Bits and Bases: An Analysis of Genetic Information Paradigms

Elebeoba E. May
Discrete Mathematics and Complex Systems Department
Sandia National Laboratories, Albuquerque, NM 87185 USA
(Email: eemay@sandia.gov)

Abstract— **The objective of this work is to form a general understanding of biological communication mechanisms by applying Shannon information theory and coding theory concepts to study the complex system of information transmission in biological organisms. We assess the viability of a biological coding theory framework by exploring coding theoretic characteristics of the genetic system that parallel traditional communication systems. We present results of channel capacity studies for prokaryotic and eukaryotic replication processes and explore connections between capacity and cellular aging.**

Keywords— **biological coding theory, genetic information, channel capacity, mutation**

I. Introduction

In his 1998 paper, "The Invention of the Genetic Code," Brian Hayes details the decade long pursuit to break the genetic code hidden inside the deoxyribonucleic acid (DNA) double helix discovered by James Watson and Francis Crick in 1953 [7]. In addition to Watson and Crick's work, Rosalind Franklin's research contributed significantly to the discovery of the structure of DNA. Hayes recounts how quickly a biochemical puzzle was reduced to an abstract problem in symbol manipulation. This all-important quest for a golden fleece of sorts attracted quantitative scientists, accomplished in their respective fields, including the physicist George Gamow and coding-theorist Solomon W. Golomb. Experimental evidence from Marshall W. Nirenberg and J. Heinrich Matthaei of the National Institutes of Health eventually led to the cracking of the genetic code. Unfortunately it also seemed to mark the end of fervent research into information and coding theoretic characteristics of biological organisms and processes.

During the past twenty years, there has been a renewed interest in the use of information and coding theory in the study of genomics [16], [6], [14], [19]. Coding theory has been used for frame determination, motif classification, oligo-nucleotide chip design, and DNA computing [2], [22], [11], [20], [8]. Additionally researchers, such as Hubert Yockey who performed fundamental investigations of coding properties of genetic systems, have explored the error control coding properties of genetic sequences [23], [10], [13], [12], [15].

From the early 1950s to the mid 1960s the focus of the genetic code-breaking enthusiast was understandably on the protein-coding portion of DNA (the region that contains triplet nucleotide bases that represent amino acids which constitute proteins). Non protein-coding DNA sequences, pejoratively referred to as "junk DNA" have until recently been overlooked. Scientist are finding that these sequences are far from "junk" but rather some serve regulatory roles for genetic processes including the control of RNA transcription and the initiation of protein translation. So it seems that the mystery of the double helix has not been completely unraveled. There is at least a second part to deciphering the information transmission protocols of biological systems, namely a need to translate the regulatory code of DNA.

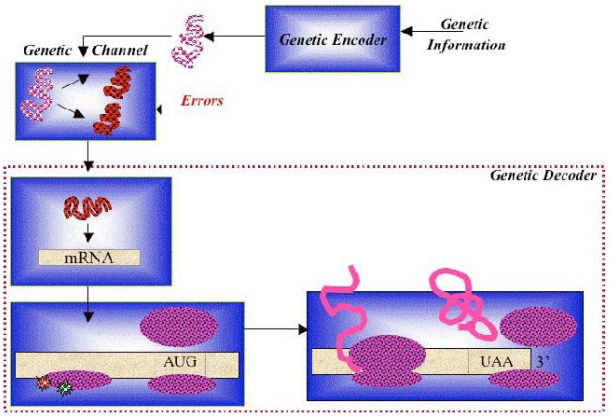

Fig. 1. The central dogma of genetics as a communication paradigm for molecular biology: DNA replication is the genetic channel; transcription, protein translation initiation, protein translation elongation plus termination constitute the genetic decoder.

Besides the potential impact on the biological sciences, learning how biological organisms are able to communicate their genetic message efficiently in the presence of noise can improve and advance current communication protocols. The underlying hypothesis of our work is that the genetic system can be operationally paralleled to an engineering communication system which transmits and operates on bases as opposed to bits. The central dogma of genetics, depicted in Figure 1 can then be viewed as the paradigm for biological communication, where an organisms redundancy containing DNA is the result of an error control encoding process. Replication is represented as the channel transmission process where transmission errors in the form of mutations are introduced into the DNA. Transcription and translation makeup the decoding process. The feasibility of applying this paradigm to the analysis of regulatory signals is the focus of current work.

