Applications of Coding Theory in Biological Systems

Amir Hesam Salavati, Algorithmics Laboratory (ALGO), I&C

***Abstract*—Reliable information exchange and processing is a crucial need in both biological and artificial systems. Without such reliability, living beings do not have much chance of survival. In artificial information processing systems, coding methods play an important role to guarantee the required reliability. Whether coding-theoretical arguments are also useful in analyzing similar biological systems is the subject of this report.**

In what follows, we will discuss some applications of coding theory in molecular biology and neuroscience. These applications include analyzing error correction algorithms in biological systems as well as the use of coding-theoretical models to analyze various mechanisms in living systems other than error correction. We also consider some biological error correction mechanisms as an inspiration to design better and faster decoders. Based on the mentioned approaches, several research topics for future works is outline, both in neuroscience and molecular biology.

***Index Terms*—Coding theory, Molecular biology, Neuroscience, Gene regulatory networks, Associative memory, Neural networks**

I. Introduction

At the first glance, coding theory and biology seems to be two completely unrelated topics. The first is a subsidiary of math and communication engineering while the second is the science of studying living beings. The former mainly deals with mathematics while the latter involves experiments over living species. At best, their relationship seems to be unidirectional in which coding theory plays a small role in engineering artifacts and devices to perform those experiments.

However, this simple description does not explain everything and the picture is changing very fast. In fact, biological systems and processes have common needs with engineered systems in many cases. The necessity of acquiring and exchanging information in an efficient and reliable manner is one of those needs. This information could the prescription of living and survival, as in DNA, or it could be information about the outside world transmitted through and processed by the neuronal system. In all cases of biological systems, be it as simple as a single cell or as complex as humans Central Nervous Systems (CNS), having a reliable efficient mechanism for exchanging and processing information is a crucial need.

In contrary, the medium through which this information exchange occurs is not reliable at all. In the case of molecular biology, the medium is the cell cytoplasm which is an unfriendly noisy environment for DNA replication process [3] and in the case of the CNS, the whole path is noisy: outside world, where information is generated, the sensory system, and neurons, through which information is transmitted, are all noisy environments [16]. For one, only 10 to 30 $\%$ of pre-synaptic spikes lead to a post-synaptic response.

Nevertheless, we enjoy quite an accurate processing system as well as a fairly reliable DNA replication process. Having such degree of reliability in presence of noise is very amazing from engineering perspective and that is where coding theory comes into play. There are two main objectives in collaboration of coding theory with biology:

1) To help biologists understand and analyze different biological processes, including the way information is reliably processed and exchanged in biological systems, by employing methods used in coding theory.
2) To inspire from error correction mechanisms implemented by living systems to design more accurate and practical codes for artificial communication systems.

Exploring the above two avenues is the main goal of this report. As discussed earlier, there are two different biological systems which are our main interests: DNA replication and neural error correction mechanism. For the former case, one must first show that error correction mechanisms in fact exist in biological sequences, i.e. DNA and proteins. Providing some evidence in favor of the existence of error correction mechanisms is the subject of [3] which is one of the three papers we will discuss in detail.

Proposal submitted to committee: June 21^{st}, 2010; Candidacy exam date: June 28^{th}, 2010; Candidacy exam committee: Prof. Bernard Moret, Prof. Amin Shokrollahi, Prof. Wulfram Gerstner.

This research plan has been approved:

Date: ____________________

Doctoral candidate: ____________________
(name and signature)

Thesis director: ____________________
(name and signature)

Thesis co-director: ____________________
(if applicable) (name and signature)

Doct. prog. director: ____________________
(R. Urbanke) (signature)

Having compelling evidence on the existence of error correction mechanisms in molecular biology, we can use coding theory to investigate such mechanisms. This is discussed in [14], in which the authors demonstrate the close relationship between statistical physics, iterative coding theory and Gene Regulatory Networks (GRNs). We will discuss their approach extensively in the next section.

While we have to show the existence of error correction mechanisms in molecular biology, we already know that error control methods are implemented at least at the network levels of CNS, i.e. in brain. Just think of the cases when you figure out the meaning of a misspelled word very easily. Therefore, although neural code is noisy at fine levels, it must be deterministic at higher levels because the same stimulus should result in the same behaviorial decisions [5].

Nevertheless, whether such mechanisms exist at the neuronal level is still not known. An interesting method with possible applications in discovering coding structures from a pool of recorded data is discussed in [5] which we will discuss in the following section. Regarding the error correction at the higher levels of CNS, we will briefly address the work of She and Cruz [20] on noise tolerant associative memory later in the report.

In addition to the works mentioned above, we will also briefly discuss several other related works, explaining the applications of coding theory both in molecular biology and neural systems. These applications mainly concern channel coding. However, we will shortly consider source coding as well.

