ERC Report - March 3, 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)

March 3, 2010

1 Introduction

During the past week, I studied more papers on the applications of coding theory in genetics and bioinformatics. The summary of these papers are given in the next section. Based on these works, it seems that coding theory has much more applications in genetics other than searching for error control coding mechanisms in DNA.

This report is divided into several subsections, based on the topic of the paper and the application of coding theory in bioinformatics. At first, I will discuss the works that address the existence of an error control coding method in genome. Later on, I will explain other possible applications of coding theory in molecular biology.

2 Coding Theory and Existence of Error Correction Codes in DNA

In the previous report, I mentioned that there is an ever growing amount of evidence suggesting the existence of an error control coding method in DNA.

In addition to the works mentioned in the previous report, there are some other papers dedicated to this subject.

Let me start by the paper that I think is one of the most interesting papers on this topic since its author is a coding theorist and not a biologist [2]. In this paper, the author argues that there are a lot of evidence suggesting the existence of error correction mechanisms in genome. The first argument in favor of such a coding is the fact that mutations in the genome replication due to chemical agents or radiations, are responsible for aging and certain diseases like cancers. Noting that genetic data is replicated several million times in evolutionary time scale, if there were no error correction mechanism, the accumulation of errors during periods million times longer would simply make genetic communication, and hence life, impossible.

Moreover, it is quite surprising that while the process of DNA replication occurs in a noisy environment, the cell, the replication error rate is as low as 10^{-9} mutations/nucleotide. This value is noticeable as DNA replication procedure alone has an error rate of 10^{-3} to 10^{-5} [11]. One might argue that the final low error rate is a result of DNA's internal proofreading mechanism: when copied, the helical structure unzips into two separate strands. RNA uses one strand to read DNA and then check the read sequence with the other strand. If they matched, it proceeds. Otherwise, it waits until the correct nucleotide is restored. This simple proofreading reduces the error rate to approximately 10^{-10}. Moreover, there are other proofreading mechanisms as well.

However, the proofreading mechanisms can at best ensure that the copy is faithful to the original. In other words, they can correct the errors which occur within the replication process but not those that may affect the original itself [2].

Another piece of evidence that suggests the role of evolution and natural selection in devising error correction methods in DNA comes from the fact that error rates are higher in simple species such as viruses and bacteria compared to that of highly developed ones such as mammals [2].

In [11], the author develops a method to uncover an error correction coding structure in the nucleotide sequence. While [11] contains some new ideas, the main results of the paper are the ones presented in the authors' earlier paper [12]. The main question that the author tries to answer is: how does DNA protect itself from error? The author suggests a method for finidng the answer of this question and if there is an error correcting code in genome. The main idea behind her approach is to check the dimensionality

of n-tuples in genome. If there is a coding method there, then there must be redundancy among the n-tuples and the dimensionality must be less than n. **Generally speaking, any method for finding the dimensionality of a subspace in a space defined by n-tuples would be useful in this area.** More specifically and as mentioned in the paper, **there is a need for a general approach to find k-parity bits placed in *any order* in *any n-size code* to discern an (n, k) block coding structure from a DNA sequence**.

Nevertheless, there are a lot of difficulties in the case of genome sequence. First of all we neither know n nor k. Furthermore, we do not even know if the coding scheme, if it exists, is a block code. As suggested by May et. al [9], convolutional codes seem to be a better model in certain cases. In addition to all these problems is the reading frame issue. In a traditional coding method, the decoder knows the beginning of each new codeword. However, in a genome we have no clue about the beginning of a frame. Assuming codewords with length n, we must consider all possible reading frames from the beginning of the genome (total of n possible ways).

While the idea considered in [11] is very interesting and promising, their model is quite unrealistic (That's probably the reason they have not yet found any sign of linear coding in genome of primal species such as E. Coli bacteri). For one, the author has assumed that the whole genome is a coded sequence and divided it into N vectors (codewords) with length n. She has then proposed a novel approach to compute the dimensionality of the resulting $N \times n$ matrix, obtained by putting the codewords and the rows of the matrix. However, in my opinion, the assumption that the whole genome is a coded sequence is rather simplistic and it is probable that some parts are coded while the other parts have left uncoded for evolutionary reasons.

On other fronts, Mac Donaill has suggested a parity check code interpretation of nucleotide composition [8]. I have not yet read this paper and its details are to be overviewed in my next report.

Another very interesting subject is mentioned in [6] where authors suggest that the current assignment of codons to amino acids are based on error minimization criteria. In other words, they claim that natural selection has chosen amino acid to codon assignment such that errors in translation process are minimized.

