC coding bound that sets the error correction threshold for a code with a minimum Hamming distance of d. An EC code with a minimum Hamming distance value of d can correct t or fewer errors for $d \geqslant 2t+1$ [53]. The error correction threshold of the code determines the power of the EC code. According to Battail, if we take n in Eq. (2) to be the length of the gene or a portion of the gene, minimum distance decoding may be used to produce a near errorless rule. Eukaryotes tendency to evolve towards increasing complexity may parallel the connection between increasing word length and increasing reliability, which is stated in the fundamental theorem of channel coding [50]. The fundamental theorem of channel coding states that coding rates that are below the channel capacity result in arbitrarily small probabilities of error ($\lambda^n \to 0$) for sufficiently large blocks lengths, n [34].

The survival of an organism necessitates the existence of a reliable information replication process. Therefore, Battail asserts that EC codes must be used in replication or in another process of information regeneration that precedes replication. Biological organisms may not use EC coding in replication or in a form that directly parallels EC methods used in engineering communication systems.

However, to ensure reliable transmission of its genetic information, similar to Battail, we hypothesize that an organism or species must incorporate a form of EC coding either implicitly through redundancies in their genetic message or explicitly by increasing their copy number. Even though the possibility of biological EC coding is still being investigated, an EC coding framework can provide valuable insight into the underlying structure of genetic regulatory sites. Battail suggests that genetic information undergoes nested encoding, where the result of a previous encoding process is combined with new information and encoded again. The more important genetic information is assumed to be in the primary coded message. Battail's nested coding model mirrors coding theory's concept of concatenated codes [54].

3.3. Biological communication system frameworks

While Battail's argument for the existence of EC codes for living systems is based on Dawkins' model of evolution and focuses on the replication process, several researchers have explored the central dogma of genetics from an information transmission viewpoint. The central premise of genetics is that genes are perpetuated in the form of nucleic acid sequences but function once expressed as proteins [55]. Investigators have developed models that attempt to capture various information theoretic aspects of the genetic system [1,19,20,30,56].

3.3.1. Gatlin's communication model

One of the earliest work on the information theoretic properties of biological systems is presented by Gatlin [19]. In the opening chapter of her work, Gatlin asserts that:

> Life may be defined operationally as an information processing system ... that has acquired through evolution the ability to store and process the information necessary for its own accurate reproduction [19].

Gatlin's interpretation of the biological information processing system is depicted in Fig. 2. In the Gatlin model, DNA base sequences are the encoded message generated by a source, an EC encoder. Gatlin suggests that extra bases in DNA may be used for error control and correction purposes. The encoded DNA goes through a channel (defined in Gatlin's model by transcription and translation) which Gatlin refers to as all the mechanics for protein production. The amino acid sequence of the

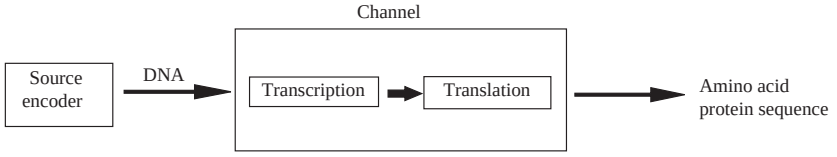

Fig. 2. Gatlin's communication theory view of the genetic system.

protein is the received message. It is unclear where DNA replication fits or whether Gatlin considers the replication process as part of the encoder. Although transcriptional and translational errors occur, replication also introduces errors that propagate beyond a single replication event. Replication is a potentially significant source of noise and should be addressed explicitly.

In addition to an information theoretic view of genetic processes, Gatlin also parallels the genetic sequence to a computer program. She proposes that the genetic code can be viewed as "part of an informational hierarchy" where the redundant DNA regions and the non-coding DNA have important programmatic control functions. It is well known that non-protein coding regions of DNA, such as promoters and the 5′ untranslated leader of messenger RNA (mRNA), have regulatory functions in the protein synthesis process [55,57].

3.3.2. Yockey's communication model

Yockey [20] performs a fundamental investigation of biological information theory and lays the foundations for developing theoretical biology from the mathematical principles of information and coding theory. Yockey's biological information framework diverges from the traditional communication system model and is based on a data storage model, where the behavior of the genetic information system is compared to the logic of a turing machine. The DNA is paralleled to the input tape where the genetic message is the bit string recorded on the tape. The computer program or internal states of the Turing machine are the RNA molecules and molecular machines that implement the protein synthesis process. The output tape, similar to Gatlin's model, is the protein families produced from the recorded message in DNA.

EC codes are used in data storage media to ensure data fidelity; hence, Yockey's model incorporates EC coding. Yockey's DNA–mRNA–protein communication system is re-created in Fig. 3 [20]. In Yockey's DNA–mRNA–protein communication system, the source code, genetic message in DNA, is stored on the DNA tape. Transcription is the encoder, transferring DNA code into mRNA code. Messenger RNA is the channel by which the genetic message is communicated to the ribosome, the decoder. Translation represents the decoding step where the information in the mRNA code is decoded into the protein message or the protein tape. Genetic noise is introduced by events such as point mutations. Yockey states that while genetic noise can occur throughout the system, all of the noise is represented in the mRNA channel. In Yockey's model, the genetic code (the codon to amino acid mapping) is the decoding process and referred to as a block code. He suggests that the redundancy in the codon to amino acid mapping is used as part of the error protection mechanism. Therefore, we can assume that the transcription step would be the EC encoding step in Yockey's model. This assumption is consistent with Yockey's inclusion of transcription in the channel code portion of his communication system diagram. Yockey's DNA–mRNA–protein system is a discrete, memoryless (probability of symbol error is statistically independent of the error history of the preceding symbols), and unconstrained system (any order of symbol transmission is permitted) [20,47].