A. Bits and Bases

A fundamental challenge for engineering systems is the problem of transmitting information from a source to a receiver over a noisy channel. This same problem exists in

978-1-4244-2110-7/08/$25.00 ©2007 IEEE

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? There are three general problems in communication, which we can loosely term packing, transmission, and security. Source codes (compression codes) help reduce the number of bits used to represent a message. Channel codes (error control codes) tackle the problem of efficiently transmitting the message over a noisy environment or channel. Cryptographic codes protect the message from eavesdroppers that can compromise the system. Years of theoretical research and funding have produced algorithms for addressing these challenges in engineering communication systems. If communication engineers recognize the necessity of protecting inorganic information, it is not difficult to imagine that organic systems also have a need to protect their genetic message - the key to their survival and the survival of the species. Our objective is to form a general understanding of biological communication mechanisms by applying Shannon information theory and coding theory concepts to study the complex system of information transmission in biological organisms.

To address the question of the existence (or non-existence) of an error control code in genetic sequences some researchers have searched for linear codes in DNA sequences [10], [15]. We take an alternate approach and analyze the channel capacity of replication to assess the necessity for error control to be incorporated into DNA and gain insight regarding the characteristics of such a code. Focusing on regulatory regions of DNA and RNA, and keeping in mind that the "cracked" or current genetic code (i.e., the nucleic acid codon to amino acid mapping) is error tolerant and redundant, hence an error control code, we theorize that the transmission of genetic information can be viewed as a biological, cellular communication system that employs some method of error control coding to protect and recognize valid information regions and to correct for "transmission" errors (see Figure 1). We assess the viability of the proposed framework by exploring coding theoretic characteristics of the genetic system that parallel traditional communication systems. Towards this end in the following sections we analyze the information capacity of replication using empirical mutagenesis data and conclude by discussing insights gained from our analyses in the final section of this work.

II. Information Capacity of the Genetic Channel

The capacity of the communication channel is a key system characteristic that governs the type of error control code used in transmission. In this work, we do not attempt to treat the general question of the capacity of molecular machines, which Schneider has elegantly addressed [17], rather we use mutagenesis data to quantify the capacity of select prokaryotic and eukaryotic organisms in order to gain insight regarding the potential for and nature of sequence-based genetic error control mechanisms.

The genetic communication system depicted in Figure 1 represents the replication process as the error introducing transmission channel. Shannon's channel coding theorem asserts that for a coding rate R less than the channel capacity C, there exists a channel code with rate $R = \frac{k}{n}$ such that the probability of a decoding error becomes arbitrarily small as n increases, where k is the number of information bits and n is the number of encoded bits [4], [21], [1]. The capacity of a transmission channel (the maximum data transmission rate) is dependent on the error rate of the channel p_{ij}, which is the probability of the channel transforming symbol i into symbol j for $i \neq j$. In order to determine appropriate error control coding parameters for genetic regulatory sequences, we characterize the replication channel in terms of the error rate (i.e., mutation rate) associated with replication. Mutations are replication errors that remain or are missed by genetic proofreading mechanisms. Mutation derived capacity values can suggest a genomic encoding rate $R_{Genetic}$ and from that plausible n and k values for genetic systems. We calculate the genetic channel capacity using mutation rates reported in Drake et al. [5]. Assuming the replication channel can be paralleled to a discrete memoryless channel (as illustrated in Figure 2), the capacity of the channel, C, is the maximum reduction in uncertainty of the input X given knowledge of Y [4]:

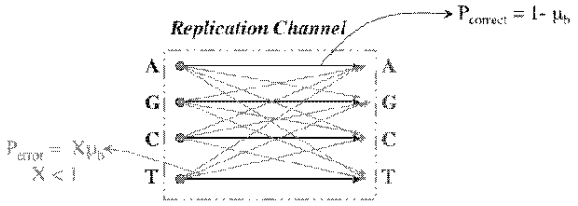

Fig. 2. Genetic replication channel as a discrete memoryless channel, where μ_b is the base mutation probability.

$$C = \max_{p(x)} I(X,Y) \tag{1}$$

where we select the maximum over all possible probability distributions, $p(x)$, of the input alphabet and

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2}$$

The Shannon entropy $H(X)$ and $H(Y|X)$ are defined as:

$$H(X) = -\sum_i p(x_i) log_2 p(x_i) \tag{3}$$

$$H(Y|X) = -\sum_k \sum_j p(x_k, y_j) log_2 p(y_j|x_k) \tag{4}$$

The probability $p(y_j|x_k)$ is the channel error probability. If $p(y|x)$ is specified by the error rate μ_b then $p(y_j|x_k) = \mu_b$, $\forall y \neq x$ and $p(y_j|x_k) = 1 - \mu_b$, $\forall y = x$ (where μ_b is the

mutation rate per base per replication [5]). The channel transition matrix, Table I, assumes all base mutations are equal, hence a transition mutation (purine to purine, $Adenine(A) \leftrightarrow Guanine(G)$ and pyrimidine to pyrimidine, $Cytosine(C) \leftrightarrow Thymine(T)$) and a transversion mutation (purine to pyrimidine, $(A, G) \rightarrow (C, T)$ and pyrimidine to purine, $(C, T) \rightarrow (A, G)$) are represented as equally probable. Our current method for calculating replication channel capacity does not distinguish between transversion and transition errors, hence Table I is a sufficient channel transition model.