In order to prove their claim, the authors have first defined a measure to evaluate the codewords of a code quantitatively, just like the hamming weight in coding literature (in this case, the authors have used polarity of

the corresponding amino acid as the measure). Then, given a codon, they do a single mutation in the codon and "decode"[1] the corresponding amino acid of the resulted codon. They then measure the distance between the new and original amino acids to asses the strength of the coding method. The ideal case is that a single mutation does not change the resulting amino acid. By repeating this process for all of the three nucleotides in a codon, and for all codons, and then averaging the results according to a weight function, we will obtain measure of how good a code is. The lower this measure is, the stronger the code is.

The authors have then build many random "codes" by arbitrarily assigning codons to codewords and evaluated their strength based on the aforementioned measure. Their result show that only 1 of codes (from a pool of 1 million random codes) performed better than the standard genetic code. Therefore, their work clearly suggest that the codon assignment is optimized by evolution in a way to minimize translation errors.

Furthermore, as mention in [3], the codons which "code" for one amino acid are more closely related to one another (in sequence) than they are related to codons that code for other amino acids. In other words, codons that code for one amino acid differ in several cases by just one nucleotide. Thus, single nucleotide mutations (especially in the third location) will often not change the resulting amino acid.

Nevertheless, there are a few drawbacks in the approach of [6]. First of all, the strength of code is their method is completely dependent on the feature used to measure the codewords. In other words, and as the authors have mentioned themselves int heir paper, while some properties may hint for support of existence of error correcting codes, it may not happen for other properties. **Therefore, it is of outmost importance to choose the measure wisely.**

A further argument in favor of the existence of ECC in genome comes from the fact that this hypothesis in genome helps us explain some puzzling phenomena very easily, which otherwise could not be explained simply [2]:

- The fact that the species are discrete and the fact that evolution proceeds by jumps: a puzzling fact in biology is the discreteness of species.

[1] I have put decode between quotation marks because in genomics literature, coding and decoding sometimes refer to the translation procedure in DNA where a gene is "decoded" to build the corresponding protein.

Why don't we have a *spectrum* of living things instead of having different species and families? Error correction codes could explain this phenomena quite easily and in a neat matter: small number of mutations (errors in close distance of current codewords) are corrected while the ones with larger distances are left uncorrected. Hence, they result in new species! In other words, genomes located in a distance less than that of the minimum distance of the code can not exist in outside world!

- The trend of evolution towards increased complexity: longer and more complex genomes means better error correction (as the length of the genome, i.e. codeword, goes to infinity, coding becomes better, which is a well known fact in coding literature.)

2.1 Soft Codes and Genetics

Battail has also introduced the concept of *soft codes* in genetics [2]. Basically, there are two ways of defining a code: specifying its construction rules (as communication engineers do) or specify the required constraints of the codewords. He suggests that the second approach is more appropriate for natural phenomena in which the assumption of deterministically specifying the construction rules seems nave.[2] Moreover, the constraints are expressed in a probabilistic matter. Therefore, the main parameters of a code, like its minimum distance, then become random variables.

Interestingly, this resembles some works on random codes (like the one we are working on with Raj and Payam) in which the generator and parity check matrices are not fixed but are drawn randomly from an ensemble of codes for which a general constraint, such as the weight of each column in generator or parity check matrix, is specified in advance.

To get a better understanding of soft codes, think of natural languages. In a natural language, there are a lot of constraints such as the properties of the vocal tract, which limits the number of possible words, constraints on the meaning of the actual words (lexicon) out of the pool of possible words, constraints of having meaningful sentences and so on. In other words, phonetics, lexicon, syntax (grammar) and overall meaning of the sentence are the constraints on natural languages. At the same time, natural languages have a great capability of error correction (both in oral and written forms).

[2]In my opinion, the first approach is similar to fixing the generator matrix while the second one is to determine the parity check matrix, in a probabilistic manner

Furthermore, a language is defined by distinct constraints acting at several hierarchical levels. For instance, phonetic constraints, which are due to the structure of the vocal tract, are more fundamental than constraints specific to a given language, which are social conventions inherited from history. The same level of hierarchy and constraints are also present in DNA: chemical constraints on nucleotides and their pairings, the lexicon (genes) and meaningful (functional) proteins. These are examples of what is called *nested soft codes* [2].

In a nested code, some parts of data are protected more than other parts. In other words, some parts of the data are first coded and then the coded sequence is coded again using a possibly different coding approach and so on (similar to raptor codes!).

There are some evidence that the error correcting mechanism in genome, if any, is a nested code. In fact, vital genes need much more protection. In fact, these genes are preserved among generations of far-related species. For instance, the HOX genes which determine the organization plan of living beings are shared by, e.g., humans and flies, which diverged from a common ancestor hundreds of millions of years ago [2]. A similar scheme has independently been used by Barbieri to describe the organic codes [1].