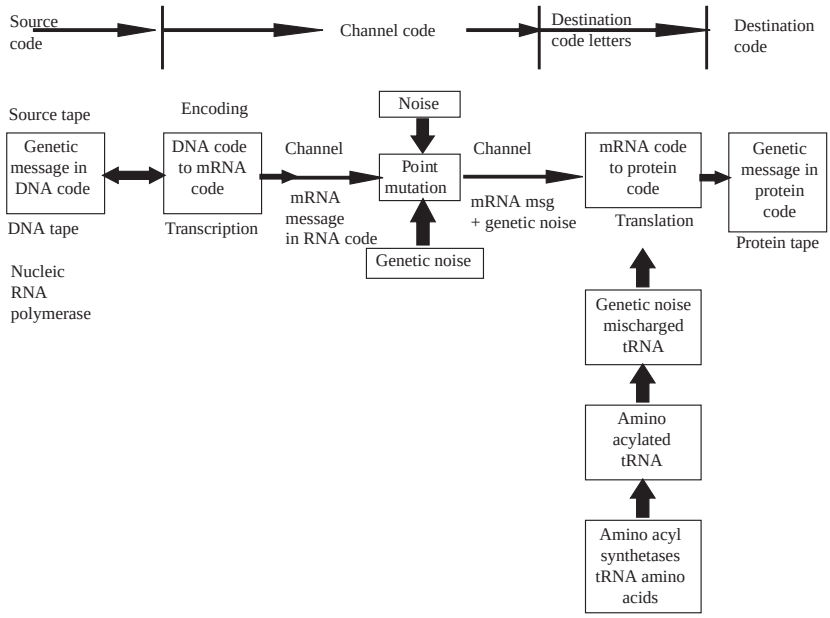

Fig. 3. Yockey's DNA–mRNA–protein communication system.

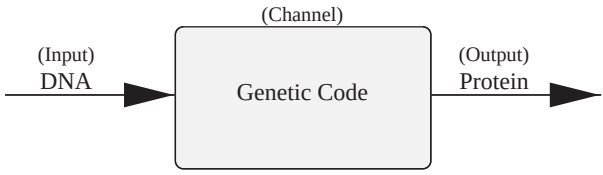

Fig. 4. Roman-Roldan et al.'s information theoretic view of protein synthesis.

3.3.3. Roman-Roldan et al.'s communication model

Roman-Roldan et al. [1] suggest that living beings can be characterized by their information processing ability and hence information based analysis can be used in their study. Viewing protein synthesis as an information processing system allows nucleotide sequences to be analyzed as messages without considering the physical–chemical elements for information processing. Similar to Gatlin, Roman-Roldan et al. models the transfer of biological information as a communication channel with the DNA sequence as the input and the amino acid sequence which forms protein as the channel output, depicted in Fig. 4. Roman-Roldan et al. define the genetic information source as an ergodic source that generates messages from a finite alphabet. An ergodic source is a source that, using a random selection criteria, generates typical messages and atypical messages. Typical or statistically homogenous messages are generated with probability close to one while atypical messages are generated with probability

close to zero [1]. Roman-Roldan et al. define the genetic information source with the following parameters:

- *Genetic alphabet*: $B = [\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}]$, where the members of the alphabet represents adenine, cytosine, guanine, and uracil, respectively.
- $p(\mathrm{A}) + p(\mathrm{C}) + p(\mathrm{G}) + p(\mathrm{U}) = 1$
- Genetic message source is modeled as a Markov source (bases in a message are not independent) with a stochastic distribution matrix

$$[p(B_i|B_j)], \quad \sum_i p(B_i|B_j) = 1.$$

The Markov source is assumed to be stationary and ergodic.

Similar to Yockey and Gatlin, Roman-Roldan et al. designate the genetic code, the process of mapping codons to amino acids, as the transmission channel through which DNA is transmitted and protein is received. If the genetic channel is noiseless, or free of genetic mutations, in Roman-Roldan et al.'s [1] model the input/output probabilities are specified as follows:

$$p(A_i/B_1, B_2, B_3) = \begin{cases} 1 & \text{if } (A_i/B_1, B_2, B_3) \text{ is part of the genetic code,} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where A_i is the ith amino acid and $B_1B_2B_3$ represents the codon. It is assumed that Roman-Roldan et al.'s model essentially parallels Gatlin's model since DNA rather than mRNA is the input into the channel. As in previous models, Roman-Roldan et al. does not address the role of DNA replication in the genetic information transmission framework. Additionally, their model does not explicitly address the presence or function of redundancy in the DNA input sequence.

3.3.4. *May et al.'s communication model*

The communication channel view proposed by Roman-Roldan et al. differs from the initial model presented by May et al. [58]. Our initial model defined the mRNA as the output of the communication channel and incorporates a decoder that translates the mRNA into protein forming amino acid chains. Originally the channel consisted of the DNA replication and transcription process during which errors are introduced into the nucleotide sequence. Based on Battail and Eigen's work, our initial communication view of the genetic system is modified as follows: (1) the replication process represents the error-introducing channel; (2) a nested genetic encoder is assumed. The genetic decoding process is separated into three phases: transcription, translation initiation, and translation elongation plus termination [21,50,58]. Fig. 5 depicts our coding theoretic view of information transmission in genetic systems. In our genetic communication system, the un-replicated DNA sequence is the output of an EC genetic encoder that adds redundancy to inherently noisy genetic information. The noise in the source can be thought of as mutations transferred from parent to offspring. Drawing from Gatlin's parallel of the genetic sequence to computer programs, we can view the non-coded genetic information the organism is communicating as instructions for protein production or control of

