Information Theoretical Approach in Quantification of DNA Storage Device Capacity

Jacob Emanuel JOSEPH (jamie@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr. Wilfred Ndifon
AIMS, South Africa

18 May 2017

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

The classical computational paradigm has difficulty dealing with the exponential growth of big data. There is an urgent need for inexpensive solutions to this big data problem. Deoxyribonucleic acid (DNA) offers a possible solution to this problem, especially because it has an extremely high data storage capacity that can last for thousands of years. Data scientists have been developing various experimental schemes for storing data in DNA. These schemes are not immune to errors, so it is of interest to understand the limits of their storage capacity. In the current work, we used information theory to estimate the theoretical limits of the storage capacity of two DNA storage schemes based on four- and six-letter nucleotide alphabets, respectively. We recover the previously reported limit of 1.83 bits/nt for a four-letter alphabet scheme and also report a new limit of 2.38 bits/nt for a six-letter alphabet scheme. This indicates that the storage capacity of the commonly employed scheme based on a four-letter alphabet is 23.11% less than one based on a six-letter alphabet, which has thus far been little studied.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Jacob Emanuel JOSEPH, 18 May 2017

Contents

1. Introduction

1.1 Background of Study

Data is very important to human kind because of its use in daily life. Humans have been using data for social and economic development. The demand for data has been increasing enormously and has created an emergent demand for massive data repositories (Yazdi et al., 2015). The increased volume of data, demands more capacity, scalability and efficient accessibility without an increase in resource demands.

There is an increasing demand in different organisations to use analytics applications to extract invisible information or those which would be impossible to derive using ordinary methods present. Likewise, different industries have been producing very large data sets for decades. This is a main reason behind the rise of big data. Today's big data does not include only the volume of the data sets, but also includes dealing with unstructured data. Therefore, the challenges is not only to come up with the storage architecture that is capable of storing this data set, but also that which will enhance the processing and analysis of the data set. Hence, technological advancement is inevitable (Computer-weekly).

Saxena et al. (2013) pointed out that, the current super computer uses the old technology which was used in the mechanically working dinosaurs in 1930s. This account for the challenge the modern electronic devices face in dealing with big data. There is a need for a new technology that will solve the problem of big data and increase the efficiency of digital devices. Current researchers are mainly focused on investigating the possibilities of using different technology to develop modern computer that is capable of handling the problems currently the digital world is facing. The new areas of computing, emerging in the field of quantum computing, optical computing and DNA(Deoxyribonucleic acid) computing, show promising results toward revolutions of entire digital media and computational models (Saxena et al., 2013). In this study our main focus is on the DNA computing technology.

DNA computing uses the information-processing capabilities of the DNA structure to counteract many drawbacks of the classical computers for example the low storage capacity. It is better to understand the new computing paradigm whose data structure and operations are different from the existing ones. The purpose of this new computational paradigm is to challenge the existing computational approaches and as a result, to bring the best solutions to the problems which the current computing paradigm is facing (Paun et al., 2005).

DNA computing is evolving with numerous range of applications in the digital world. It is employed to solve many combinatorial problems and offers a fundamental building block for building large scale nano-structures. DNA computing forms a basis of using the reversible logic gates in circuit synthesis. In addition to that, DNA computing gives high-density storage capacity (Ezziane, 2005; Tagore et al., 2010). In this work we concentrate on the later application.

The high-density storage capacity of DNA storage devices is very promising candidate to build upon an urgent solution of big data problem (Blawat et al., 2016). The DNA storage devices have unbelievable storage capacity compared to the ordinary storage devices such as CDs, DVDs, HDDs and Blu-rays discs. For example 1 gram of DNA holds 700 TB of data, this requires 233 pieces of 3 TB hard drives, weighing 151 kg in total (Saxena et al., 2013).Therefore, the classical computational paradigm requires an enormous expansion of the available resources to face the challenge of the observed exponential growth of data.

The DNA storage has been considered to be a potential medium for digital data storage since it has got many advantages compared to other forms of storage medium. DNA storage is extremely dense with the theoretical limit above 1 exabytes/mm^3. It has longer term storage solution and is often readable despite the degradation in harsh environments (Church et al., 2012; Bornholt et al., 2016). Moreover, DNA storage devices, consume significantly less energy compared to other electronic storage devices (De Silva and Ganegoda, 2016).

Despite the fact that DNA storage promises to provide better solution in the big data problem, there are still several challenges to overcome. None of the available encoding models is not associated with errors. Most of these models are cost infeasible and take a significant amount of time to read and write data onto DNA. Other developed model are completely inaccurate and not distinctively decodable (De Silva and Ganegoda, 2016; Bornholt et al., 2016).

Existing work aim at achieving the maximum net information density as possible has focused on controlling redundancy in their schemes to come out with best results. Church et al. (2012) used the binary scheme which has no error correcting code or fold redundancy[1] encoded 659 megabytes(MB) and achieve the net information density of 0.83 bits/nt. Goldman et al. (2013) used a ternary scheme that has one parity nucleotide for error detection and two nucleotides to detect the reverse complement and a four redundancy, has achieved to encode 757 MB and achieve a net information density of 0.34 bits/nt. Blawat et al. (2016) used a scheme that employs Reed Solomon code, encoded 22 MB with net information density of 0.92 bits/nt (Erlich and Zielinski, 2017).

The experimental schemes are characterised by high cost, inaccuracy, data breakage during encoding, need of PCR (Polymerase chain reaction) and take significant amount of time to read and write data onto DNA (De Silva and Ganegoda, 2016). These challenges make the process of developing the DNA storage systems harder with small chance of success. Due to the many advantages that DNA storage systems offer to the digital universe, there is a great need to come out with an approach that will serve as the alternative solution to the experimental schemes and offer a good foundation of the experimental schemes that will reduce the cost and increase their efficient. By the same reasoning, we formulate DNA storage as an information transmission channel and use mutual information theory to quantify the storage capacities of two DNA storage schemes, one based on a four-letter nucleotide alphabet and another based on a six-letter alphabet.

1.2 Organisation of the Essay

We discuss the rise of the big data and the emergent of DNA storage system in Chapter 1. We show the need of a new sophisticated architecture for data storage which does not require an expansion of resources. We describe the challenges of this architecture and the possible solutions have been tried thus far. We discuss the information theory in Chapter 2. We describe into details the theory behind the communication channel capacity. We make use the tools discussed in Chapter 2 in DNA storage device architectures in Chapter 3. We calculate the net information density of different schemes of DNA storage devices using the theoretical approach we have discussed and compare the results with other works. We then conclude our work in Chapter 4.

[1] Fold redundancy is defined as the number of reads which contribute to each consensus base e.g. a 4 fold redundancy indicate the that an average of 4 sequencing reads spanned each base (Bouck et al., 1998).