The rest of this report is organized as follows: the three assigned papers are discussed in section II. In sections III and IV, we briefly explain some related works in molecular biology and neuroscience, respectively. The proposed research plan is given in section V. Finally, section VI concludes the report.

II. Summary of the Three Papers

In this section, we will study the three assigned papers in more detail.

A. Should Genetics Get an Information-Theoretic Education? [3]

The main purpose of this paper is to initiate mutual collaboration between geneticists and coding theorists by giving various pieces of evidence on existence of error correction mechanisms in DNA. The author proposes a new hypothesis, suggesting that error correction mechanisms exist in DNA, and analyzes the consequences of such hypothesis. By comparing the predictions of the proposition with experimentally verified facts, he shows that the model is actually very compatible with biological findings.

Having demonstrated the consistency of an error control model with reality, the author proceeds by proposing a new framework, called soft codes, to investigate such error control structures. In what follows, we will first review the pieces of evidence suggesting that there are error correction means in DNA. The concept of soft codes is discussed later in this section.

1) Evidences on Existence of Error Correction Mechanisms in DNA: Occurring in the cell's noisy environment, DNA replication process is not error-free. In many cases, mutation due to chemical agents or radiation results in certain diseases, like cancer, and is also responsible for aging. DNA is replicated several million times in a lifetime of a species and if there were no error correction mechanism, the accumulation of errors during its lifetime and, on a larger scale, over millions of years of evolution would simply make genetic communication, and hence life, impossible [3].

It would then be surprising to find out that the total error rate in decoding genes is as low as 10^{-9} per nucleic base and per replication. This value is noticeable as DNA replication procedure alone has an error rate of 10^{-3} to 10^{-5} [17]. 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. DNA polymerase uses one strand to copy DNA and the other one to check the copied segment. Although this simple proofreading reduces the error rate to approximately 10^{-6}, it is still not sufficient to explain low error probabilities observed experimentally.

Furthermore, the aforementioned and other similar proofreading methods can only ensure the fidelity of replication process. They can not correct errors occurring in DNA itself. Note that in any case, the capability of any error correction mechanism is limited and if the number of errors increase a certain threshold (denoted as the code's minimum distance) the errors could not be detected or corrected. Therefore, mutation could be viewed as *uncorrected* errors in this regards.

Taking into account that mutation is necessary for natural selection, we are interested in error control methods that are neither very strict, such that they correct most of probable errors and prevent mutations completely, nor very weak, in the sense that they result in an extremely unreliable replication process.

Another interesting argument comes from evolution. Because of natural selection, only the well-fitted species survive. However, having a well-fitted phenotype is only one side of the story. In order for a well-fitted living being to survive, it should also preserve its genome. This seems to be impossible with an unreliable DNA replication process. In other word, a well-fitted genotype (which represents a well-fitted phenotype) only survives in light of a reliable replication process.

A further argument in favor of the existence of Error Correction Codes (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 [3]:

- Evolution proceeds by jumps: Error correction codes could explain this phenomena quite neatly by arguing that small number of errors (errors in close distance of current codewords) are corrected while the ones with larger distances are left uncorrected.
- The trend of evolution towards increased complexity: As information theory suggests, longer and more complex genomes could provide better error correction. This may suggest the role of natural selection in developing error correction mechanisms in DNA during evolutionary time

scales.

2) Soft Codes and Genetics: There are two ways of defining a code: specifying its construction rules (as communication engineers usually do) or specify the required constraints of the codewords. The second approach is more appropriate for natural phenomena. In some sense, soft codes are similar to define the parity check matrix instead of the generator matrix in coding theory.

To get a better understanding of soft codes, think of natural languages. In a natural language, there are a lot of constraints including the properties of the vocal tract (phonetics), constraints on the meaning of the actual words (lexicon), grammatical rules (syntax) and requiring meaningful sentences. At the same time, natural languages have a great capability of error correction, both in oral and written forms.

Furthermore, a language is defined by distinct constraints acting at several hierarchical levels. For instance, phonetic constraints, are more fundamental and rigid than constraints specific to a given language, which are mainly based on social conventions.

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* [3]. In a nested code, some parts of data are protected more than other parts. More specifically, 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.

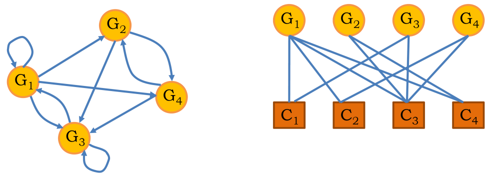
There are some evidence that the error correcting mechanism in genome, if any, should be a nested code. There exist some genes that are preserved among generations of far-related species. Many of such genes correspond to vital functions of species. For instance, the HOX genes which determine the organization plan of living beings are shared by humans and flies [3].