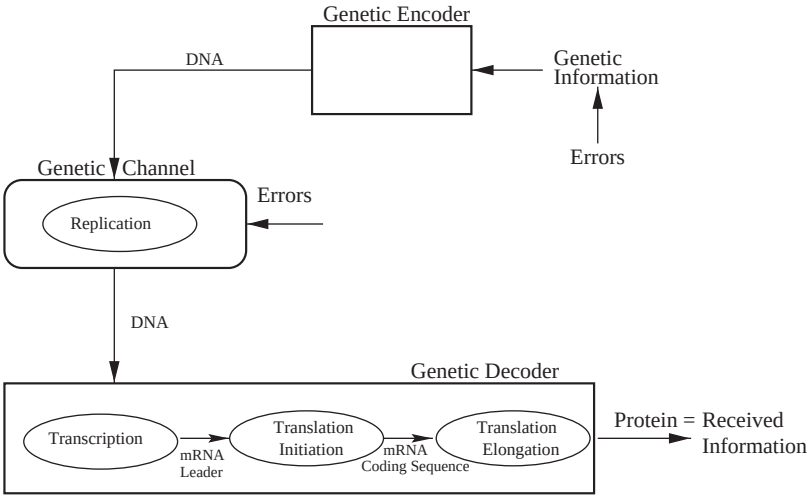

Fig. 5. May et al.'s coding theoretic view of the central dogma of genetics.

protein production. In contrast to Yockey's model, our encoder does not parallel any aspect of the central dogma of genetics. Unlike replication, transcription, and translation, the encoding process does not and must not introduce errors into the message. Therefore, none of these genetic processes can adequately mirror an EC encoder. Defining the genetic encoder in our model would require addressing the question of the origin of the genetic code. As Yockey states in reference to this issue:

> ... the reason for the difficulty in speculating on the origin of the genetic code is that there seems to be no place to begin. There is no trace in physics or chemistry of the control of chemical reactions by a sequence of any sort or of a code between sequences [20].

Perhaps, additional insight into potential biological functions corresponding to the encoder will emerge as researchers continue investigating the evolution of the genetic code and speculating on the origins of the code [59,60].

Neither of Gatlin, Yockey, or Roman-Roldan et al.'s models explicitly addresses replication in their models. All three frameworks represent the noise introducing channel as the genetic mechanisms responsible for protein synthesis, namely transcription and translation in Gatlin and Roman-Roldan et al.'s frameworks and the mRNA itself in Yockey's framework. We define the genetic channel as the DNA replication process during which errors are introduced into the nucleotide sequence [56]. Similar to Roman-Roldan et al., we assume the transmission channel to be stationary, memoryless and discrete. This is also in line with Yockey's assumptions about the genetic channel. Assuming the channel is memoryless simplifies the model but may not accurately represent biological systems and events such as the existence of mutation hotspots. Further investigation is necessary to determine the channel characteristics of the replication channel. Additionally, our model does not explicitly address errors in transcription and translation. Although

these errors do not propagate like replication errors, a more detailed analysis of their impact on the genetic communication system is needed.

Incorporating Battail's nested coding idea, EC decoding occurs in three phases represented by transcription, translation initiation, and translation elongation plus termination. Similar to Yockey, the ribosome is paralleled to an EC decoder. Given the similarities between the translation and transcription mechanisms we represent transcription as a decoding step and view the RNA polymerase as the EC decoder. While Gatlin addressed the potential function of non-protein coding regions, she does not specifically highlight these regions in her communication model of genetics. Our model distinguishes between the EC decoding of protein coding regions and the decoding of non-protein-coding regions that regulate translation initiation, typically the rate-limiting aspect of the protein production process [55,61,62]. Given the importance of regulating protein synthesis and the redundancy present in regulatory regions such as the ribosome binding site (RBS) it seems plausible that regulatory information is EC encoded [63]. We have applied our EC coding framework to the analysis of translation initiation sites in *E. coli* K-12 [29,64,65].

Development of coding theoretic frameworks for molecular biology is an ongoing endeavor. Although the existence of redundancy in genetic sequences is accepted and the possibility of that redundancy for error correction and control is being explored and exploited, mathematically determining the encoding algorithm particularly for regulatory regions remains a major research challenge.

4. EC coding models for genetic regulatory sites

Advances in genetic sequencing have provided large amounts of genomic data for developing computational tools for recognition and modeling of protein coding regions and, in the recent past, identification of translation start sites [17,63,66–70]. Probabilistic methods, such as Suzek et al.'s [71] RBSFinder and Yada et al.'s [72] GeneHacker Plus return the location of the initiation codon for prokaryotic genes. Besemer et al.'s [73] GeneMarkS uses iterative hidden Markov models (HMMs) to locate translation start sites with relatively high accuracy. Walker et al. couple statistical methods with comparative genomics to identify start sites. Hannenhalli et al. incorporate several biological factors into their quantitative description of translational start sites, including the binding energy at the RBS, distance between RBS and initiator, and the initiator codon. They use a mixed integer linear program to determine parameters for their discriminatory model. Other computational techniques including support vector machine, machine learning, combinatorial approaches, free energy calculations and information theory have also been used in quantifying and classifying translational start sites [18,43,70,74,75].