2. Theory

In this chapter we reviewed the present theories used to describe a communication channel. We discuss various aspects of entropy and mutual information in the communication channel. We define the entropy and discuss its properties. We then describe the key concepts of the communication channel by defining the channel capacity. We describe the quantification of the channel capacity of the channel using the mutual information. We describe the properties of the channel capacity and discuss some examples of channel capacity. We then discuss the DNA storage device architecture and show its basic structure and then develop a mathematical formulation of finding the net information density of the channel.

2.1 Information Theory

Several communication channels are controlled by the signals they receive from their inputs and/or the feedback from their outputs. The signals carry information about the sources of the communication channel. Thus, we can understand the behaviour of the communication channel by understanding the mutual relationship between the inputs and outputs of the channel through the information that they carry and convey. (Godfrey-Smith and Sterelny, 2016). To make inference about the information conveyed by these input signals, requires a thorough examination of the output signals(Lopes et al., 2011).

In 1948, Claude Shannon published a mathematical theory of communication currently known as information theory(IT)(Mousavian et al., 2016). According to Shannon, information is regarded as knowledge that distinguishes a particular state of the system(signal or input) from other many available potential states. For example, among the genes expressed in a particular disease, the knowledge of which sets of genes are up-regulated and which sets of genes are down-regulated or not altered at all, can be regarded as information. In Shannon's information theory, the system that is capable of taking multiple states is considered to be a source of information. The state of the system is presented mathematically as a random variable(Rhee et al., 2012). The random variable is here defined as a mathematical object that takes on a finite number of different states of the system with specific probabilities(Mousavian et al., 2016). The random variable can gain, reduce or retain the amount of information it carries. This amount of information is quantified by the Shannon entropy. The communication channel in which information is transmitted consists of input and output random variables which depend on each other. Therefore, we can resolve the input value by measurement of the output value and the amount of information gained is then quantified using the mutual information(Rhee et al., 2012).

In the following sections, we discuss the concept of Shannon's entropy, mutual information and channel capacity. We later apply these concepts in different schemes of DNA storage devices.

2.2 The Shannon Entropy

Entropy can be defined as a measure of uncertainty about the state of the object in the system. It quantifies the unpredictability of the value of a random variable(Adami, 2012; Rhee et al., 2012). In the following section, we define the entropy, with examples, of the discrete random variable and discuss its properties.

2.2.1 Definition. A discrete random variable X that take values from $x_1, ..., x_n$ with probabilities $p(x_1), ..., p(x_n)$ have a measure of uncertainty(entropy) given by

$$H(X) = \sum_{i=1}^{N} p(x_i)\text{log}_2 \frac{1}{p(x_i)}. \tag{2.2.1}$$

To measure the entropy in bits, we use the normal convention in equation 2.2.1 by defining the entropy in base 2 logarithm. Entropy is non-negative value. We provide the proof of this property here below;

2.2.2 Lemma.

$$H(X) \geq 0$$

Proof. Since $p(x_i)$ is the probability that a discrete random variable X takes a value x_i, then

$$0 \leq p(x_i) \leq 1 \text{ for all } i$$

This implies that

$$p(x_i)\text{log}_2\, p(x_i) \leq 0$$

This implies that

$$-\sum_{i=1}^{N} p(x_i)\text{log}_2\, p(x_i) \geq 0$$

$$\therefore H(X) \geq 0$$

□

Let us consider some few examples in system biology that describes the properties of entropy. The first example, we consider the simplest case in which a particular gene (a random variable X) is either turned on (state x_1) or off(state x_2) with equal probabilities $(p(x_1) = p(x_2) = \frac{1}{2})$ to transcribe a certain kind of protein. Deducing the entropy we have

$$H(X_{on,off}) = \frac{1}{2}\text{log}_2 4 = 1\text{bit}$$

Another example is the Jukes–Cantor substitution model, which assumes the four bases Adenine(A), Cytosine(C), Thymine(T) and Guanine(G) of DNA are substituted with the same probabilities $(p_A = p_C = p_T = p_G = \frac{1}{4})$. In this particular case we calculate the entropy as follows

$$H(X_{A,C,T,G}) = 4(\frac{1}{4}\text{log}_2 4) = 2 \text{ bits}$$

The last example is the case in which the four bases A, C, T and G have different probabilities i.e $p_A \neq p_C \neq p_T \neq p_G$ a practical example being a *Saccharomyces cerevisiae*(Yeast) with the following probabilities

$$p_A = 0.3090$$
$$p_T = 0.3080$$
$$p_G = 0.1917$$
$$p_C = 0.1913$$

The entropy becomes

$$H(X_{A,C,T,G}) = 1.96 \text{ bits}$$

The first example involves only two states and the entropy is found to be 1 bit whereas in the second example and third example increase in states has increased the entropy. In equal probable states such as in example 2 the entropy is high compared to unequal probable states such as in example 3. We develop a mathematical proof for this statement here below.

2.2.3 Proposition. Suppose X takes on N values. The entropy $H(X)$ is maximized when X is uniformly distributed on values of N

We use the Lagrange multipliers method to prove that entropy is maximum when X is uniformly distributed.

Proof. Recall equation 2.2.1

$$H(X) = \sum_{i=1}^{N} p(x_i)\text{log}_2 \frac{1}{p(x_i)}. \tag{2.2.2}$$

$H(X)$ is maximized under the constraint

$$\sum_{i=1}^{N} p(x_i) = 1. \tag{2.2.3}$$

Introducing the Lagrange multiplier in the equation 2.2.3 we get

$$H^*(X) = \sum_{i=1}^{N} p(x_i)\text{log}_2 \frac{1}{p(x_i)} + \lambda(\sum_{i=1}^{N} p(x_i) - 1). \tag{2.2.4}$$

We then find the derivative of $H^*(X)$ with respect to $p(x_i)$

$$\begin{aligned} \frac{\partial}{\partial p(x_i)} H^*(X) &= \frac{\partial}{\partial p(x_i)} \Big(\sum_{i=1}^{N} p(x_i)\text{log}_2 \frac{1}{p(x_i)} + \lambda(\sum_{i=1}^{N} p(x_i) - 1) \Big), \\ \frac{\partial}{\partial p(x_i)} H^*(X) &= \Big(-\text{log}_2 p(x_i) - \frac{1}{\text{ln}2} + \lambda \Big). \end{aligned} \tag{2.2.5}$$

We set $\frac{\partial}{\partial p(x_i)} H^*(X) = 0$ for maximization

$$\Big(-\frac{\text{ln } p(x_i)}{\text{ln } 2} - \frac{1}{\text{ln } 2} + \lambda \Big) = 0. \tag{2.2.6}$$

We solve for $p(x_i)$ in the equation 2.2.6 above we get

$$p(x_i) = \text{exp}\Big(\lambda \text{ln } 2 - 1\Big). \tag{2.2.7}$$

We then substitute $p(x_i)$ into the equation 2.2.3 as follows

$$\begin{aligned} \sum_{i=1}^{N} \text{exp}\Big(\lambda \text{ln } 2 - 1\Big) &= 1, \\ N\text{exp}\Big(\lambda \text{ln } 2 - 1\Big) &= 1, \\ \text{exp}\Big(\lambda \text{ln } 2 - 1\Big) &= \frac{1}{N}. \end{aligned} \tag{2.2.8}$$

Therefore,

$$p^*(x_i) = \frac{1}{N} \tag{2.2.9}$$

□

Thus, the entropy $H(X)$ is maximum when X is uniformly distributed.