Therefore, the author suggest that there are error control mechanisms implemented in DNA. Moreover, he proposes the framework of nested soft codes to investigate these structures. While the author has provided a number of evidences in favor of the proposed hypothesis, whether such error control means exist in living beings is still not confirmed and certainly requires close collaboration of biologists and coding theorists.

B. Information Theory and Coding Problems in Genetics [14]

Although it is possible to have error correction mechanisms at nucleotide level, and although robustness of codons to errors was shown in [8], it is more likely to have error correction codes, if any, at genes network level. Because genetic ”code” is highly constrained by physicochemical rules. Using the metaphor of natural languages, error correction is implemented in the lexicon and upper levels, not in phonetics and alphabet levels.

Milenkovic et al. have focused on this issue in [14], by considering the applications of coding theory in Gene Regulatory Networks (GRNs). 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. If a mismatch occurs during the replication, a fast iterative decoding algorithm is used to correct the erroneously copied genes.



a. An example genetic graph

b. The corresponding factor graph

Fig. 1. Gene network and its corresponding gene factor graph

They show that there is a direct connection between certain models of GRNs, Boolean Network (BN) model in particular, and Low Density Parity Check (LDPC) codes. In fact, BNs contain LDPC codes as a special case. To review their work, we will first do a quick overview on basics of GRNs and iterative decoding techniques.

1) Gene Regulatory Networks: Genes are not independent of each other as the expression of certain genes affects the others. A gene is said to be expressed if its mRNA or protein can be found in the cell cytoplasm. Whenever a gene is expressed, it could affect the expression or suppressions of other genes by producing enzymes or proteins that bind to the regulatory sites of other genes. Hence, expression of one gene results in a series of reactions which could be modeled by GRNs.

To investigate GRNs, we consider the genetic graph in which we have N genes as the nodes. The effect of genes on each other is indicated by directed edges between nodes. An edge from node G_i to G_j means that G_i affect the state of G_j. The state of gene G_i is a function f_i of the set of genes that affect it. In other word, $f_i : \Pi_i \rightarrow \{0, 1\}$ where $\Pi_i = \{G_{i_1}, \dots, G_{i_k}\}$ is the set of genes affecting G_i. If $f_i = 1$, G_i is expressed and suppressed otherwise.

Usually, in a genetic graph, the in-degree of nodes is relatively small, i.e. the genetic graph is sparse. Moreover, there are occasionally long edges in this graph meaning that distant genes could depend on each other.

For the purpose of analysis,genetic graph could be transformed into a bipartite factor graph, in which left and nodes represent genes and control nodes for each gene, respectively. Genes connected to the control node C_j exhibit an influence on the state of gene G_j [14]. This transformation is shown in figure 1.

2) GRNs and Error Control Mechanisms: Consider a GRN with input state x and assume it converges to some stable state, y, if there are no perturbation, i.e. y is a point attractor of x. Now, we introduce a perturbation d which corresponds to errors in x. We would like to see how network evolves. In particular, we are interested in networks such that for small perturbations, the network still converges to the correct point attractor y after a certain number of steps.

In order to characterize the evolution of the GRN, we need to define the Boolean Jacobian F which is an $N \times N$ matrix with entry (i, j) equal to $\partial f_i / \partial x_j$, in which:

$$\frac{\partial f_i}{\partial x_j} = f(x_1, \ldots, x_j = 0, \ldots, x_N) \oplus f(x_1, \ldots, x_j = 1, \ldots, x_N)$$

In the above equation $\oplus$ indicates XOR operation. Having $F(t)$ and the perturbation at time t, $d(t)$, equation (1) shows how the perturbation evolves over time:

$$d(t+1) = F(t) \otimes d(t) \tag{1}$$

where all operations are in GF(2). Therefore, in order to have a vanishing perturbation we must have $d(t) = 0$ for some t, which means $d(0) \otimes F(1) \otimes \ldots \otimes F(t-1) = 0$.

Comparing the model for Gallager's LDPC codes and that of GRNs, their similarity becomes apparent immediately. The decoding graph of LDPC codes could easily be transformed to an equivalent genetic factor graph. In this graph, the control node for variable node i has as its input all variable nodes at distance exactly two from node i.

To see the application of the proposed model in understanding the proofreading mechanism of DNA, suppose an error occurs in decoding gene G_i and instead of the correct protein, which binds to the regulatory site of gene G_j, the newly produced protein binds to another gene's regulatory site, say G'_j, which may result in its expression. This corresponds to a perturbation in the network. Now if network possesses certain amount of error correction capability, it could overcome the perturbation in a few steps. This may be very helpful in preventing or even curing certain diseases. Therefore, 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.