Channel code models of regulatory systems focus on analyzing the regulatory sites using coding theoretic approaches rather than statistical methods. They do not attempt to explicitly describe interactions between regulatory macromolecules or capture gene networks. Unlike statistical methods, effective coding theoretic models can provide quantitative insight on the effects of individual bases on regulatory

efficacy and a framework for understanding the interaction of the regulatory site with macromolecular "decoders." Beyond the work of Forsdyke, Bermel et al., and our investigation of block and convolutional code models for translation initiation, there is little known research into the development of channel coding models for genetic regulatory processes.

4.1. EC coding models for introns and promoters

Forsdyke [27] analyzes introns from an EC coding framework. He designates error-correction due to double stranded DNA as error-correction in-parallel and EC in single-stranded nucleic acid sequences as error-correction in-series. Forsdyke proposes that introns serve as parity checks for the information being transmitted in exons. He suggests that the parity check intron sequence does not need to be located next to the exon sequence it checks but can reside in a separate part of the genome. Forsdyke also proposes a two-dimensional view of the intron error-checking system. Bermel et al. [76] investigates table-based convolutional code models for *E. coli* promotors. Based on the information content of the promoters, Bermel et al. approximates a $\frac{1}{9}$ coding rate for the *E. coli* promoter and devised a $\frac{1}{5}$ binary convolutional code model for the region.

4.2. EC code models for translation initiation

Each codeword, v, in an (n,k) block code or an (n,k,m) convolutional code's codebook can be produced using a generator matrix, G, which encodes the information vector, u, in a deterministic manner [49]. In general, the relationship between u, v, and G is as follows:

$$v = uG, \tag{4}$$

where G is $k \times n$ for block codes, u is $1 \times k$, and v $1 \times n$. Based on the May et al. model of genetic transmission, Fig. 5, we assume that the messenger RNA, r, is the received noisy version of v. The objective is to determine the encoding model, G, using r. The code model is either specified by the coefficients of G as in the convolutional code model for translation initiation (Section 4.2.2) or the set of all valid codewords (referred to as the codebook) as in the block code model (Section 4.2.1).

Although one does not know the exact mechanism employed by the genetic decoder, by analyzing key elements involved in initiating protein translation, we hope to gain insight into possible decoding schemes used in the initiation of translation in prokaryotic organisms. The key elements considered are: the 3′ end of the 16S ribosomal RNA, the common features of bacterial ribosomal binding sites (such as the existence and location of the Shine–Dalgarno sequence), and RNA/DNA base-pairing principles. A block coding model and convolutional coding model for the translation initiation system are explored [58,64,65]. Assuming an encoding method, the corresponding decoding algorithm is designed.

4.2.1. The (5, 2) *block code model*

In the block code model, the genetic encoder is modeled as an (n, k) block code whose output is a systematic zero parity check code [47,58]. Codewords of length $n = 5$ and 8 are developed based on the last 13 bases of the 3′ end of 16S ribosomal RNA, which contains the hexamer complementary to the Shine–Dalgarno sequence [55]. The model employed a minimum distance decoder to verify the block coding model for translation initiation.

The *E. coli* K-12 strain MG1655 sequence data (downloaded from the NIH ftp site: ncbi.nlm.nih.gov) is used to test the model. Fig. 6 shows the resulting mean minimum Hamming distance by position for the (5, 2) block code model. The smaller the value on the vertical axis, the stronger the bond formed between the ribosome and the mRNA. Zero on the horizontal axis corresponds to the alignment of the first base of a codeword with the first base of the initiation codon. As Fig. 6 illustrates, there is a significant difference among the translated, hypothetical, and the non-translated sequence groups. For the translated and hypothetically translated sequence groups, a minimum distance trough occurs between the −15 and −10 regions. The −15 to 0 region contains large synchronization signals which can be used to determine valid protein coding sequences or frames. There are also smaller synchronization signals outside the −15 to 0 region which seem to oscillate with a frequency of 3. These oscillations may reflect valid and invalid reading frames inside the protein coding region.

The (5, 2) and (8, 2) model (discussed in [29]) are able to distinguish between translated sequence groups and non-translated sequence groups from *E. coli* K-12 genome. When applied to mRNA leader regions of other prokaryotic organisms (*Salmonella typhimurium* LT2, *Bacillus subtilis*, and *Staphylococcus aureus* Mu50), similar results are observed [29].

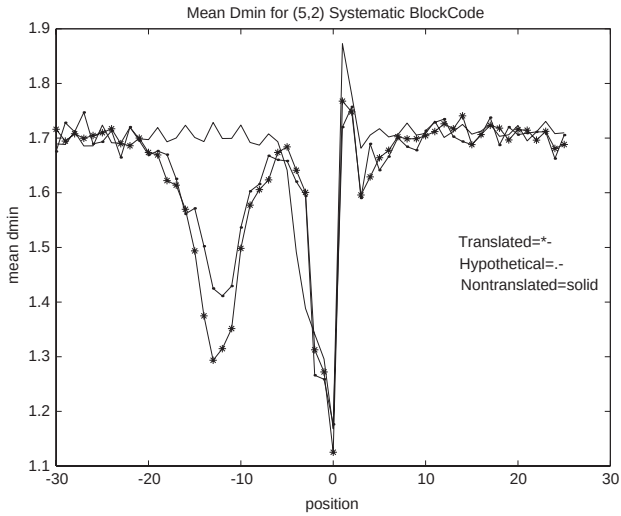

Fig. 6. Results of minimum distance block decoding model for (5, 2) code.

4.2.2. The $(3,1,4)$ *convolutional code model*

Convolutional coding produces encoded blocks based on present and past information bits or blocks. The modeling assumption is that genetic operations such as initiation and translation may involve "decisions" which are based on immediate past and immediate future information. This would allow error correction and other related functions. The convolutional code model views the ribosome as a mechanism with memory. Evaluating the messenger RNA as convolutionally encoded data allowed the model to capture the inter-relatedness between the bases in a mRNA sequence.