2.2.4 Remark. The upper bound of $H(X) = \log_2 N$. i.e.

$$\max H(X) = \log_2 N = \log_2 |\mathcal{X}| \tag{2.2.10}$$

In general, the entropy is a *"simple function of the number of possible states and the probabilities"*(Rhee et al., 2012) as we have shown using the three examples above. We now define the concept of joint entropy and conditional entropy as follows;

2.2.5 Definition. A pair of discrete random variables with a joint distribution $p(x, y)$ has the joint entropy $H(X, Y)$ defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y). \tag{2.2.11}$$

2.2.6 Definition. The conditional entropy $H(Y|X)$ can be defined as

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X = x). \tag{2.2.12}$$

But,

$$H(Y|X = x) = -\sum_{y\in\mathcal{Y}} p(Y|X = x)\log_2 p(Y|X = x).$$

Therefore, the equation 2.2.12 can be written as

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|x). \tag{2.2.13}$$

2.2.7 Theorem.

$$H(X, Y) = H(X) + H(Y|X). \tag{2.2.14}$$

Proof. Recalling equation 2.2.11 we have

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y).$$

We can express the above equation as

$$\begin{aligned}
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x)p(y|x),\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|x)
\end{aligned}$$

$\sum_{y\in\mathcal{Y}} p(x, y)$ gives the marginal distribution of x and therefore,

$$\sum_{y\in\mathcal{Y}} p(x, y) = p(x).$$

Hence,

$$= -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(y|x). \tag{2.2.15}$$

We then substitute the equation 2.2.1 and 2.2.13 into 2.2.15 we get

$$H(X, Y) = H(X) + H(Y|X).$$

□

Thus;

$$H(X,Y) = H(X) + H(Y|X). \tag{2.2.16}$$

2.3 Mutual Information

Mutual information is defined as the amount of information that can be obtained about a random variable X by observing another random variable Y. This implies that the information that Y provides about X reduces uncertainty about X(Farhangmehr et al., 2014). Mutual information measures the dependence between the variables X and Y(Zhang et al., 2012). It is always a non-negative value and is equal to zero when the two variable are completely independent from each other.

2.3.1 Definition. Given two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X; Y)$ is defined as

$$I(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\text{log}_2\frac{p(x,y)}{p(x)p(y).} \tag{2.3.1}$$

Considering the equation 2.3.1 above we have

$$I(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\text{log}_2\frac{p(x,y)}{p(x)p(y)}.$$

The above equation can be simplified to

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\text{log}_2\frac{p(x|y)}{p(x)},$$

which is further simplified to

$$\begin{aligned} &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\text{log}_2\, p(x) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\text{log}_2\, p(x|y), \\ &= -\sum_{x\in\mathcal{X}} p(x)\text{log}_2\, p(x) - \Big(-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\text{log}_2\, p(x|y)\Big). \end{aligned}$$

We then use the equations 2.2.1 and 2.2.13 in the above equation to get

$$= H(X) - H(X|Y)$$

$$\therefore I(X;Y) = H(X) - H(X|Y). \tag{2.3.2}$$

By symmetry, we can write the equation 2.3.2 as follows,

$$I(X;Y) = H(Y) - H(Y|X). \tag{2.3.3}$$

Also, the self mutual information can defined as

$$I(X;X) = H(X) - H(X|X) = H(X). \tag{2.3.4}$$

We recall equation 2.2.16 we have

$$H(Y|X) = H(X,Y) - H(X). \tag{2.3.5}$$

Substituting equation 2.3.5 in 2.3.3 yields

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{2.3.6}$$

2.4 Channel Capacity

In a given communication channel the measure of how much information can be transmitted through the channel under given constraints is called **channel capacity**. Suppose a discrete channel of communication has an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. In this channel, the probability of observing the output y given the input x is defined by $p(y|x)$. If the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs, then the channel is said to be **memoryless discrete channel** (Cover and Thomas, 2012).

2.4.1 Definition. Information channel capacity is the maximum mutual information taken over all possible input distributions $p(x)$

$$C = \max_{p(x)} I(X;Y). \tag{2.4.1}$$

2.4.2 Properties of Channel Capacity. Channel capacity has got various properties as discussed here below (Cover and Thomas, 2012);

1. $C \geq 0$ since $I(X;Y) \geq 0$
2. $C \leq \log_2 |\mathcal{X}|$ since $C = \max I(X;Y) \leq \max H(X) = \log_2 |\mathcal{X}|$
3. $C \leq \log_2 |\mathcal{Y}|$ since $C = \max I(X;Y) \leq \max H(Y) = \log_2 |\mathcal{Y}|$

Channel capacity exists in various forms. Examples of which are;

1. noiseless binary channel,
2. noisy channel with non-overlapping outputs,
3. binary symmetric channel,
4. binary erasure channel.

We discuss briefly the above examples of channel capacity in the following section.

2.4.3 Noiseless Binary Channel. Noiseless binary channel is the one whose binary input is reproduced exactly at the output. Example if the input is 1 or 0 the output is also 1 or 0 respectively. Therefore, any transmitted bit is received without error.

2.4.4 Noisy Channel with Non-overlapping Outputs. Noisy channel with non-overlapping outputs is the one with two possible outputs corresponding to each of the two inputs. For example, in a channel with input alphabet $\mathcal{X}$ which takes two values $\{0,2\}$ and the output alphabet $\mathcal{Y}$ that is either the input value (e.g. 0) recovered without error or the input value plus one (e.g. $0+1=1$) due to error(noisy) in the transmission channel, will contain the input values $\{0,2\}$ and the output $\{0,1,2,3\}$.

2.4.5 Binary Symmetric Channel. This is the kind of channel whose outputs are complemented with small probability p. Example, a channel in which, when error occurs, the input 0 changes to output 1 and the input 1 changes to output 0.

2.4.6 Binary Erasure Channel. Is a kind of channel in which the bits are lost rather than being changed as in case of the binary symmetric channel. Example, a channel in which some fraction of data α is being lost in the output.

Using one or more types of the information channels we have discussed in this section various DNA storage architectures can be constructed. In the following section we discuss into details the fundamental structure of the DNA storage device and show its construction using the types of channel we have discussed so far.