Nevertheless, the authors have neither verified the proposed model experimentally nor provided sufficient biological findings in its favor. They only mention some biological evidence, which are not compelling in my opinion. These evidences include the finding that nucleases move in saltatory manner, suggesting that the genome is a globally connected structure and the error correction code could operate on the global structure [14]. Another one is the fact that malfunctioning of proof-reading mechanisms cause certain diseases such as cancer. Thus, verifying the suggested model and analyzing the performance of the error control mechanism from a coding theoretical point of view is an interesting subject for further research.

C. Neural Coding and Decoding: Communication Channels and Quantization [5]

The goal of this paper is to study neural coding from a new perspective by modeling the neural sensory systems as a communication channel. Based on this model, stimulus is encoded by spike trains. Furthermore, since the neural channel is noisy, the spike trains that could be confused with each other are grouped into equivalency classes. In this way, certain amount of noise reduction is achieved which translates into more robust codes.

Therefore, according to the model neural code has a structure similar to a dictionary: there are sets of stimulus-response pairs that are synonyms. Moreover, these sets are independent except for a few common members between each set. The rationale behind this model comes from the theory of typical sequences, which implies that stimuli belonging to a certain class results in a response in the same equivalency class with high probability. Note that this general approach contains some of currently used models for neural coding as special cases. For example in rate codes, all sequences having the same number of spikes in a given interval are members of the same equivalency class.

In order to identify the neural dictionary, information maximization techniques and quantization are used to reduce the size of data set. Among all possible quantization, the authors have considered the one which preserves as much information between stimulus-response pairs as possible. Quantization has the additional desirable property of helping us deal with the limited number of samples to fully characterize input-output probability distribution.

1) Model: The model considered in [5] is similar to models in other problems of information theory: we have an input with probability distribution $p(x)$. The input is transmitted over a channel with probability distribution of $p(y|x)$ which results in the output y with probability distribution of $p(y)$. In neural systems, input is the stimulus, channel is the neuron and output is the spike train.

From information and coding theory, we expect that the received sequence, Y[1], and the transmitted sequence, X, constitute a joint typical set. The jointly typical input-output pairs form equivalency classes. To identify the equivalent classes, we have to find the mutual information, I, and conditional entropy, H, between X and Y. A big challenge in estimating H and I is the large size of the bulk of data needed to closely approximate the probability distribution.

To overcome this issue, the quantized version of channel input and output are considered. Among all possible quantization schemes, the one which preserves as much information between stimulus and response as possible is selected. A stochastic quantizer is used to map the output sequence Y to the quantized version Y_N with N quantization levels according to the (to be determined) distribution $q(Y_N|Y)$. In this regard, rate code is equivalent to a deterministic quantizer. The quality of quantization is assessed with the following measure [5]:

$$D_I(Y; Y_N) = I(X; Y) - I(X; Y_N) \tag{2}$$

Now the problem is to minimize D_I, which is equivalent to the following problem [5]:

$$\max_{q(Y_N|Y)} H(Y_N|Y) \tag{3}$$

subject to

$$D_{eff} = I(X; Y_N) \geqslant I_0 \tag{4a}$$

$$\sum_{\nu=1}^{N} q(y_\nu|y) = 1, \ \forall\ y \tag{4b}$$

Where I_0 is a threshold determining the precision of the quantizer.

[1] Capital letters denote vectors.

This is a concave problem so the solution lies at the boundaries, i.e. $I(X;Y_N) = I_0$. Therefore, in order to find the solution of (3), we need to find the quantizer that carries at least I_0 bits of information about X. By pushing I_0 further and further, we reach a point, I_0^{max}, beyond which the above problem does not have a solution. This is the maximum possible information we can get for a fixed number of quantization levels, N.

Authors have used Lagrange multipliers to solve (3). In particular, Karush-Kuhn-Tucker (KKT) conditions are used to set the derivative of the Lagrangian objective function to zero and get the closed form of $q(Y_N|Y)$, which is given in equation (5).

$$q(y_N^{\nu}|y_k) = \frac{e^{-\beta \frac{\nabla D_I^{\nu k}}{p(y_k)}}}{\sum_{\nu=1}^{N} e^{-\beta \frac{\nabla D_I^{\nu k}}{p(y_k)}}} \tag{5}$$

where β is the Lagrangian multiplier corresponding to the constraint (4a), $q(y_N^{\nu}|y_k)$ gives the probability that y_k belongs to the equivalency class ν and $\nabla D_I^{\nu k} = \partial D_I / \partial q(y_N^{\nu}|y_k)$.

