ERC Report - February 19, 2010

Amir Hesam Salavati
E-mail hesam.salavati@epfl.ch

Supervisor Prof. Amin Shokrollahi
E-mail amin.shokrollahi@epfl.ch
Algorithmics Laboratory (ALGO)
Ecole Polytechnique Federale de Lausanne (EPFL)

May 5, 2010

1 Introduction

In the past week, I read some papers on applications of coding theory in biology, the summary of which is given in the next section.

2 Coding and Molecular Biology

During my research on applications of information theory in bioinformatics, I also encountered some promising applications for coding theory. Therefore, I have started studying possible applications of coding theory in biology. In particular, I would like to investigate possible coding patterns in **bioinformatics** and **neural propagation**.

Coding and information theory for biological systems are still in their infancy. At this point, there are more questions than answers in this area. The redundancy in genome and DNA is an accepted fact. However, whether this redundance is because of an error correction code is still not known. Therefore, Development of coding-theoretic frameworks for molecular biology is an ongoing endeavor [2]. We need to search for classical and quantum codes

and, in some cases, different types of coding [13]. According to [13], we can divide the field into four groups, according to their research subject"

1. Applications of information theory to biology
2. Existence of error correction in biological information processing
3. Applications of coding theory to biomolecular computing (e.g., DNA computing)
4. Applications of coding theory to computational molecular biology and bioinformatics.

There are different mechanisms of information exchange in the genetic systems. Therefore, many researchers have attempted to model these processes from information theoretical point of view. Three of these models have become more popular. The first of is suggested by Gatlin [1]. In Gatlin's model, DNA is an coded sequence which encodes vital information. The decoded messages are the amino acids which build proteins in the end.

The second model is due to Yockey [3]. His model is based on data storage systems and Turing machine. DNA is assumed to be the input tape where sequence of bits are stored. The tape is feeded into the Turing machine equivalents, RNA molecules. The output is the amino acids, just like the Gatlin's model [1].

The third and the most promising model is the one suggested by May et. al. [4]. In their proposed framework, DNA is the output of and encoded that codes biological information. The DNA replication process is the communication channel which is used to transfer genetic data between two generations. The decoding process is done by RNA molecules after which amino acids are given as the output messages.

None of the above models discuss the origins of such error control coding. In fact, even the existence of such coding scheme is at question as no one has been able to prove its existence. Nevertheless, we have an ever increasing amount of evidence that suggest error control codes must be present in genetic systems. For instance, as Battail argues, who is also a coding theorist, error control becomes obvious when one notes that the number of errors in a k-symbol message that has been replicated r times is approximately equal to the number of errors in an unreplicated message with $r \times k$ symbols. Thus, in order to have a reliable message during the life cycle of an organism (let alone

during the evolutionary time scale!) the message must have good methods of error correction.

Another very interesting price of evidence lies in the proof-reading mechanism of DNA replication process. Proof-reading mechanisms are observed during DNA replication, and when the activity of these polymerase mechanisms are blocked, error rates increase from $10^{-}6$ to 10^{-3} [7].

As a matter of fact, a number of researchers have made an effort to identify block codes in the DNA [8], [7]. None of them were able to find such codes. However, as they have already mentioned, their approaches were too simplistic and limited. Moreover, neither of them addressed the existence of convolutional codes. In fact, Liebovitch et al. [8] suggest that a more comprehensive examination would be required.

Nevertheless, Rosen et al. have used an interesting approach in using finite fields to represent the four nucleotides and then use finite field arithmetic to find the parity check matrix of genetic code. They have assumed the existence of a linear $(n, n-1)$ code [7] and divide the whole DNA sequence into frames of n nucleotides. Using Gram-Schmidt algorithm, they identify the basis of the subspace formed by these vectors of length n. Having built the basis, the single vector of the parity check matrix could be found by looking for a base vector whose corresponding coordinate is zero when all vectors are expressed as a linear combination of basis.

Schmidt and May [10] exploited graph-theoretic methods to analyze error correction and detection properties of Escherichia coli K-12 translation initiation sequences. They first prove that in contrast to binary random sequences, binary block codes form distinctive cluster graphs. Then, they have applied their method to Escherichia coli K-12 translation initiation sequences. Their results show that non-initiation sites fail to cluster into distinct groups. However, cluster formations in valid initiation sequences are clearly observed, suggesting the possibility of existence of an error control coding mechanism in E. coli's translation initiation sites.

Another important point in analyzing coding properties of DNA is identifying system characteristics such as the channel capacity [2]. To calculate the capacity, we must have the error probability of the channel. This translates into mutation rate in May et al.'s model in which the DNA replication process is modeled as a communication channel. Some results on the rate of mutation in different species could be found in [11] and [12]. Based on these numerical values on mutation rate, channel capacities of different organisms are illustrated in figures 1 and 2 [2].