2.5 DNA Storage Device Architecture

2.5.1 Fundamental Structure of DNA. DNA is a molecule that carries the genetic instructions necessary for growth, development, functioning and reproduction of all living organisms(Wikipedia, a). DNA consists of four nitrogenous bases that are basically being divided into two main groups the **purines** and the **pyrimidines**. The purines consist of the Guanine (**G**) and Adenine (**A**) whereas the pyrimidines consist of Thymine (**T**) and Cytosine (**C**). These bases pair up with each other to form units called base pairs. A pairs with T and G pairs with C. The bases are also attached to sugar and phosphate molecules. The base pair, sugar molecule and the phosphate molecule make a **nucleotide**. Nucleotides form two long strands like spiral, that form a double helix structure(Genetic Home Reference). Figure 2.1 shows the fundamental structure of DNA together with the base pairing mechanism.

A short sequence of nucleotide usually 10 to 30 nucleotides forms an **oligonucleotide (oligo)**. Oligonucleotides are commonly made in the laboratory by solid-phase chemical synthesis and they are vital for artificial genes synthesis, polymerase chain reaction (PCR) and DNA sequencing. Oligonucleotides have a wide range of applications in genetic testing, research, forensics and DNA storage devices(Wikipedia, b).

2.5.2 DNA Storage Device. The base-pairing mechanism gives DNA the ability to carry information by means of the linear sequence of its nucleotides. Each nucleotide is said to write a biological message in linear form (Saxena et al., 2013). The bases in the nucleotide can be converted in digital code of 1's and 0's and make a storage device that can be used to store data that can be read by modern computer. The DNA storage system is made up of a DNA synthesizer, a storage container and a DNA sequencer as shown in the Figure 2.2. The DNA synthesizer is used to encode data to be stored in DNA. The storage container has compartments in which pools of DNA that map to a volume are stored and the DNA sequencer reads the DNA sequences and makes a conversion into a digital data (Bornholt et al., 2016).

Basically the DNA storage device is can be regarded as a communication channel that transmit information by synthesis of DNA oligos and receive the transmitted information by sequencing the oligos and decoding the sequencing data. Erlich and Zielinski (2017) describes an example of the DNA storage device that behaves as a constrained channel concatenated to an erasure channel as shown in the Figure 2.3. Number of experimental factors including the DNA synthesis imperfection, degradation of DNA molecule over time, stutter noise[2] and PCR dropout make the channel prone to errors and noisy (Erlich and Zielinski, 2017). Currently, various works are focused on how to achieve an error free DNA storage device that can store as maximum information as possible.

2.5.3 Net Information Density of DNA Storage Device . The amount of information that can be recovered(transmitted without error) per nucleotide is called net information density. It is also called the information capacity per nucleotide. In the DNA storage devices, it is crucial to quantify the net information density of the device so as we can tell the capabilities of our communication channel (Erlich and Zielinski, 2017). The net information density is calculated using the following formula

$$C_{nt} = C/l \tag{2.5.1}$$

[2] Stutter noise is the kind of error that arise during PCR amplification and lead to addition or deletion of copies of the repeated unit in observed sequencing reads (Gymrek, 2016)

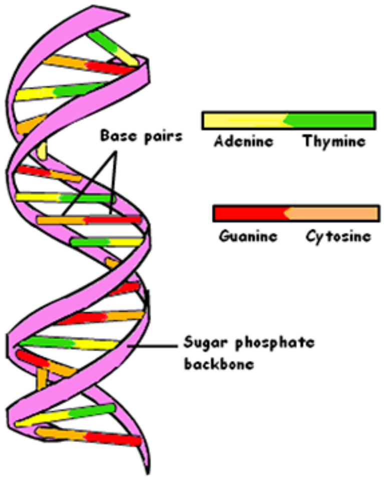

Figure 2.1: Fundamental structure of DNA (Tutorpace).

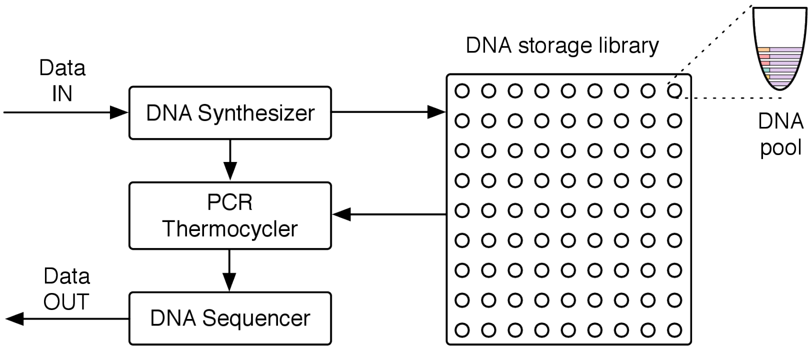

Figure 2.2: Components of DNA storage system (Bornholt et al., 2016).

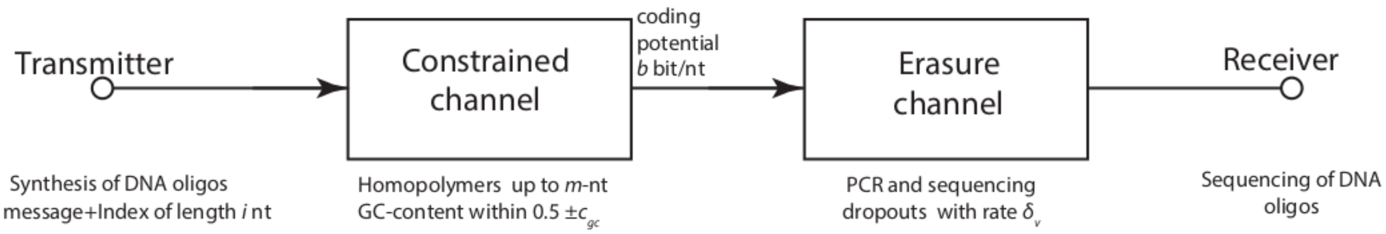
Transmitter
Constrained channel
coding potential b bit/nt
Erasure channel
Receiver
Synthesis of DNA oligos message+Index of length i nt
Homopolymers up to m-nt GC-content within 0.5 ±c_gc
PCR and sequencing dropouts with rate δ_v
Sequencing of DNA oligos

Figure 2.3: The DNA storage device transmission channel(Erlich and Zielinski, 2017).

Whereby

$$\begin{aligned} C_{nt} &= \text{Net information density} \\ C &= \text{Channel capacity} \\ l &= \text{length of the nucleotide} \end{aligned} \tag{2.5.2}$$

We will use the equation 2.5.2 to calculate the the net information density of various channel of DNA storage device in the next chapter. The inputs and the outputs of the channel of the DNA storage device are discrete entities. Therefore, in all of this work our estimation of the value of C_{nt} will base on discrete random variables of the inputs and outputs.