2) Deducing Codebook: To deduce the codebook, we fix an acceptable amount of distortion, D_I, which gives us N. Having N, we start building $q(y_N^{\nu}|y_k)$. Then, we can build a similar distribution $q(y_N^{\nu}|x)$ which is the same probability for input sequences. If D_I is small, then the responses associated with class y_N^{ν} are

$$\mathrm{y}^{\nu} = \{y_k | q(y_N^{\nu}|y_k) \simeq 1,\ \forall k\} \tag{6}$$

In this case, if we group all those x's whose output belongs to the same equivalency class, i.e.:

$$\mathrm{x}^{\nu} = \{x | q(y_N^{\nu}|x) \simeq 1\} \tag{7}$$

We end up having equivalency classes in x's. Therefore, we will have probability distribution $q(y_{|}^{\nu}x^{j})$, which is the probability that a member of the j^{th} equivalency class be mapped to a spike train from equivalency class ν. This gives us the codebook for the quantized model.

3) Results: The above method was applied to many sets of synthetic data, among which was the case where x and y are input and output of a Hamming(7,4) code. Then, noise is applied to *both* x and y to get x_0 and y_0. This is just to make a non-deterministic relationship between x and y which gives us something like equivalency classes as we already discussed. [2]

The results illustrated in figure 2 show that the proposed approach is able to identify equivalent classes which will be extremely helpful in analyzing neural codes. Note that a permutation of the rows and columns was applied to the generator matrix which changes the order of equivalent sequences of Hamming(7,4) code to be easier to understand.

Furthermore, in another paper, the authors have described the applications of the proposed method to real neural data [15]. The results show that the suggested approach is able to successfully identify stimulus-response equivalent classes with biologically meaningful relationships. Therefore, main contribution of this work is a framework to analyze the dependencies in spike trains. Such dependencies could provide a hint to existence of error correction mechanism at neuronal level. Extending this approach to neural population codes is an interesting topic for further research.

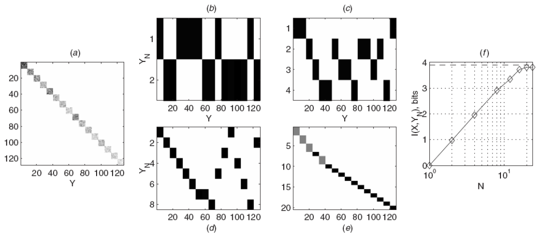


Fig. 2. The joint probability between X and Y after permuting rows and columns (a) and optimal quantization for different number of classes (b)(e). The behavior of the mutual information with increasing N can be seen in the loglinear plot (f). The dashed curve is $I(X;Y)$ [5].

III. Related Works on Genetics

A. Evidence in Favor of Existence Error Correction Mechanisms in Genome

The redundancy in genome and DNA is an accepted fact [2]. Moreover, redundancy is much higher in more evolved organisms. Although not proven yet, the addition redundancy may suggest existence of error correction codes, specially in more evolved organisms.

In addition to *qualitative* evidence we discussed in the previous section, some researchers tackled the problem of error control mechanisms in DNA *quantitatively*. MacDonaill has discovered a parity check code interpretation of nucleotide composition [12]. In this model, nucleotides form a 4-bit even parity code.

At a different level, there are several works suggesting that ”genetic code”, i.e. codon assignments to amino acids, are optimized in a way to admit error minimization during translation process. As an example, in [8] authors claim that natural selection has chosen amino acid to codon assignment such that translational errors are minimized. To prove their claim, many random ”codes” were generated by arbitrarily assigning codons to amino acids. Then, the authors evaluated the average resistance of each ”code” against random errors (mutations) among codons. 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.

May et al. [13] have modeled mRNA as a 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.

This approach was tested over the mRNA of E. Coli. Both the $(5, 2)$ and $(8, 2)$ models were able to distinguish coding and non-coding parts of DNA. They both also indicate the existence of key regions within the mRNA leader sequence.

[2] Therefore, this has nothing to do with evaluating the error correction properties of code. Instead, the goal is to find the structure of the neural code (the mapping between stimulus and response).

Hence, this approach is also suitable for finding coding and non-coding regions in DNA.

B. Arguments Against the Existence Error Correction Mechanisms in Genome

Although there is some evidence about the existence of error control methods in DNA, there are also several arguments against such mechanisms. First of all, neither of the works mentioned in the previous section is a proof of the existence of error correction in DNA. They all *suggest* that such mechanisms must exist.

Furthermore, some of the preliminary works on finding error control coding in DNA were unsuccessful [11], [17]. Both approaches considered block codes and proposed algorithms to search for signs of block-coding in DNA. However, as they have already mentioned, their approaches were too simplistic and limited. Neither of the mentioned approaches addressed the existence of convolutional codes. In fact, Liebovitch et al. [11] suggest that a more comprehensive examination would be required.

C. Applications of Coding in Gene Regulatory Networks

In addition to analyzing possible error corrections codes in DNA, other applications of coding theory in modeling different phenomena in genomics have been considered in recent years. A promising example of such applications is to use coding theory in order to model GRNs.