Genetic algorithms (GAs), an evolutionary computing method, are used to search for table-based convolutional codes whose decoding masks (gmasks) recognize individual mRNA leader sequences [77,78]. A gmask is a sequence vector, derived from the EC code, used to calculate the syndrome vector for a received sequence (the mRNA is the received sequence). The syndrome vector is used in the decoding process. The GA search space consists of all possible $(n=3,k=1,m=4)$ convolutional codes or individuals. The fitness of each individual in the population (a set of potential solutions) is based on the syndrome values produced when the code's gmasks are applied to the mRNA parity sequence. There are two gmasks, *gmask*1 and *gmask*2, for a $(3,1,4)$ convolutional code. An all zero syndrome value indicates that no errors within the code's error detection capability occurred. In the GA approach, random selection and target sampling rates are used to select highly fit individuals for reproduction. New populations are created using parameterized uniform crossover. Mutation is used to preserve population diversity and elitism ensures that the most fit solution is not discarded. Usually, the definition of a good convolutional code is based on memory length, error detecting, and error correcting capabilities. For "genetic" convolutional codes the most important feature of a good code is how well it distinguishes errors from non-errors, non-ribosome binding sites from RBSs.

Messenger RNA leader sequences from *E. coli* K-12 strain MG1655 (down-loaded from the NIH ftp site: ncbi.nlm.nih.gov and parsed by Rosnick [79]) are used as training sequences for constructing the best candidate code model. There were three types of models: horizontal, horizontal motif, and vertical. Horizontal coding models produced the most probable generator matrix for each mRNA parity sequence (266 sequences in the training set). The horizontal motif model weights biologically significant regions (like the Shine–Dalgarno domain and the non-random domain) higher than other regions in the mRNA leader sequence. It mirrors a nested error control coding model. The vertical or positional code model attempts to find a single generator matrix for each of the 48 positions in the leader region of all 266 mRNA sequences in the training set. The syndrome distance vector for each code model is calculated and indicates how well the associated decoder recognizes the training subsequence. If the GA found the perfect code, the convolutional coding system that produced the exact sequence, then the syndrome distance vector should be the all zero vector and the fitness value would be 1. Fig. 7 shows the average syndrome distance value for the optimal codes discovered using the *E. coli* training set. In Fig. 7, the horizontal axis is position relative to the first base in the initiation

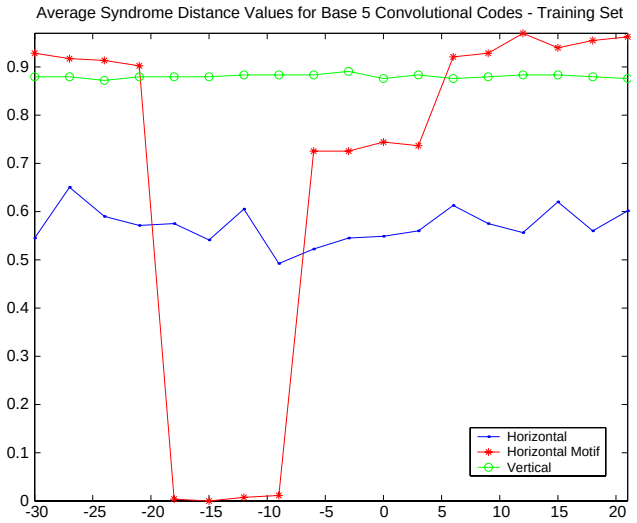

Fig. 7. Average syndrome distance of table-based convolutional code models for translation initiation.

codon and the vertical axis is the average syndrome distance value. For the horizontal codes the individual syndrome distance values for each code model are averaged over 266 models. The vertical code model is the average syndrome value for each of the 48 positional models. As Fig. 7 illustrates, when compared to the horizontal code models, the average syndrome distance for the vertical code models does not indicate any regions of significant activity. The lowest average syndrome distance value for the equal weight horizontal code models occurs at position -9 while the motif-based horizontal code models have approximately zero average syndrome distance values from position -18 to position -9. These positions correspond to the non-random domain and the Shine–Dalgarno domain, key regions in the translation initiation process.

If the ribosome functionally parallels a table-based decoder, the gmask of the code models may resemble the exposed part of the 16S rRNA. Convolutional codes with high similarity to the last 13 bases of the 16S rRNA and low syndrome distance (i.e. high fitness) values would, from a biological perspective, be more plausible models for translation initiation. Fig. 8 depicts the relationship between each code model's fitness score and the model's similarity to the last thirteen bases of the 16S rRNA. The Hamming distance between the $3'$ end of the 16S rRNA and the code model's gmask is used to measure the degree of similarity. The horizontal axis in Fig. 8 is fitness and the vertical axis is percent similarity. In Fig. 8, for gmask1, the equal weight horizontal code model and the vertical code model achieve the highest percent similarity to the 16S rRNA. But, for gmask2 the motif-based horizontal code model and the vertical code model achieve the highest percent similarity scores. In all code model groups, there exist individuals with high similarity values and relatively high fitness values.