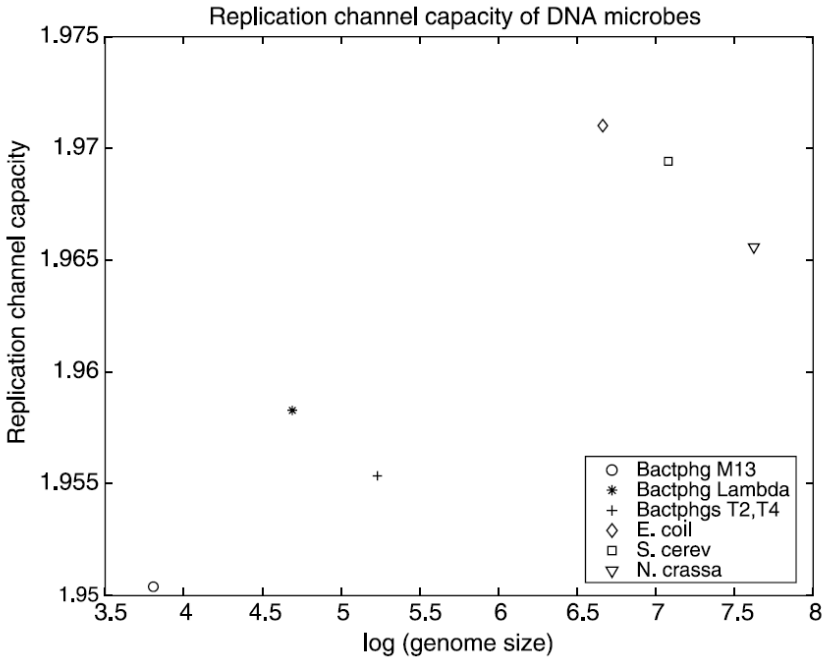

Figure 1: Capacity of prokaryotic replication channels [2].

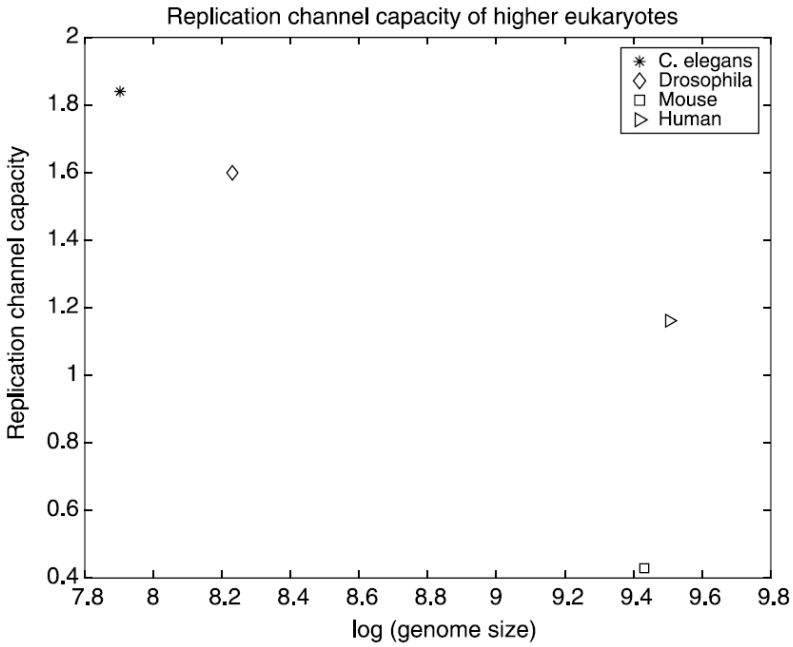

Figure 2: Capacity of eukaryotic replication channels [2].

At this point, an important question arises: even if the existence of error correction codes in DNA are proved, how can we identify the coding scheme? Several researchers have started to answer this question by proposing different coding methods and testing them against biological findings. May et al. [5], [6] have modeled mRNA as a noisy, systematic zero-parity encoded signal and the ribosome as an (n,k) minimum-distance block decoder. Their idea is to build the parity check matrix of the (n,k) code by finding the set of vectors that are orthogonal to codewords. Then choose the *best* possible parity check matrix. Here, the *best* matrix is the one with the highest fitness, which is a function of the number of zeros in H and its syndrome S.

The applications of spin glasses and statistical physics are quite well known in coding theory. On the other hand, there exists an interesting relationship between spin glasses and origin of life [13].

2.1 Iterative Decoding and Genomics

Another close relationship between coding theory and genomics is iterative codes such as LDPC and similar codes, in which we have a sparse matrix and fast decoding. The performance of the decoder improves over time and iteration by iteration. While this seems bizarre, there is growing amount of evidence that the coding techniques in genome, if any, are very similar to that of LDPC-like codes. For one, as Gupta points out [13]: "The need to be greedy speaks against the employment of long block and convolutional codes in the case of sensory perception/pattern recognition, too. However, there are now iterative decoders of the turbo code type for certain simple convolutional code and other code families (MDPC families, e.g.) with relatively short constraint lengths which have the advantage that a turbo (i.e., iterative, effectively maximum-likelihood-seeking) decoder (if there can be said to be a decoder in the brain and anyone can locate where) would be capable of producing a suboptimum decision at a moments notice. Moreover, this suboptimal decision would become increasingly close to optimum over time if it does turn out that the situation allows for sufficient time prior to decision making".