Recently, coding theory has been employed in refining current models of GRNs [6]. In the proposed method, the main interest is inferring the influence functions of a GRNs from biological findings. A coding-theoretic methods is advantageous here by being able to handle various issues including noise. In addition, even though it is assumed that the topology of the network is perfectly known, the proposed framework can be extended to deal with uncertainties in the network topology [6].

D. Source Coding and DNA

While our main focus in this report is on channel coding, there also exists some approaches based on source coding to model different phenomena in molecular biology. In [19], authors have investigated compression capabilities of DNA and its similarities to multiplexed codes. They show that the cardinality of the equivalent codon classes (the number of codons matched to each amino acid) is very close to what is predicted by multiplex codes. Furthermore, the expected length of the optimal multiplexed code for the ”genetic code” is shown to be only very slightly greater than that of the multiplexed codes [19].

IV. Related Works on Neuroscience

A. Information Theory and Neural Codes

Information theory has been used in neuroscience for a long time. Because of its close relationship with coding theory, we briefly review some of the applications of information theory in analyzing neural systems in the following.

In [4], a necessary and sufficient condition for the neuron’s firing profile is given so that the neural code is capacity achieving. This criterion has been tested via experiments on the visual system of the fly and it was verified that its firing pattern satisfies capacity achieving conditions [4]. This is an example of information maximization approach. A similar approach is used to show that population coding in cat’s visual system is also capacity achieving [4].

Another work on information maximization is due to [21] in which authors maximize mutual information between input synapses of a neuron and its outputs subject to the constraint that post-synaptic average firing rate stay as close as possible to its typical value. Following this approach, authors obtain the optimal updating rule for synaptic weights which is very similar to the experimentally verified Bienenstock-Cooper-Munro rule [21].

B. Error Correction in Neural Systems

In contrast to information theory, it is only a few years that coding theory has been employed in investigating neural systems. In this section, we discuss error correction in neural systems at two levels: the neuron’s level and the network level.

1) Neuronal Error Correction: The main job of sensory neurons is to encode stimuli efficiently and reliably. Such encoding in presence of noise was addressed in [7], where a group of neurons use block coding techniques to enhance error correction abilities of the neural code. To be more specific, a system in which data is encoded linearly via an encoding matrix, W, is considered. Encoded data is transmitted over a channel with additive noise and then decoded to give an estimate $\widehat{x}$ of input using a decoding matrix, A. The goal is to construct A such that the reconstruction error, $E[(\widehat{x}-x)^2]$ is minimized subject the power constraint in an AWGN channel. The authors have determined the best A and verified their prediction via simulations. However, they have not tested their work on real neurons to see if neurons employ the same approach.

2) Error Correction in Neural Networks: In conventional memories, like RAM, recall is based on address. However, in neural networks, recall is based on content [4]. In particular, there is a mechanism in brain, called associative memory, which is responsible for recalling different memorized patterns even if the input is noisy or contains errors (think of a blurry picture for example).

The pattern matching procedure which is performed by associative memory iteratively is very similar to decoding algorithms used in communication systems. Therefore, it is possible to use associative memory as an inspiration in designing better codes or employ coding-theoretical methods to better analyze such mechanisms.

An interesting example of the former case is the work of She and Cruz [20]. Inspired by the concept of associative memories, their goal is to design an artificial Bidirectional Associative Memory (BAM) which is capable of correcting the maximum number of errors. Although their work is focused on artificial neural network, their analysis could also be extended to neuronal networks.

BAM is a two-layer feedback neural network in which we have an input layer L_A with n neurons and an output layer L_B with m neurons. We have both forward and backward connections between these two layers but no intra-layer connections. Moreover, input and output of BAM are binary.

To overcome the effect of noise, we have to use weighted BAM in which correlation matrix is the weight sum of input patterns:

$$M = \sum_{i=1}^{N} w_i X_i^T Y_i, \tag{8}$$

where N is the number of training pairs and X_i and Y_i are bipolar representation of input and output, i.e. we have $+1/-1$ in X_i instead of $0/1$.

Now the goal is to find W such that BAM is able to tolerate largest amount of noise, i.e. be able to correct errors with Hamming distances as large as possible. The authors have specified the necessary conditions for W to accomplish this goal and show that if these conditions are satisfied, BAM is able to correct any errors within a Hamming distance less than a maximum determined by the training pairs. Moreover, there is always at least one pattern with a Hamming distance larger than the determined maximum that can not be recalled correctly.

The result were also experimentally tested on an artificial BAM to distinguish between three patterns. In the simulations, the maximum correctable error had a Hamming distance of 4 and all noisy input pairs within this bound were recalled correctly. Moreover, the proposed method was also compared to other approaches and it was shown that the suggested approach can find the maximum noise tolerance set, which is not guaranteed in other algorithms [20].