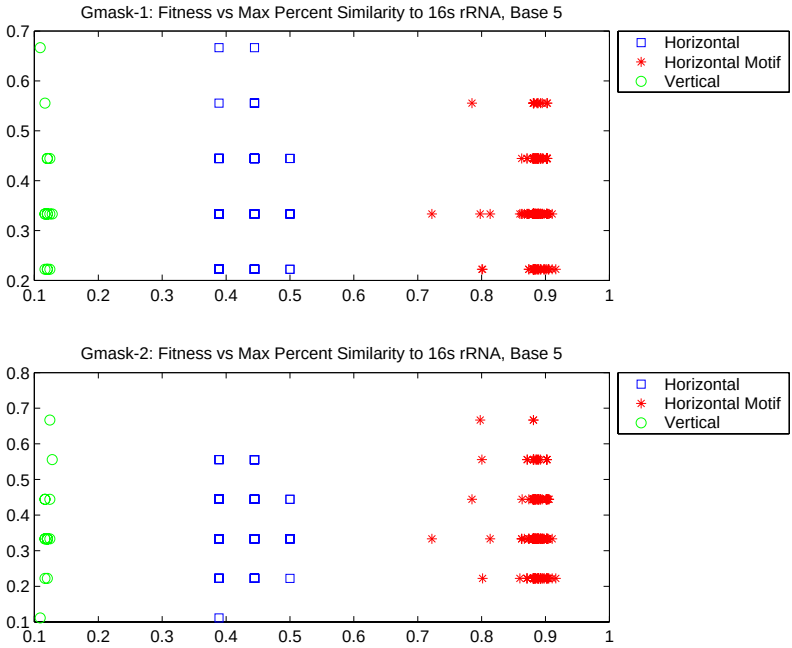

Fig. 8. Individual fitness versus individual similarity values for code models: gmask1 (top) and gmask2 (bottom).

The results of the block and convolutional error-control coding models suggest that it is possible to develop coding-based heuristic for distinguishing between protein coding and non-protein coding genomic sequences by "decoding" the mRNA leader region. Most importantly, results also imply that genetic systems may use methods which are functionally parallel to channel coding techniques to protect, transmit, store, and detect genetic signals; this demonstrates the plausibility of our EC coding framework for genetic communication. The successful development and implementation of a channel coding model for genetic systems can lead to the development of powerful methods for identifying and manipulating protein coding sequences within a genome as well as further our understanding of translation regulatory mechanisms.

5. Conclusion

Effective and secure data transmission remains an ongoing engineering endeavor, but biological systems have learned how to efficiently and effectively combine all three aspects of communication seamlessly (this includes compression, encryption, and error-correction coding). Existing information and coding theoretic frameworks for understanding biological information processing provide foundations that can be expanded to increase our quantitative understanding of how genetic information is

packaged, transmitted, and expressed with relatively high accuracy in an error introducing environment. Such understanding will enable researchers to quantify key processes that govern information transmission in living systems and understand the relationship between genetic errors and disease. In addition to contributions to the biological sciences, the engineering and mathematical sciences stand to benefit from the knowledge gained from modeling and deciphering the rules and algorithms that govern biological communication. Insights gained from studying biological coding theory can contribute to the development of more effective coding theory algorithms for engineering communication systems. Ongoing work includes quantitative characterization of the genetic communication system and improvement of the EC coding models for translation initiation.

Acknowledgements

The authors would like to thank the reviewers for their insightful comments and suggestions for improving the initial manuscript. This work was supported in part by a National Science Foundation Minority Graduate Fellowship and the Ford Foundation Dissertation Fellowship for Minorities. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

References

[1] R. Roman-Roldan, P. Bernaola-Galvan, J.L. Oliver, Application of information theory to DNA sequence analysis: a review, Pattern Recognition 29 (7) (1996) 1187–1194.
[2] R. Sarkar, A.B. Roy, P.K. Sarkar, Topological information content of genetic molecules—I, Math. Biosci. 39 (1978) 299–312.
[3] T.B. Fowler, Computation as a thermodynamic process applied to biological systems, Int. J. Biomed. Comput. 10 (6) (1979) 477–489.
[4] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1949.
[5] K. Palaniappan, M.E. Jernigan, Pattern analysis of biological sequences, in: Proceedings of the 1984 IEEE International Conference on Systems, Man, and Cybernetics, Halifax, Canada, 10–12 October 1984.
[6] H. Almagor, Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach, J. Theor. Biol. 117 (1985) 127–136.
[7] T.D. Schneider, Theory of molecular machines. II. Energy dissipation from molecular machines, J. Theor. Biol. 148 (1991) 125–137.
[8] T.D. Schneider, Theory of molecular machines. I. Channel capacity of molecular machines, J. Theor. Biol. 148 (1991) 83–123.
[9] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, J. Mol. Biol. 219 (1991) 555–565.
[10] P. Salamon, A.K. Konopka, A maximum entropy principle for the distribution of local complexity in naturally occuring nucleotide sequences, Comput. Chem. 16 (2) (1992) 117–124.