In order to understand how iterative decoding in genome works, one must become familiar with the concept of Regulator Network of Gene Interactions (RNGI). According to findings of biologists, genes form a network meaning that the activity of one gene affects those of others. More precisely, genes act as a switch: when one of them is expressed (switched on), it may also

switch some other genes on or off. This is what called RNGI [14].

There are various mathematical models for modeling RNGIs. However, the most widely used model is the so called NK-model [15], [16]. Here is how NK-model works: we have N genes and each gene is affected by K other genes. As already mentioned, these genes act as binary switches: they are either on or off. Their state depends on the activities of other genes. This control action is known as epistasis and is described in terms of a *genetic graph*. The nodes of this graph represent genes and a directed edge exists between two interacting genes. The average *incoming degree* of the nodes is K. For the purpose of analysis, one can describe the genetic graph by a bipartite graph, where the left hand side nodes represent genes, and the nodes on the right hand side represent control nodes for each gene. Genes connected to control node C_j exhibit an influence on the operation of gene G [14].

NK-model is a special case of Boolean Networks (BN) which has applications in biology, mathematics and communications. In fact, BNs contain LDPC codes as a special case [14].

There exist many biological indications that the proofreading mechanism of DNA transcription is intimately connected to the RNGl mechanism. Furthermore, malfunctions of the proofreading mechanism are known to cause disease such as cancer. Therefore, in [14] the authors conjecture that during the process of DNA replication, nodes of the genetic network are arranged in a special form that only allows valid genes as the codewords and an error-control code. If a mismatch occurs during the replication, some sort of fast decoding method is used to identify the erroneously copied genes. Then, these genes are undergone another level of internal error control which leads to determining the erroneous nucleotides. In other words, both genes and nucleotides are involved in the error control process. The global Genetic Error Control (GEC) code uses genes rather than three nucleotides as its symbol while the local (internal) GEC uses the nucleotides as its symbols and acts within a gene. As they have mentioned in their paper: "The error control capability of a genome does not lie in the DNA code structure alone, but primarily in the way genes interact. There are several biological characteristics that seem to support this conjecture, but the most interesting one is based on the binding patterns of nucleases. The base excision repair pathway that provides for most of the error control consists of arrays of enzymes known as nucleases. The nucleases enzymes monitor DNA for the presence of damaged binding sites. Nucleases jump from one position in the genome to another,

in order to collect information about the erroneous sites. Hence, the genome seems to be globally connected and the proofreading code could operate on the global structure".

References

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

[2] E. E. May, "Error Control Codes and the Genome", in: M. Akay, "Genomics and proteomics engineering in medicine and biology", Wiley-IEEE Press, 2007.

[3] H. Yockey, "Information Theory and Molecular Biology", Cambridge University Press, New York, 1992.

[4] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, An error-correcting code framework for genetic sequence analysis, J. Franklin Inst., 341: 89109, 2004

[5] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, Coding model for translation in E. coli K-12, Paper presented at the First Joint Conference of EMBS-BMES, Atlanta, GA, 1999. 49.

[6] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, Coding theory based models for protein translation initiation in prokaryotic organisms, BioSystems, 76: 249260, 2004.

[7] G. L. Rosen, J. D. Moore, "Investigation of Coding Structures in DNA", Proc. IEEE int. conf. Acoustics, Speech, and Signal Processing, 2003, vol. 2, pp. 361-364

[8] L. S. Liebovitch, Y. Tao, A. Todorov, and L. Levine, Is there an error correcting code in DNA? Biophys. J., 71: 15391544, 1996.

[9] G. Battail, Does information theory explain biological evolution? Europhys. Lett., 40(3): 343348, November 1997.

[10] D.C. Schmidt and E. E. May, Visualizing ECC properties of E. coli K-12 translation initiation sites, Paper presented at the Workshop on Genomic Signal Processing and Statistics, Baltimore, MD, 2004.

[11] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow, Rates of spontaneous mutation, Genetics, 148: 16671686, 1998.

[12] A. Bebenek, G. T. Carver, H. Kloos Dressman, F. A. Kadyrov, J. K. Haseman, V. Petrov, W. H. Konigsberg, J. D. Karam, and J. W. Drake, Dissecting the fidelity of bacteriophage RB69 DNA polymerase: Site-specific modulation of fidelity by polymerase accessory proteins, Genetics, 162: 10031018, 2002.

[13] M. K. Gupta, "The Quest for Error Correction in Biology", IEEE Engineering in Medicine and Biology Magazine, Vol. 25, No. 1, pp. 46-53, 2006

[14] O. Milenkovic, B. Vasic, "Information Theory and Coding Problems in Genetics", ITW 2004, San Antonio, Texas, October 24.29,2004

[15] S. Kauftinan, Adaptation on Rugged Fimess Landscapes, in Lectwzs in the Science of Complexity, SFI Studies in the Science of Complexity, Lecture Volume I, ed. D. Stein, pp. 527-618, Addison-Wesley, Redwood City, 1989.

[16] Y. Gao, J. Culberson, "An Analysis of Phase Transition in NK Landscapes", Journal of Artificial Intelligence Research 17 (2002) 309-332