3. Results and Discussion

We use the developed tools in Chapter 2 to quantify the amount of information that can be stored in DNA storage devices. Church et al. (2012) achieved an error free channel using high throughput sequencing methods. On other hand, other works(Goldman et al., 2013; Grass et al., 2015; Erlich and Zielinski, 2017) have used different schemes and their results differ according to the schemes they have used and the constraints under consideration. Our objective is to estimate theoretically the amount of information that can be transmitted in the DNA storage channel with minimum error. We apply the mutual information theory in two kind of DNA storage device i.e. one with four bases discovered by Francis Crick in the year 1954 and the synthesized six bases DNA. In both cases, we consider the DNA storage device as a communication channel described in the previous chapter(Chapter 2).

3.1 Theoretical Capacity Quantification of DNA Storage Device

A theoretical quantification of the capacity of DNA storage device can be done by analysing the mutual information of the channel input and output. In this section we apply this approach in two types of DNA storage device i.e. the one with four bases and the one with six bases separately.

3.1.1 Four Alphabet DNA Device. The DNA four letters storage device uses the four bases of the DNA molecule to store its information.The oligonucleotides are used to encode and decode data and hence makes a storage device. Encoding and decoding processes make a communication channel similar to the one described in chapter 2. This communication channel is prone to random errors and consequently the channel information capacity is reduced. We use the information theoretical approach to estimate the amount of information that can be transmitted without/ with minimum error. We then compare our theoretical results with other related works.

Coding Capacity of Error Free Channel. We consider the channel whose input $\mathcal{X}$ and output $\mathcal{Y}$ can take **(A,T,G,C)**. In this particular case A and G are encoded as A which is expressed in binary as 00 while T and C are encoded as T which is expressed in binary as 01. The channel input $\mathcal{X}$ transmit A and T and since it is error free channel, A and T are received perfectly in the output $\mathcal{Y}$ as shown in the Figure 3.1

Let

$$p(X = 0) = \frac{1}{2}$$
$$p(X = 1) = \frac{1}{2}$$

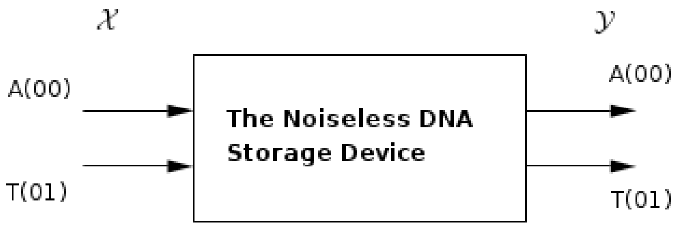

Figure 3.1: Error free (noiseless) 4 alphabet DNA communication channel.

$\mathcal{Y}$ \ $\mathcal{X}$	A	T
A	1	0
T	0	1

Table 3.1: Joint probability distribution

We calculate the entropy $H(X)$ using the equation 2.2.15

$$H(\frac{1}{2},\frac{1}{2}) = 1\text{bit}$$

The joint distribution for this channel is shown in the Table 3.1. From this distribution we calculate $H(X|Y)$ as follows

$$\begin{aligned} H(X|Y) &= \sum_{i=A,T} p(Y=i)H(X|Y=i), \\ &= p(Y=A)H(X|Y=A) + p(Y=T)H(X|Y=T), \\ &= -\text{log}_2 1 - 0\text{log}_2 0 - 0\text{log}_2 0 - \text{log}_2 1, \\ &= 0 \text{ bit}, \\ \therefore H(X|Y) = 0 \text{ bit}. \end{aligned}$$

$$C = \max_{p(x)} I(X;Y).$$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y), \\ &= 1\text{bit}. \end{aligned}$$

Thus,

$$C = 1\text{bit}$$

We can obtain similar result as above using equation 2.2.10

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} H(X) \\ &= \text{log}|\text{A}_X| \\ &= \text{log } 2 \\ &= 1\text{bit} \end{aligned}$$

Net Information Density of Error Free Channel. We let δ_v be the probability that the valid sequence is reduced during the encoding process. We suppose that, the encoded sequence is valid and is perfectly received with the probability of $1-\delta_v$ hence the $H(X|Y) = \delta_v H(X)$(Erlich and Zielinski, 2017). Recall

$$\begin{aligned} C_{nt} &= C/l, \\ &= \frac{H(X) - H(X|Y)}{l}, \\ &= \frac{H(X) - \delta_v H(X)}{l}, \\ &= (1-\delta_v)\frac{\text{log}_2|\text{A}_X|}{l}. \end{aligned}$$

But,

$$b = \frac{\log_2 |\mathrm{A}_X|}{l}.$$

Therefore, when a valid sequence is transmitted perfectly without considering the dropouts rate[3] the net information density becomes;

$$C_{nt} = (1 - \delta_v) b. \tag{3.1.1}$$

We choose a realistic value(Erlich and Zielinski, 2017) of $\delta_v = 0.005$ and solve for equation 3.1.1 we get

$$C_{nt} = (1 - 0.005) = 0.995 \text{ bits/nucleotide.}$$

Thus, the net information density of error free channel is approximately equal to 1 bits/nucleotide. Church and others (Church et al., 2012), experimented with error free channel using the coding capacity $b = 1$ bits/nucleotide were able to encode 659 kilobytes using oligos of length 115nt with the index length of 19bits and found out that the net information density $= 0.83$ bits/nucleotide. In their work, Church and others(Church et al., 2012), did not mention the value of δ_v they used even though their net information density is also approximately equal to 1 bits/nucleotide similar to what we have obtained in our theoretical approximation.

Coding Capacity of Non- Error Free Channel. In non-error free channel we consider a different scheme that will maximize the net information density of the channel. Consider the channel whose input $\mathcal{X}$ and output $\mathcal{Y}$ can take **(A,T,G,C)**. We restrict ourself to the mutation that occurs only within the same types of the base pairs and not otherwise. In other words, we neglect the mutations that may cause transition from one type of base to another by assuming that they have small chance to occur. To maximize the coding capacity we need to maximize $H(X)$ and minimize $H(X|Y)$. $H(X)$ is maximized with uniform distribution of $p(x_i)$ as proved in section 2.2 of chapter 2 above. We let

$$p(X = A) = p(X = T) = p(X = G) = p(X = C) = \frac{1}{4}$$

We compute the entropy using equation 2.2.1 and we obtain

$$\begin{aligned} H(X) &= \sum_{i=A,T,G,C} p(x_i) \log_2 \frac{1}{p(x_i)} \\ H(X) &= 2 \text{ bits} \end{aligned}$$

We generate a joint probability distribution that will minimize as much as possible the value of $H(X|Y)$ but it should not be equal zero because in practise there is a depence of $\mathcal{X}$ and $\mathcal{Y}$ and due to mutation, there will be some changes of bases in the output. In this case we have employed the methods used by Rivas (2005) to generate the joint probability distribution in table 3.2.