2.2 Searching for ECC in DNA

Searching for error correction codes is much more difficult than one might imagine. First of all, what is the coding alphabet? The trivial answer to this question is that the alphabet are the four nucleotides used in construction of DNA. However, this is similar to saying that English alphabet is composed the set of vertical and horizontal lines that are used to build the actual letters. My colleagues and I have worked on this issue before and our results suggest that the letters of genetic language, if there is any language, is actually not the four nucleotides but a combination of them, just like the situation in natural languages [13].

Another challenge comes from the codes being nested. If this hypothesis is correct, i.e. the ECC in genome is nested, the set of alphabets used in each of the encoders in a nested code may be different from the other codes. Therefore, we have to find set of physiologically meaningful alphabets.

3 Coding Theory and Modeling Gene Regulatory Networks

Gene Regulatory Networks (GRNs) (or Regulator Network of Gene Interactions (RNGI), as I stated in my previous report) are an important field in system biology [7] because complex interaction patterns among genes in a genome has a big impact on our understanding of several biological procedures including disease development (particularly cancer).

In the most general model of GRNs, the influence of genes $1, \ldots, n$ on gene i is captured by the following equation:

$$\frac{dv_i}{dt} = f_i(v_1, \ldots, v_n) \tag{1}$$

where f_i is an arbitrary function depending on the model and v_j represent the expression level of gene j.

Recently, a new idea has been introduced, mainly due to Milenkovic[3] and her team (see for example [4]), which suggests novel applications of coding theory in genetics. While investigating the possible existence of an error control coding mechanism in DNA seems very interesting and promising, the mechanisms based on which this code is implemented in genome is not known.

As I mentioned in my previous report, Milenkovic et. al [10] have suggested that iterative decoding methods might be present in GRNs. In a typical genome, genes form a network in a way that they act as a switches: when one of them is expressed (switched on), it may also switch some other genes on or off. This is what called GRNs. As mentioned before, in [10] 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.

However, the applications of coding theory in GRNs is not limited to iterative decoding. Based on a recent work by Milenkovic and her team,

[3] I intentionally mentioned her name here since her areas of specialty and research interests are very close to what we are working on, both from a traditional coding theory viewpoint (LDPC coding, coding for storage systems, etc.) and applications of coding and information theory in genetics. Here is her website: http://faculty.ece.illinois.edu/milenkov

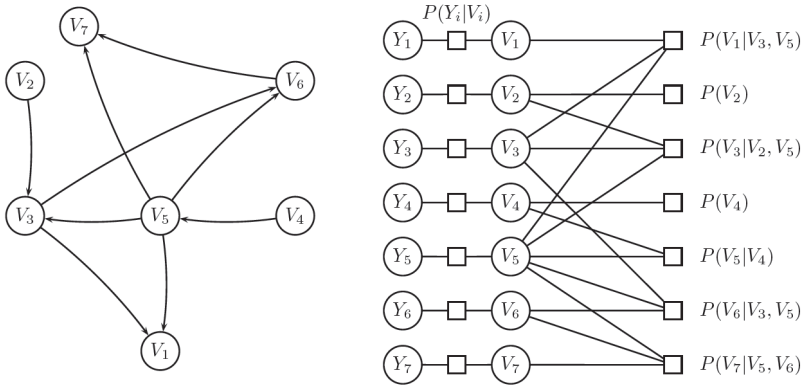

Figure 1: Gene network and its corresponding gene factor graph [4]

coding-theoretic methods can also be used to refine current models of GRNs. Here, the main problem we are interested in is inferring the structure of a GRNs from biological findings. There are various approaches for this purpose (for a brief survey see [4]). One of these methods is the probabilistic method in which Bayesian networks are used to model the behavior of genes. In this model, the directed interaction network is transformed into a bipartite factor graph. In the factor graph, the set of left vertices represent the nodes in the original network and the set of right nodes indicate the way the nodes interact. For example, if there is an edge between the vertices L_i and R_j on the left and right side, respectively, then gene V_i affects gene V_j, i.e. if V_i is switched on, V_j is switched on or off. This transformation is shown in figure 1. In the figure, the variables Y_i represent the measured expression levels of the genes and one can think of $P(Y_i|V_i)$ **as describing the unknown and noisy measurement channel.**

In [4], authors propose a coding-theoretic method to model the interactions between genes. They assume that the interaction network is known from biological findings, which is not a very limiting assumption, while the interaction functions f_i's in equation (1) must be determined. They use polynomial interpolation techniques in coding theory to determine f_i's. In their

approach, f_i is related to the polynomials used in Redd Muller codes and show that the same approach used in decoding of Redd-Muller codes could be used here to determine f_i if the number of observed biological data is sufficient, i.e. it is bigger than a threshold which depends on the minimum distance of RM codes.