C. Source Coding and Neural Systems

In sensory parts of the nervous system, source coding plays an important role to encode stimuli efficiently and with high fidelity. Optimal source coding is achieved when the output of a group of neurons are independent. However, due to correlations in the stimulus, this not accomplished unless neurons cooperate. Using rate distortion theory and Kullback-Leibler divergence, it was respectively shown in [1] and [9] that population coding can perform as a better source coder compared to cases where neurons act independently.

V. Future Works and Research Plan

In this section, I will explain some of the research topics which seems worth pursuing. I will then mentioned the ones that I would like to work on personally.

A. Research Topics on Applications of Coding Theory in Molecular Biology

In molecular biology, one has to first show that error correction mechanisms exist indeed. Toward this end, any method for detection of coding structures in a set of given noisy sequences is of outmost importance, both in molecular biology and neuroscience. We explained some of these approaches [17], [13]. However, as discussed earlier, the mentioned techniques suffer from over-simplification. They only consider block codes and assume the genome to be composed of equal length codewords. Furthermore, mutation rate was assumed constant for different parts of the genome, an assumption known to be incorrect. Finally, coding and non-coding region were treated similarly while according to several suggestions, it is more likely to find coding structures in non-coding parts of genome and introns [3], [17]. Refining current models according to the mentioned directions is a topic for further research.

Another interesting approach is to focus on the set of genes that are shared by diverse species (such as HOX genes as already mentioned). By measuring the error (mutation) rate of such genes, one could verify the possibility of having nested codes in DNA, as suggested by Battail [3]. In addition, one can measure and compare the error rate in various species with different levels of evolution. The result could confirm or reject the conjecture due to Battail according to which error correction is a result of evolution and exists in higher species [3].

Nevertheless, searching for error correction codes in DNA is much more difficult than one might imagine. First of all, we do not know the encoding alphabet. The trivial assumption that the alphabet is composed of the four DNA nucleotides is not necessarily correct as the English alphabet is not composed of a set of vertical and horizontal lines that are used to build the actual letters. In fact, our previous work suggests that genome letters are composed of a combination of nucleotides [18]. Moreover, we even do not know which type of error correction methods are used in genome. So far, only block codes were considered. However, as suggested in [13], convolutional codes seem to be a better model in certain cases.

Investigation of error correction at GRN level is another subject which worths deeper explorations. Error correction mechanisms seems more likely at the network level when one notices the fact that nucleotides and codons are highly constrained by physicochemical rules. Using the language metaphor, it is unlikely to find error correction at phonetics or alphabets level. However, we have powerful error correction means at lexicon and upper levels.

As discussed earlier, Milenkovic et al. have employed iterative codes and codes on graph to model GRNs. Based on the proposed model, they conjecture that iterative error correction methods exist in GRNs. However, they do not provide enough evidence to prove their conjecture. Verification of the mentioned conjecture is certainly an interesting research topic. As a beginning step, coding-theoretical arguments could be used to specify necessary conditions for a feasible, fast and relatively accurate error control mechanism which provides some testable criterions for experimentalists. In addition, combination of evolutionary game theory with error correction mechanisms in GRNs may provide further evidence in favor of ECC in the realm of molecular biology.

Nevertheless, whether error correction mechanisms exist in DNA or not, coding-theoretical methods could also be employed in refining GRNs models. This is due to the similarity between the approaches used in modeling GRNs and codes on graph. We already mentioned an example in section III.

However, there are still a lot of work to be done towards this end.

On a separate topic, we have addressed phase transition in mutual information for a family of codes in communication systems [10]. The model used in our approach is similar to soft codes, in that we only specifies constraints on the generator matrix, and to NK-model, which is widely used to model GRNs [14]. Extension of this and similar approaches to address issues in molecular biology seems an interesting topic for future works.

B. Research Topics on Applications of Coding Theory in Neuroscience

Error correction codes may be present at two levels in the neural system: neuronal and network level. In the neuronal level, we must first gather enough evidence to show the existence of such methods. As mentioned before, any method that could extract coding structures from a given set of data is definitely helpful in this regard. We mentioned one such approach in section II. Although the method proposed in [5] is based on quantization and, therefore, closer to source coding than channel coding, the method could easily be extended for gathering evidence of channel coding, if any, in neurons. Because, in the end, the suggested approach identifies dependencies among the spike trains. Application of this method to actual neural data and its extension to neural population code seems a natural step in investigating coding properties of neurons.

In the network level, we have enough evidence on the existence of error correction mechanisms. Therefore, we may use coding-theoretical methods to analyze such mechanisms as well as inspiring from them to design better and faster decoding algorithms. For the analysis purpose, the method of [20] and similar other works on associative memory provides us with a good starting point. However, the method used in [20] to derive update rule for weights seems unappropriate for real neuronal networks as neurons can not use genetic algorithm to find the optimal weight. Coding theory may be able to come up with practical weight update algorithms.