We calculate $H(X|Y)$ as follows

$$\begin{aligned} H(X|Y) = &\sum_{i=A,T,G,C} p(Y = i) H(X|Y = i) \\ &= p(Y = A)H(X|Y = A) + p(Y = T)H(X|Y = T) + p(Y = G)H(X|Y = G) \\ &+ p(Y = C)H(X|Y = C) \\ &= 0.507(0.07988) + 0.159(0.20160) + 0.184(0.15110) + 0.150(0.37824) \\ &= 0.156\text{bits} \end{aligned}$$

[3] Dropout rate is the probability of the oligo dropouts during decoding process (Erlich and Zielinski, 2017).

$\mathcal{X}$

$\mathcal{Y}$	A	T	G	C
A	0.502	0.005	0.000	0.000
T	0.005	0.154	0.000	0.000
G	0.000	0.000	0.180	0.004
C	0.000	0.000	0.011	0.139

Table 3.2: Joint probability distribution

From

$$C = \max_{p(x)} I(X;Y)$$

Substituting 2.3.2 in the above equation we have

$$C = (2 - 0.156)\text{bits} = 1.844\text{bits}.$$

Net Information Density of Non-Error Free Channel. We use equation 3.1.1 to calculate the net information density with $\delta_v = 0.005$. But since our channel is non-error free the capacity per nucleotide is reduced directly proportional to the dropout rate. In most storage architectures the value of δ has been found to be 0.5% (Erlich and Zielinski, 2017). Hence the net information density will be

$$\begin{aligned} C_{nt} &= (1 - 0.005)^2 1.844 \text{ bits per nucleotide} \\ &= 1.83 \text{ bits per nucleotide.} \end{aligned}$$

Therefore the net information density, $C_{nt} = 1.83$ bits per nucleotide. We will compare these results with other related work in the next section.

To achieve this results, we have used a reasonable joint probability distribution that maximizes $H(X)$ and minimize as much as possible the $H(X|Y)$. One can obtain different results if he is using another probability distribution. Our results are reasonable due to the fact that $H(X)$ is maximize if and only if the distribution is uniform as we have proved in Chapter 2, and the observed transition matrices used in most model give A a higher frequency compared to other bases which result to the similar pattern as the one we have obtained in the joint probability distribution as shown in the Table 3.2.

3.1.2 Six Alphabet DNA Device. In the six alphabet DNA device there are two more bases which makes the nucleotide to look like **(A,T,G,C,$\mathscr{X}$,$\mathscr{Y}$)**. We suppose an input $\mathcal{X}$ and $\mathcal{Y}$ take **(A,T,G,C,$\mathscr{X}$,$\mathscr{Y}$)**. The goal is to achieve the minimal error or error free channel and to calculate the maximum amount of information which can be transmitted.

Coding Capacity of Error Free Channel. The encoding process maps A and G to A which can be expressed in binary as 000, C and T to T expressed as 001 in binary and $\mathscr{X}$ and $\mathscr{Y}$ to $\mathscr{X}$ which is expressed as 010 in binary. Therefore we construct an error free binary channel as shown in the Figure whose capacity is calculated as shown below.

$$\begin{aligned} p(X = A) &= \frac{1}{3} \\ p(X = T) &= \frac{1}{3} \\ p(X = \mathscr{X}) &= \frac{1}{3}. \end{aligned}$$

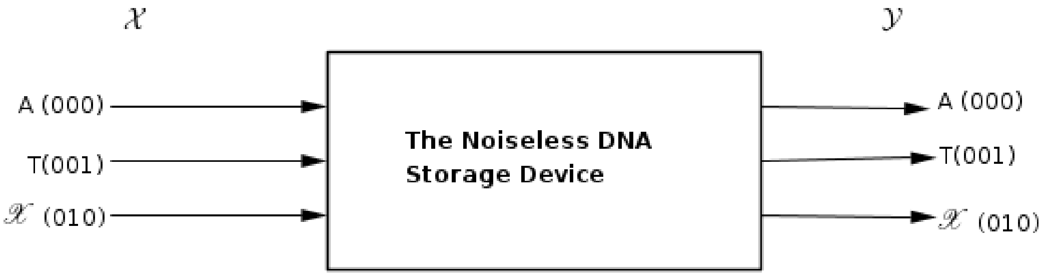

Figure 3.2: Error free (noiseless) 6- alphabet DNA communication channel.

$\mathcal{X}$

		A	T	$\mathscr{X}$
$\mathcal{Y}$	A	1	0	0
	T	0	1	0
	$\mathscr{X}$	0	0	1

Table 3.3: Joint probability distribution for noiseless 6 alphabet DNA device.

We caclute the entropy $H(X)$ using the equation 2.2.1 we get

$$H(\frac{1}{3},\frac{1}{3},\frac{1}{3}) = 1.58\text{bits}.$$

We calculate the conditional entropy $H(X|Y)$ using the joint probability distribution shown in the Table 3.3.

$$\begin{aligned} H(X|Y) &= \sum_{i=A,T,\mathscr{X}} p(Y=i)H(X|Y=i) \\ &= p(Y=A)H(X|Y=A) + p(Y=T)H(X|Y=T) + p(Y=\mathscr{X})H(X|Y=\mathscr{X}) \\ &= 0 \text{ bit} \end{aligned}$$

$\therefore H(X|Y) = 0$ bit.

The information channel capacity becomes

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} H(X) \text{ since } H(X|Y) = 0 \\ &= 1.58\text{bits}. \end{aligned}$$

Alternatively;

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} H(X) \\ &= \log_2 |\mathrm{A}_X| \\ &= \log_2 3 \\ &= 1.58\text{bits}. \end{aligned}$$

Thus, the coding capacity of this channel $b = 1.58\text{bits}/\text{nucleotide}$.

This result, agree with the theory we have discussed in Chapter 2. Using the three examples, we were able to show that increase in number of states increases the entropy. The value of the coding capacity obtained here is relatively bigger compared to the one obtained in the similar case (in the previous section), when four bases of the DNA sequence were used. Therefore, with many states we are able to achieve maximum coding capacity.

Net Information Density of Error Free Channel. The channel is error free and therefore the encoded sequence is perfectly received with the probability $1 - \delta_v$ without dropout rate. We use the same value of δ_v as described above. The net information density is therefore estimated to be;

$$\begin{aligned} C_{nt} &= (1 - 0.005)1.58 \text{ bits per nucleotide} \\ &= 1.57 \text{ bits per nucleotide} \end{aligned}$$

Thus, the error free channel with six DNA bases has the net information density $C_{nt} = 1.57$ bits per nucleotide. There is no an experimental result to compare with. But with the same logic and method we have used in the four alphabets DNA device in section 3.1.1, we conclude that the value we have approximated is reasonable value as the one we have approximated using the four alphabets DNA device.