Therefore, coding theory is used here to fully specify the parameters of the GRN model. interestingly, their approach has the advantage that it works in the noisy models in which the GRN is not a deterministic network but a probabilistic model which correctly reflects the stochastic behavior of biological phenomena. Furthermore, the data we use in our biological analysis comes from DNA micro-arrays which scan the DNA of species. The micro-array itself introduces some noise in the data as well.

4 Coding Theory for DNA Classification

Finding out the ancestors of current species and building the phylogentic tree of life is a very important filed in bioinformatics. The main problem here can be stated as follows: we would like to a build a tree in a way that its leaves are current species. The one before last level, i.e. the parents of these leaves, are the ancestors of common species. For example, monkey, chimpanzee and humans probably have a common ancestor which be their parent in the tree of life. This ancestor would be the parent node of these leaves in the phylogenetic tree, and so on until we end up with the first living species on earth as the root of the tree. Since most of these ancestors are extinct now, we would like to construct this tree merely based on the genome of current species.

There are a lot of algorithms for constructing the phylogenetic tree, all of which are heuristics as the problem is NP-complete. Nevertheless, coding theory can have a very nice application here **if we model the evolution as a communication channel in which an input (the ancestor of some species) is transmitted via the noisy communication channel (evolutionary time in this case). We have then sampled the outputs of the channel at different times (different noise values) to get different species. Our goal is to deduce the transmitted data based on the received noisy channel output.** A similar idea is mentioned in [3]. In this case, noise is mutations that occur in a genome of a species during evolutionary time scales.

5 Conclusion

Based on the literature review so far, it seems that there are a lot of evidence on existence of error correction schemes in DNA. Furthermore, no contradictions were found between the hypothesis that natural genomic error-correcting means exist and the properties of the living world. On the contrary, it seems to account for a number of facts, especially of evolution, that conventional theories fail to explain. In addition, as discussed above, coding theory could have many other applications in this field other than finding ECC in DNA.

Nevertheless, in my opinion it is quite soon to conclude that ECC exists in genome. Because all the phenomena that could be explained using coding theory could also be possibly justified by the help of other hypotheses. Nevertheless, coding theory is a strong candidate here and requires a deeper and more comprehensive study, which I will do in the following weeks.

References

[1] M. Barbieri, "The Organic Codes", Cambridge, UK: Cambridge Univ. Press, 2003.

[2] G. Battail, "Should Genetics Get an Information-Theoretic Education? ", IEEE ENGINEERING IN MEDICINE AND BIOLOGY MAGAZINE, pp. 34-45

[3] Z. Dawy, P. Hanus, J. Weindl, J. Dingel, F. Morcos, "On Genomic Coding Theory", European Transactions on Telecommunications, Volume 18 Issue 8,Pages873-879, 2007

[4] J. Dingel, O. Milenkovic, "Coding-Theoretic Methods for Reverse Engineering of Gene Regulatory Networks", Proc. IEEE Information Theory Workshop, 2008, pp. 114-118.

[5] I. Gat-Viks, A. Tanay, D. Raijman, and R. Shamir, A probabilistic methodology for integrating knowledge and experiments on biological networks. J Comput Biol, vol. 13, no. 2, pp. 16581, mar 2006.

[6] S. J. Freeland, T. Wu, N. Keulmann, "The Case for an Error Minimizing Standard Genetic Code", Journal of Origins of Life and Evolution of Biospheres , Volume 33, Numbers 4-5 / October, 2003 , pp. 457-477

[7] H. de Jong, Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol, vol. 9, no. 1, pp. 67103, 2002.

[8] D. A. Mac Donnaill, "Why nature chose A, C, G and U/T: An error-coding perspective of nucleotide alphabet composition", Origins of Life and Evolution of the Biosphere, 33:433455, October 2003.

[9] 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.

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

[11] G. L. Rosen, "Examining Coding Structure and Redundancy in DNA", IEEE engineering in medicine and biology magazine, 2006, pp. 62-68

[12] 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

[13] Amir Hesam Salavati, Masih Nilchian, Mahnoosh Alizadeh, Saeid Bagheri, Mohammad Javad Emadi, HAdi Kiapour, Mehdi Sadeghi, Mohammad Reza Aref, Kaveh Kavousi, Mehdi Pakdaman, Genome Alphabet: Are the Letters of Genome Language the Four Nucleotides or a Combination of Them?, To be submitted to the Journal of Theoretical Biology