[11] J.L. Oliver, P. Bernaola-Galvan, J. Guerrero-Garcia, R. Roman-Roldan, Entropic profiles of DNA sequences through chaos-game-derived images, J. Theor. Biol. 160 (1993) 457–470.
[12] F.M. De La vega, C. Cerpa, G. Guarneros, A mutual information analysis of tRNA sequence and modification patterns distinctive of species and phylogenetic domain, in: Pacific Symposium on Biocomputing, Hawaii, USA, 1996, pp. 710–711.
[13] T.D. Schneider, D.N. Mastronarde, Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method, Discrete Appl. Math. 71 (1996) 259–268.
[14] B.J. Strait, T.G. Dewey, The Shannon information entropy of protein sequences, Biophys. J. 71 (1996) 148–155.
[15] A. Pavesi, B. De Iaco, M.I. Granero, A. Porati, On the informational content of overlapping genes in prokaryotic and eukaryotic viruses, J. Mol. Evol. 44 (6) (1997) 625–631.
[16] D. Loewenstern, P.N. Yianilos, Significantly lower entropy estimates for natural DNA sequences, in: Proceedings of the Data Compression Conference, Snowbird, UT, USA, 25–27 March 1997.
[17] T.D. Schneider, Information content of individual genetic sequences, J. Theor. Biol. 189 (1997) 427–441.
[18] T.D. Schneider, Measuring molecular information, J. Theor. Biol. 201 (1999) 87–92.
[19] L.L. Gatlin, Information Theory and the Living System, Columbia University Press, New York, NY, 1972.
[20] H. Yockey, Information Theory and Molecular Biology, Cambridge University Press, New York, NY, 1992.
[21] M. Eigen, The origin of genetic information: viruses as models, Gene 135 (1993) 37–47.
[22] D.G. Arques, C.J. Michel, A code in the protein coding genes, Biosystems 44 (1997) 107–134.
[23] N. Stambuk, On circular coding properties of gene and protein sequences, Croat. Chem. Acta 72 (4) (1999) 999–1008.
[24] R. Sengupta, M. Tompa, Quality control in manufacturing oligo arrays: a combinatorial design approach, J. Comput. Biol. 9 (1) (2002) 1–22.
[25] L. Kari, J. Kari, L.F. Landweber, Reversible molecular computation in ciliates, in: J. Karhumaki, H. Maurer, G. Paun, G. Rozenberg (Eds.), Jewels are Forever, Contributions on Theoretical Computer Science in Honor of Arto Salomaa, Springer, Berlin, 1999, pp. 353–363.
[26] D. MacDonaill, A parity code interpretation of nucleotide alphabet composition, Chem. Commun. 18 (2002) 2062–2063.
[27] D.R. Forsdyke, Are introns in-series error-detecting sequences? J. Theor. Biol. 93 (1981) 861–866.
[28] L.S. Liebovitch, Y. Tao, A. Todorov, L. Levine, Is there an error correcting code in DNA? Biophys. J. 71 (1996) 1539–1544.
[29] E. May, M. Vouk, D. Bitzer, D. Rosnick, Analysis of coding theory based models for initiating protein translation in prokaryotic organisms, in: Fifth International Workshop on Information Processing in Cells and Tissues, Lausanne, Switzerland, September 2003.
[30] G. Rosen, J. Moore, Investigation of coding structure in DNA, in: ICASSP 2003, Hong Kong, 2003.
[31] E.R. Berlekamp, Algebraic Coding Theory, McGraw-Hill, New York, NY, 1968.
[32] N. Stambuk, Symbolic cantor algorithm (SCA): a method for analysis of gene and protein coding, Period. Biol. 101 (4) (1999) 355–361.
[33] D.R. Powell, D.L. Dowe, L. Allison L, T.I. Dix, Discovering simple DNA sequences by compression, in: Pacific Symposium on Biocomputing, Hawaii, USA, 1998, pp. 597–608.
[34] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, NY, 1991.
[35] D.M. Loewenstern, H.M. Berman, H. Hirsh, Maximum a posteriori classification of DNA structure from sequence information, in: Pacific Symposium on Biocomputing, Hawaii, USA, 1998, pp. 669–680.
[36] D. Loewenstern, P.N. Yianilos, Significantly lower entropy estimates for natural DNA sequences, J. Comput. Biol. 6 (1) (1999) 125–142.
[37] O. Delgrange, M. Dauchet, E. Rivals, Location of repetitive regions in sequences by optimizing a compression method, in: Pacific Symposium on Biocomputing, Hawaii, USA, 1999, pp. 254–265.
[38] C.C. Maley, DNA computation: theory, practice, and prospects, Evol. Comput. 6 (3) (1998) 201–229.