TABLE I
CHANNEL TRANSITION PROBABILITY ASSUMING *p(Transition Mutation)=p(Transversion Mutation)*

	A	G	C	T
A	$1-\mu_b$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$
G	$\frac{\mu_b}{3}$	$1-\mu_b$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$
C	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$1-\mu_b$	$\frac{\mu_b}{3}$
T	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$\frac{\mu_b}{3}$	$1-\mu_b$

Initial capacity results based on μ_b, the single base mutation rate, suggested a near optimal transmission channel and very little reduction in the two bit capacity of the replication channel. This was misleading. Genome replication is not accomplished through a single use of the replication channel. The replication of a genome of size G requires G uses of the channel. Therefore, we can model genome replication as a channel with error probability μ_{b_G}, the probability of one or more errors in G uses of the channel. If X_i represents the transmitted base at channel use i and Y_i represents the corresponding received base, where $i = 1...G$, the error probability for the genome replication channel is related to μ_b and defined as follows:

$$\begin{aligned} \mu_{b_G} &= Prob(Y_i \neq X_i), for\ \ 1\ \ or\ \ more\ \ (X_i, Y_i)\ \ pairings \\ &= 1 - Prob(Y_i = X_i, \forall i) \\ &= 1 - (1-\mu_b)^G \end{aligned} \tag{5}$$

Equation 5 assumes $\mu_{b_i} = \mu_b$ for all i.

Figures 3, 4, and 5 show the capacity of the genome replication channel as a function of the log_{10} of the genome size for DNA microbes and prokaryotes (organisms include Bacteriophage M13, Bacteriophage Lambda, Bacteriophages T2 and T4, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Neurospora crassa*), higher eukaryotes (*Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* (mouse), *Homo sapiens*), and the effective genome of higher eukaryotes, respectively, using μ_b values from Drake et al. [5]. The effective genome is the portion of the genome where mutations, if they occur, can be the most lethal (i.e. genes or exons) [5]. Prokaryotic organisms have larger channel capacity values, ranging from 1.95 to 1.975 bits, than the higher eukaryotes with capacity values ranging from 0.4 to 1.85 bits. This suggests that for DNA microbes the maximum coding rate R is closer to $\frac{n-1}{n}$, leaving few bases for error control coding. In contrast, the channel capacity values for higher eukaryotes imply a distinctly smaller max-

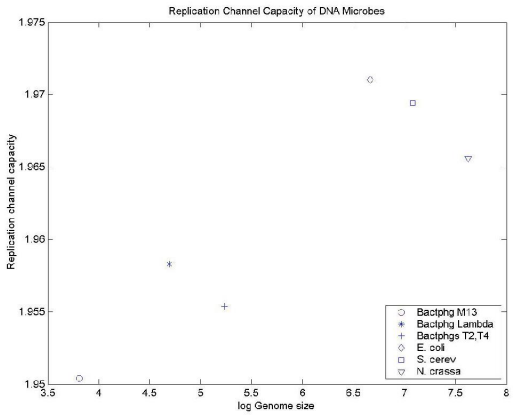

Fig. 3. Capacity of prokaryotic replication channels.

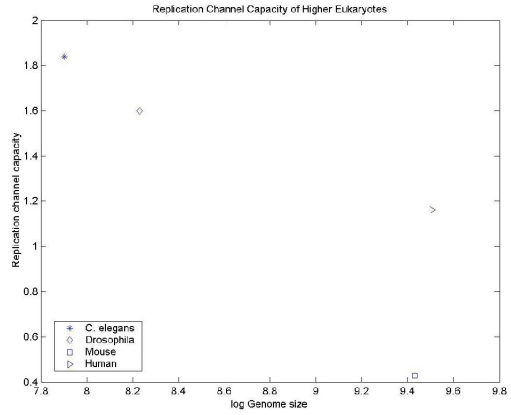

Fig. 4. Capacity of eukaryotic replication channels.