The results show that, we have achieved the large net information density than in the previous case above when we used four bases DNA storage device. This implies that, we can encode large amount of data in the DNA storage device if the six bases are used instead of the four bases. This gives an important future application of the recently discovered synthetic DNA with six bases i.e. **(A,T,G,C,$\mathscr{X}$,$\mathscr{Y}$)**

Coding Capacity of Non- Error Free Channel. As we have discussed in the four alphabets DNA device, the objective is to maximize $H(X)$ and minimize as much as possible $H(X|Y)$ so as to increase the coding capacity. We maximize the $H(X)$ by ensuring the maximum distribution of $p(x_i)$ in the input alphabet $\mathcal{X}$

$$p(X = A) = p(X = T) = p(X = G) = p(X = C) = p(X = \mathscr{X}) = p(X = \mathscr{Y}) = \frac{1}{6}.$$

Hence,

$$H(X) = \log_2(6) = 2.58\text{bits}.$$

And we minimize $H(X|Y)$ with the joint probability distribution shown in the Table 3.4. The probabilities have been obtained using the method discussed by Rivas (2005).

		$\mathcal{X}$					
		A	T	G	C	$\mathscr{X}$	$\mathscr{Y}$
$\mathcal{Y}$	A	0.002	0.005	0.000	0.000	0.000	0.000
	T	0.005	0.134	0.000	0.000	0.000	0.000
	G	0.000	0.000	0.162	0.004	0.000	0.000
	C	0.000	0.000	0.011	0.139	0.000	0.000
	$\mathscr{X}$	0.000	0.000	0.000	0.000	0.187	0.003
	$\mathscr{Y}$	0.000	0.000	0.000	0.000	0.001	0.149

Table 3.4: Joint probability distribution

$$\begin{aligned} H(X|Y) = & \sum_{i=A,T,G,C,\mathscr{X},\mathscr{Y}} p(Y=i)H(X|Y=i) \\ & = p(Y=A)H(X|Y=A) + p(Y=T)H(X|Y=T) + p(Y=G)H(X|Y=G) \\ & + p(Y=C)H(X|Y=C) + p(Y=\mathscr{X})H(X|Y=\mathscr{X}) + p(Y=\mathscr{Y})H(X|Y=\mathscr{Y}) \\ & = 0.205(0.16543) + 0.139(0.2235) + 0.166(0.16386) + 0.150(0.37824) \\ & + 0.19(0.11709) + 0.150(0.057778) \\ & = 0.189\text{bits} \end{aligned}$$

Recall

$$C = \max_{p(x)} I(X;Y)$$

Thus

$$C = (2.58 - 0.189) = 2.39 \text{ bits/nt}$$

Therefore, the coding capacity of the six alphabets DNA channel $C = 2.39$ bits/nt.

Net Information Density of Non-Error Free Channel. We use the $\delta_v = 0.005$ and $\delta = 0.5\%$ in equation 3.1.1 we get

$$\begin{aligned} C_{nt} &= (1 - 0.005)^2 2.58 \text{ bits per nucleotide} \\ &= 2.38 \text{ bits per nucleotide.} \end{aligned}$$

Therefore, under given constraints we can achieve to make a six alphabets device with the net information density of 2.38 bits per nucleotide.

Similarly, we have obtained the values of net information density and coding capacity larger than in the four bases DNA storage device. This confirm that, the coding capacity and the net information density depends also in the number of independent bases in our storage device.

3.2 Related Works

Experimental results differ depending on various schemes used during the encoding and decoding processes(Goldman et al., 2013; Grass et al., 2015; Erlich and Zielinski, 2017). Despite the fact that different works have reported different results of the DNA storage device capacity, but all their results depends on common factors such as the length of oligos used, the size of the homopolymers, the index length and the dropout rates. Erlich and Zielinski (2017) used the DNA storage architecture that makes use of 150nt oligos, with the cost effective reduction rate $\delta_v = 7\%$ and the dropout rate δ of 5% . Their scheme provides best results in terms of retrieved data and storage capacity compared to the pre-existing ones. The scheme screens potential oligos to obtain a maximum coding capacity of $b = 1.98$bit/nt. Mathematical approximation of b provided in their work is given below;

$$b \approx \log_2 n - \frac{3\log_2 e}{4^{m+1}} + \frac{\log_2[2\Phi(2\sqrt{l}c_g c) - 1]}{l}. \tag{3.2.1}$$

Whereby;

$l =$ length of the oligo.

$n =$ number of independent events observed(number of bases in a sequence).

$\Phi =$ The cumulative function of standard normal distribution.

$c_{gc} =$ Maximum deviation of GC content from 0.5.

The derivation of the above approximation is given in Erlich and Zielinski (2017). We focus on the practical application of the approximation above to approximate the coding capacity.

Coding capacity in four alphabet channel. Erlich and Zielinski (2017) used $m = 3$, $l = 150$nt and a conservative value of $c_{gc} = 0.05$ followed some other previous works which were proved experimentally to be reasonable. For this choice we found out the value of b by substituting the given parameters in the equation 3.2.1. The value of n in this case is 4 because we have four types of bases in our sequence.

$$b = 2 - 0.017 - 0.002 = 1.98 \text{ bits/nt}$$

Net information density of four alphabet channel. Using the scheme suggested by Erlich and Zielinski (2017) which has been proven to be experimentally cost effective and reasonable, we calculate the net information density of the four alphabet channel. Net information density, given by equation 3.1.1, is found to be

$$C_{nt} = (1 - 0.07)1.98 = 1.841 \text{ bits/nt.}$$

Putting into consideration the dropout rate of 5% during decoding we found the net information density to be

$$C_{nt} = (1 - 0.07)(1 - 0.005)1.98 = 1.832 \text{ bits/nt.}$$

Hence, with a reasonable scheme of four alphabet DNA storage device, the net information density becomes 1.832 bits/nt. We found the same value in the previous section where we used a different approach based on mutual information theory to quantify the the net information density in the four alphabets channel. We therefore conclude that, our results agree with the estimate in the scheme used by Erlich and Zielinski (2017).

Coding capacity in six alphabet channel. Unfortunately, there is no an experimental scheme yet that has used the six alphabets channel to quantify the storage capacity of the DNA device. Intuitively, we use the similar approach as the one we have used in the four alphabets DNA device to calculate the coding capacity and the net information density with the same parameters. We substitute the values of the parameters given in equation 3.2.1 but in this case we use $n = 6$ because we are using a six alphabets DNA storage device.

$$b = \log_2(6) - 0.017 \, + \, 0.002 = 2.57 \text{ bits/nt.}$$

Thus, we can achieve the coding capacity of 2.57 bits/nt with six alphabets DNA device.

Net information density of six alphabet channel. Similary, we calculate the net information density using the equation 3.1.1

$$C_{nt} = (1 - 0.07)2.57 = 2.39 \text{ bits/nt.}$$

Considering the dropout rate in the channel we get

$$C_{nt} = (1 - 0.07)(1 - 0.005)2.57 = 2.38 \text{ bits/nt.}$$

Hence, with the same scheme used in previous section but with six alphabets DNA storage device, the net information density becomes 2.38 bits/nt. The net information density obtained here, is the same value we have obtained in our theoretical estimation. We therefore conclude that, the two approximation we have shown give same value of net information density.