[39] L. Adleman, P. Rothemund, S. Roweis, E. Winfree, On applying molecular computation to the data encryption standard, J. Comput. Biol. 6 (1) (1999) 53–63.
[40] L.F. Landweber, L. Kari, The evolution of cellular computing: nature's solution to a computational problem, Biosystems 52 (1999) 3–13.
[41] D. Boneh, C. Dunworth, R.J. Lipton, J. Sgall, Making DNA computers error resistant, in: L. Landweber, E. Baum (Eds.), DNA Based Computers II, DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, Vol. 44, American Mathematical Society, DIMACS, Providence, RI, 1999, pp. 165–172.
[42] D.H. Wood, Applying error correcting codes to DNA computing, in: Fourth DIMACS Workshop on DNA Based Computers, Philadelphia, PA, USA, June 1998.
[43] B. Hayes, The invention of the genetic code, Am. Sci. 86 (1) (1998) 8–14.
[44] S.W. Golomb, Efficient coding for the desoxyribonucleic channel, Proceedings of the Symposia in Applied Mathematics, New York, NY, USA, Mathematical Problems in the Biological Sciences, Vol. 14, American Mathematical Society, Providence, RI, 5–8 April 1961, pp. 87–100.
[45] A.K. Konopka, Theory of degenerate coding and informational parameters of protein coding genes, Biochimie 67 (1985) 455–468.
[46] J. Reif, T. LaBean, Computationally inspired biotechnologies: improved DNA synthesis and associative search using error-correcting codes and vector-quantization, in: A. Condon (Ed.), DNA Computing: Sixth International Meeting on DNA-Based Computers (DNA6), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Leiden, The Netherlands, June 2000.
[47] P. Sweeney, Error Control Coding an Introduction, Prentice-Hall, New York, NY, 1991.
[48] A. Dholakia, Introduction to Convolutional Codes with Applications, Kluwer Academic Publishers, Norwell, MA, 1994.
[49] Shu Lin, D.J. Costello Jr., Error Control Coding: Fundamentals and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[50] G. Battail, Does information theory explain biological evolution? Europhys. Lett. 40 (3) (1997) 343–348.
[51] R. Dawkins, The Selfish Gene, Oxford University Press, Oxford, 1976.
[52] R. Dawkins, The Blind Watchmaker, Longman, New York, 1986.
[53] J.B. Anderson, S. Mohan, Source and Channel Coding An Algorithmic Approach, Kluwer Academic Publishers, Boston, MA, 1991.
[54] R.E. Blahut, Theory and Practice of Error Control Codes, Addison-Wesley, Reading, MA, 1983.
[55] B. Lewin, Genes V, Oxford University Press, New York, NY, 1995.
[56] E.E. May, Analysis of coding theory based models for initiating protein translation in prokaryotic organisms, Ph.D. Thesis, North Carolina State University, Raleigh, NC, March 2002.
[57] J. Watson, N. Hopkins, J. Roberts, J. Steitz, A. Weiner, Molecular Biology of the Gene, The Benjamin Cummings Publishing Company, Inc., Menlo Park, CA, 1987.
[58] E.E. May, Comparative analysis of information based models for initiating protein translation in *Escherichia coli* K-12, M.S. Thesis, NCSU, December 1998.
[59] E. Szathmary, The origin of the genetic code: amino acids as cofactors in an RNA world, Trends Genet. 16 (1) (2000) 17–19.
[60] R.D. Knight, L.F. Landweber, Guilt by association: the arginine case revisited, RNA 6 (4) (2000) 499–510.
[61] L. Gold, G. Stormo, Translational initiation, in: F. Neidhandt, J. Ingraham, K. Low, B. Magasanik, M. Schaechter, H. Umbarger (Eds.), *Escherichia coli and Salmonella typhimurium*: Cellular and Molecular Biology, ASM Press, Washington, DC, USA, 1987, pp. 1302–1307.
[62] D.E. Draper, Translation initiation, in: F.C. Neidhardt (Ed.), *Escherichia coli* and *Salmonella*, Cellular and Molecular Biology, American Society for Microbiology, Washington, DC, USA, September 1999.
[63] T.D. Schneider, G.D. Stormo, L. Gold, A. Dhrenfeucht, Information content of binding sites on nucleotide sequences, J. Mol. Biol. 188 (1986) 415–431.
[64] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick, Coding model for translation in *E. coli* K-12, in: First Joint Conference of EMBS-BMES, Atlanta, GA, USA, 1999.

[65] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick, The ribosome as a table-driven convolutional decoder for the *Escherichia coli* K-12 translation initiation system, in: World Congress on Medical Physics and Biomedical Engineering Conference, Chicago, IL, USA, 2000.
[66] D. Frishman, A. Mironov, M. Gelfand, Starts of bacterial genes: estimating the reliability of computer predictions, Gene 234 (2) (1999) 257–265.
[67] M. Tompa, An exact method for finding short motifs in sequences, with application to the ribosome binding site problem, in: ISMB, Heidelberg, Germany, 1999.
[68] S.S. Hannenhalli, W.S. Hayes, A.G. Hatzigeorgiou, J.W. Fickett, Bacterial start site prediction, Nucleic Acids Res. 27 (17) (1999) 3577–3582.
[69] M.M. Walker, V. Pavlovic, S. Kasif, A comparative genomic method for computational identification of prokaryotic translation initiation sites, Nucleic Acids Res. 30 (14) (2002) 3181–3191.
[70] A. Zien, G. Ratsch, S. Mika, B. Scholkopf, T. Lengauer, K.R. Muller, Engineering support vector machine kernels that recognize translation initiation sites, Bioinformatics 16 (9) (2000) 799–807.
[71] B.E. Suzek, M.D. Ermolaeva, M. Schreiber, S.L. Salzberg, A probabilistic method for identifying start codons in bacterial genomes, Bioinformatics 17 (12) (2001) 1123–1130.
[72] T. Yada, Y. Totoki, T. Takagi, K. Nakai, A novel bacterial gene-finding system with improved accuracy in locating start codons, DNA Res. 8 (3) (2001) 97–106.
[73] J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, Nucleic Acids Res. 29 (12) (2001) 2607–2618.
[74] A.G. Pedersen, H. Nielsen, Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis, in: Proceedings of the International Conference on Intelligent Systems and Molecular Biology, Halkidika, Greece, 21–26 June 1997, pp. 226–233.
[75] Y. Osada, R. Saito, M. Tomita, Analysis of base-pairing potentials between 16S rRNA and 5′ UTR for translation initiation in various prokaryotes, Bioinformatics 15 (1999) 578–581.
[76] P. Bermel, E. Eni, M. Vouk, D. Bitzer, On the Import of the Shine–Dalgarno Series to the Expression of mRNA Sequences, Department of Computer Science, North Carolina State University, NC.
[77] D.L. Bitzer, M.A. Vouk, A table-driven (feedback) decoder, in: Tenth Annual International Phoenix Conference on Computers and Communications, Phoenix, AZ, USA, 1991, pp. 385–392.
[78] T.M. Barnes, Using Genetic Algorithms to Find the Best Generators for Half-Rate Convolutional Coding, North Carolina State University, Raleigh, NC, 1994.
[79] David I. Rosnick, Free Energy Periodicity and Memory Model for *E. coli* Codings, Ph.D. Thesis, North Carolina State University, Raleigh, NC, 2001.