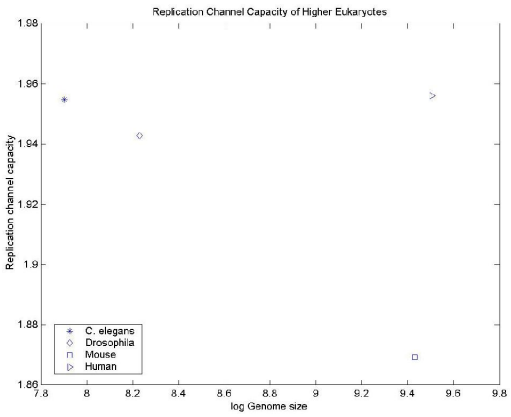

Fig. 5. Capacity of eukaryotic replication channels for the effective genome.

imum value for R, suggesting that eukaryotic genomes have more bases available for error control coding, i.e. more redundancy in their genome. It is generally accepted that eukaryotic organisms have a greater number of "extra" bases in their genome (bases that are not used to specify amino acids) than prokaryotes; our findings are consistent with this belief.

A. Replication Capacity and Cellular Aging

Cell division and mitosis are critical to the growth and survival of multicellular organisms. During mitosis a single cell produces two daughter cells that are identical copies of the parent cell. Vital to the successful production of daughter cells is error-free replication of the DNA contained in the chromosomes of the parent cell. Although biological replication is highly accurate with minimal error or mutation, if an error is introduced during the replication process in mitosis the error can propagate to daughter cells. As daughter cells become parent cells and replicate, additional mutations may occur during the generation of grand-daughter cells thereby reducing the overall fidelity of the original transmitted DNA. The number of times a chromosome is copied is bounded, which arguably limits the propagation of error-containing DNA. Telomeres, the ends of chromosomes, are shortened during each cycle of cell division. Once a cell's chromosomes are shortened to a critical length, that cell can no longer produce daughter cells nor propagate any accumulated mutations [9]. The enzyme telomerase prevents the shortening of telomeres. In normal, adult somatic cells, telomerase is turned off but in some cancerous cells, the telomerase gene is reactivated.

Given the probability of reduced replication fidelity as the number of times a chromosome is copied increases, and the existence of biological mechanisms that prevent the continued transmission of error containing chromosomes, it may be possible to view aging and related mutation engendered diseases as inevitable communication failures. Extending the genome replication channel model, it is evident that for a fixed $\mu_b < 1$, as G increases the quantity $(1-\mu_b)^G$ decreases. Consequently the channel error probability $(1-(1-\mu_b)^G)$ increases. The result is a reduction in overall channel capacity. Equation 6 is a simplified representation of the probability of error for an organism's replication channel after N_{CD} cell divisions. We equate multiple cell divisions to the transmission of a genome of size $G * N_{CD}$, where N_{CD} is the number of cell divisions. This line of reasoning parallels Battail's statement that "... the number of errors in a k-symbol message replicated r times is the same as that in an $(r*k)$-symbol message without replication[3]."

$$\mu_{b_G} \quad = \quad 1-(1-\mu_b)^{G*N_{CD}} \tag{6}$$

As illustrated in Figure 6 there is a reduction in the replication channel capacity for $N_{CD} = 1 \ldots 75$ cellular generations for higher eukaryotic organisms, substantiating the need for error control within DNA in order to ensure the survivability of an organism and ultimately the species.

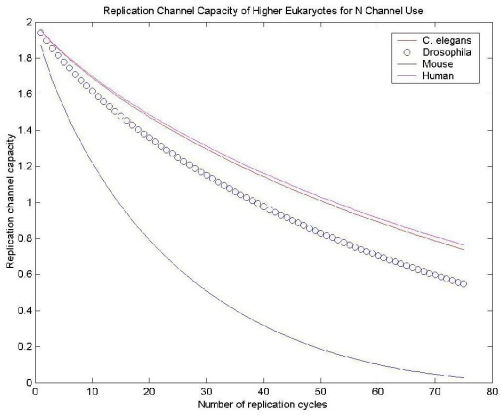

Fig. 6. Capacity of eukaryotic replication channel after M cell divisions.