3.3 Summary

We summarize the results in the tables below. The table 3.5 represents the values obtained in error free channel and table 3.6 for non-error free channel.

Architecture	Coding capacity b		Reduction probability δ_v		Net information density C_{nt}	
	Current work	Church et al.	Current work	Church et al.	Current work	Church et al.
4 alphabet	1 bit/nt	1 bit/nt	0.005	-	0.995 bits/nt	0.83 bits/nt

Table 3.5: Error free DNA-storage device

Architecture	Coding Capacity values b		Reduction probability δ_v		Net information density C_{nt}	
	Current work	Erlich et al.	Current work	Erlich et al.	Current work	Erlich et al.
4 Alphabet	1.844 bits/nt	1.98 bits/nt	0.005	0.07	1.83 bits/nt	1.83 bits/nt
6 Alphabet	2.39 bits/nt	2.39 bits/nt	0.005	0.005	2.38 bits/nt	2.38 bits/nt

Table 3.6: Non- error free DNA-storage device

4. Conclusions and Recommendations

DNA storage devices have proven so far to be better devices that offer many advantages compared to the available traditional storage devices in the market. The technology required in encoding and decoding data in DNA devices, is significantly expensive, sophisticated and prone to error. The development of DNA storage devices, requires design of complicated experimental schemes which consume a lot of time and difficult to handle. However, DNA storage devices, is a promising solution in dealing with the enormous growth of data.

We have developed a theoretical tool using the information theory that allows us to formulate the DNA storage device as an information transmission channel that maximize as much as possible its capacity and minimize as much as possible the amount of errors in the channel. We have tested the approach we have developed with other related works and confirm an excellent match.

We have found that, four bases DNA storage device gives a storage architecture with lesser capacity than the the modern synthetic DNA which has six bases i.e. **(A,T,G,C,**$\mathscr{X}$**,**$\mathscr{Y}$**)**. The the six bases DNA architecture offers the storage capacity that is 23.11% higher than the . We suggest that, future experimental work may concentrate on using the six bases DNA architecture instead of the four bases architecture because it offers higher capacity than the four bases architecture.

Acknowledgements

I am grateful to the God for the good health and wellbeing that were necessary to complete this work.

I wish to express my sincere thanks to my tutor Carine Umulisa for her sincere and valuable guidance and encouragement that extended me to the accomplishment of this work.

I am also grateful to my supervisor Dr. Ndifon for his professional guidance, encouragement, constructive criticisms and comments that have resulted to the success of this work. I take this opportunity to express gratitude to all my colleagues and friends for their help and support.

I also thank my parents for their unceasing encouragement, support and attention.

I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in the succession of this work.

References

C. Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1(1):3–22, 2004.

C. Adami. The use of information theory in evolutionary biology. *Annals of the New York Academy of Sciences*, 1256(1):49–65, 2012.

M. Blawat, K. Gaedke, I. Huetter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church. Forward error correction for DNA data storage. *Procedia Computer Science*, 80:1011–1022, 2016.

J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss. A DNA-based archival storage system. *ACM SIGOPS Operating Systems Review*, 50(2):637–649, 2016.

J. Bouck, W. Miller, J. H. Gorrell, D. Muzny, and R. A. Gibbs. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Research*, 8(10):1074–1084, 1998.

G. M. Church, Y. Gao, and S. Kosuri. Next-generation digital information storage in DNA. *Science*, 337(6102):1628–1628, 2012.

Computer-weekly. Big data storage choices. Data choices, http://www.computerweekly.com/feature/Big-data-storage-choices, Retrieved 11 Mar. 2017.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

P. Y. De Silva and G. U. Ganegoda. New trends of digital data storage in DNA. *BioMed Research International*, 2016, 2016.

Y. Erlich and D. Zielinski. DNA fountain enables a robust and efficient storage architecture. *Science*, 355(6328):950–954, 2017.

Z. Ezziane. DNA computing: applications and challenges. *Nanotechnology*, 17(2):R27, 2005.

F. Farhangmehr, M. R. Maurya, D. M. Tartakovsky, and S. Subramaniam. Information theoretic approach to complex biological network reconstruction: application to cytokine release in raw 264.7 macrophages. *BMC systems biology*, 8(1):77, 2014.

W. Feller. *An introduction to probability theory and its applications: volume I*, volume 3.

Genetic Home Reference. What is DNA? Wikipedia, https://ghr.nlm.nih.gov/primer/basics/dna, Retrieved 20 April. 2017.

P. Godfrey-Smith and K. Sterelny. Biological information. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2016 edition, 2016.

N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435): 77–80, 2013.

R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8):2552–2555, 2015.

M. Gymrek. Pcr-free library preparation greatly reduces stutter noise at short tandem repeats. *bioRxiv*, page 043448, 2016.

E. G. Learned-Miller. Entropy and mutual information. *Department of Computer Science, University of Massachusetts, Amherst*, 2013.

F. M. Lopes, E. A. de Oliveira, and R. M. Cesar. Inference of gene regulatory networks from time series by tsallis entropy. *BMC systems biology*, 5(1):61, 2011.

Z. Mousavian, K. Kavousi, and A. Masoudi-Nejad. Information theory in systems biology. part i: Gene regulatory and metabolic networks. In *Seminars in cell & developmental biology*, volume 51, pages 3–13. Elsevier, 2016.

G. Paun, G. Rozenberg, and A. Salomaa. *DNA computing: new computing paradigms*. Springer Science & Business Media, 2005.

A. Rhee, R. Cheong, and A. Levchenko. The application of information theory to biochemical signaling systems. *Physical biology*, 9(4):045011, 2012.

E. Rivas. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC bioinformatics*, 6(1):63, 2005.

P. Saxena, A. Singh, and S. Lalwani. Use of DNA for computation, storage and cryptography of information. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 3 (2):2278–3075, 2013.

C. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

C. Shannon. The best detection of pulses. In N. J. A. Sloane and A. D. Wyner, editors, *Collected Papers of Claude Shannon*, pages 148–150. IEEE Press, New York, 1993.

S. Tagore, S. Bhattacharya, M. Islam, and M. Islam. DNA computation: application and perspectives. *J. Proteomics Bioinform*, 3:234–343, 2010.

Tutorpace. DNA. Tutorpace, http://biology.tutorpace.com/images/DNA-Diagram-1.png, Retrieved 20 April. 2017.

UBS. LTI. Digital Data, https://foresight.ubs.com/longer-term-investments-digital-data/, Retrieved 20 April. 2017.

Wikipedia. DNA. Wikipedia, https://en.wikipedia.org/wiki/DNA, Retrieved 20 April. 2017a.

Wikipedia. Oligonucleotide. Wikipedia, https://en.wikipedia.org/wiki/Oligonucleotide, Retrieved 20 April. 2017b.

S. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic. A rewritable, random-access DNA-based storage system. *Scientific reports*, 5, 2015.

X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104, 2012.