Personally speaking, I would like to focus on applications of coding theory at network level, both in neural and gene regulatory networks. In my opinion, the models employed in these two areas are very similar to those used in coding theory to analyze codes on graphs and iterative coding algorithms.

I would also like to work on developing efficient algorithms for discovering coding structures among a given set of sequences. This approach would be absolutely helpful in proving/rejecting the existence of error correction mechanisms both in molecular biology and neuronal level of nervous system.

VI. Conclusion

In this report, we briefly discussed several applications of coding theory in biological systems. We considered two categories, namely, molecular biology and neural sciences. In the first case, we mentioned arguments both in favor of and against existence of error correction methods in DNA. The issue still remains an open question and requires further research. We also discussed other applications in which coding theoretical approaches are used to refine mathematical models of biological systems. For the case of neuroscience, we showed that there exists error correction means at the network level and there might be similar mechanisms in neurons as well. Coding theory may be helpful in analyzing such mechanisms as we already discussed.

In brief, although coding theory and biology seems two far apart fields, there are a lot of similar problems in both of them which requires closer collaboration of coding-theorists and biologists. Several such problems were outlined in the report which provide promising topics for further research.

References

[1] M. Aghagolzadeh, S. Eldawlatly, K. Oweiss, ”Coding stimulus information with cooperative neural populations”, Proc. ISIT, 2009, pp. 1594-1598.

[2] M. Akay, ”Genomics and proteomics engineering in medicine and biology”, Wiley-IEEE Press, 2007.

[3] G. Battail, ”Should Genetics Get an Information-Theoretic Education? ”, IEEE Eng. Med. and Bio. Mag., Vol. 25, No. 1, 2006, pp. 34-45.

[4] P. Dayan, L.F. Abbott, ”Theoretical neuroscience: computational and mathematical modeling of neural systems”, MIT Press, 2004.

[5] A. G. Dimitrov, J. P. Miller, ”Neural coding and decoding: communication channels and quantization”, Network: Comput. Neural Syst., Vol. 12, 2001, pp. 441472.

[6] J. Dingel, O. Milenkovic, ”A list-decoding approach for inferring the dynamics of gene regulatory networks”, Proc. ISIT 2008, pp. 2282-2286.

[7] E. Doi, D. C. Balcan, M. S. Lewicki, ”A theoretical analysis of robust coding over noisy over-complete channels”, Adv. in Neur. Inf. Proc. Sys., Vol. 18, 2006, pp. 307-314.

[8] S. J. Freeland, T. Wu, N. Keulmann, ”The case for an error minimizing standard genetic code”, J. Origins of Life and Evolution of Biospheres , Vol. 33, No. 4-5, 2003, pp. 457-477.

[9] D. H. Johnson, W. Ray, ”Optimal stimulus coding by neural populations using rate codes”, J. Comp. Neuroscience, Vol. 16, 2004, pp. 129138.

[10] K. R. Kumar, P. Pakzad, A. H. Salavati, M. A. Shokrollahi, ”Phase transition for mutual information”, To be submitted to IEEE Trans. Information Theory

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

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

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

[14] O. Milenkovic, B. Vasic, ”Information theory and coding problems in genetics”, Proc. Information Theory Workshop, 2004, pp. 24-29.

[15] A. E. Parker, A. G. Dimitrov, T. Gedeon, ”Symmetry breaking in soft clustering decoding of neural codes”, IEEE Tran. Inf. Th., Vol. 56, No. 2, 2010, pp. 901-927.

[16] F. Rieke, D. Warland, R. R. van Steveninck, W. Bialek, ”Spikes: exploring the neural code”, MIT Press, 1999.

[17] G. L. Rosen, ”Examining coding structure and redundancy in DNA”, IEEE Eng. Med. and Bio. Mag., Vol. 25, No. 1, 2006, pp. 62-68.

[18] A. H. Salavati, M. Nilchian, M. Alizadeh, S. Bagheri, M. Sadeghi, M. R. Aref, Genome alphabet: are the letters of genome language the four nucleotides or a combination of them?, To be submitted to the J. Theo. Bio.

[19] G. Sicot, R. Pyndiah, ”Study on the Genetic Code: Comparison with Multiplexed Codes”, Proc. IEEE International Symposium on Information Theory (ISIT), pp. 2666-2670, 2007.

[20] D. She, J. B. Cruz, ”Encoding strategy for maximum noise tolerance bidirectional associative memory”, IEEE Tran. Neur. Net., Vol. 16, No. 2, pp. 293-300.

[21] T. Toyoizumi, J. P. Pfister, K. Aihara, W. Gerstner, ”Generalized Bienenstock-Cooper-Munro rule for spiking neurons that maximizes information transmission”, Proc. Nat. Aca. Sci., 2005 , pp. 5239-5244.