III. Conclusions and Implications

Constructing parallels between engineering frameworks for communication and biological mechanisms is a challenge that if embraced can contribute to our quantitative understanding of complex biological systems. Our current coding theory analysis of biological systems and sequences has yielded compelling insights. The eukaryotic channel capacity, calculated based on mutation rates, is consistently lower than that of the prokaryotes examined. This leads to the supposition that the achievable coding rate, $R = \frac{k}{n}$, for the prokaryotic channel should be larger than that for the eukaryotic channel. Intuitively we can assume that for larger coding rates k and n are closer in value than for smaller coding rates, which implies less redundancy in the larger coding rate. Capacity calculations for prokaryotic and eukaryotic replication channels suggest that the channel coding rate for microbial organisms is closer to $\frac{n-1}{n}$ while eukaryotic systems have significantly lower, hence more redundant error control mechanisms. The obvious need for greater redundancy in eukaryotic species and the hypothesis that such redundancy is necessary for survival is supported when we consider the degradation of the replication channel over an organisms life span, measured by the number of cellular divisions. We observe that numerous replication cycles can reduce the fidelity of genome replication, further evidence that an appropriate error control code is required for reliable communication of an organisms blue print.

Our quantitative exploration of biological capacity has provided a unique insight and support for our assertion and those of numerous other researchers that error control is an integral part of genetic systems [23], [17], [18], [3], [12] . Given the efficiency of bacterial and viral organisms, we suspect that prokaryotic life forms may have achieved the Shannon limit for information transmission rates. If that is the case, research into biological coding methods could yield valuable returns not only for computational and experimental biology, but for communication engineering as well.

Acknowledgments

The author would like to thank John W. Drake of the National Institute of Environmental Health Sciences for providing additional insight with regard to mutation rates and mutagenesis studies.

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] John B. Anderson and Seshadri Mohan. *Source and Channel Coding An Algorithmic Approach*. Kluwer Academic Publishers, Boston, MA, 1991.

[2] Didier G. Arques and Christian J. Michel. A code in the protein coding genes. *BioSystems*, 44:107–134, 1997.

[3] G. Battail. Does information theory explain biological evolution? *Europhysics Letters*, 40(3):343–348, November 1997.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory* . John Wiley and Sons, Inc., New York, N.Y., 1991.

[5] John W. Drake, Brian Charlesworth, Deborah Charlesworth, and James F. Crow. Rates of spontaneous mutation. *Genetics*, 148:1667–1686, 1998.

[6] T. B. Fowler. Computation as a thermodynamic process applied to biological systems . *International Journal of Bio-Medical Computing*, 10(6):477–489, 1979.

[7] B. Hayes. The Invention of the Genetic Code. *American Scientist*, 86(1):8–14, 1998.

[8] L. Kari, J. Kari, and L. F. Landweber. Reversible molecular computation in ciliates. In *Jewels are Forever*, pages 353–363, 1999.

[9] Benjamin Lewin. *Genes V*. Oxford University Press, New York, NY, 1995.

[10] L. S. Liebovitch, Y. Tao, A. Todorov, and L. Levine. Is there an Error Correcting Code in DNA? *Biophysical Journal*, 71:1539–1544, 1996.

[11] David Loewenstern and Peter N. Yianilos. Significantly lower entropy estimates for natural DNA sequences. In *Proceedings of the Data Compression Conference*, 1997.

[12] D. MacDonaill. A Parity Code Interpretation of Nucleotide Alphabet Composition. *Chem Communic*, pages 2062–2063, 2002.

[13] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick. An error-correcting code framework for genetic sequence analysis. *Journal of the Franklin Institute*, 341:89–109, 2004.

[14] Ramon Roman-Roldan, Pedro Bernaola-Galvan, and Jose L. Oliver. Application of information theory to DNA sequence analysis: a review. *Pattern Recognition*, 29(7):1187–1194, 1996.

[15] G. Rosen and J. Moore. Investigation of coding structure in DNA. In *ICASSP 2003*, 2003.

[16] Rina Sarkar, A. B. Roy, and P. K. Sarkar. Topological information content of genetic molecules – I . *Mathematical Biosciences*, 39:299–312, 1978.

[17] Thomas D. Schneider. Theory of molecular machines. I. Channel capacity of molecular machines. *Journal of Theoretical Biology*, 148:83–123, 1991.

[18] Thomas D. Schneider. Theory of molecular machines. II. Energy dissipation from molecular machines. *Journal of Theoretical Biology*, 148:125–137, 1991.

[19] Thomas D. Schneider. Information content of individual genetic sequences . *Journal of Theoretical Biology*, 189:427–441, 1997.

[20] R. Sengupta and M. Tompa. Quality control in manufacturing oligo arrays: A combinatorial design approach. *Journal of Computational Biology*, 9(1):1–22, 2002.

[21] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, 1949.

[22] Nikola Stambuk. On the genetic origin of complementary protein coding . *Croatica Chemica ACTA*, 71(3):573–589, 1998.

[23] Hubert Yockey. *Information Theory and Molecular Biology* . Cambridge University Press, NY, NY, 1992.