2006

Information analysis of DNA sequences

Riyazuddin Mohammed
Louisiana State University and Agricultural and Mechanical College

INFORMATION ANALYSIS OF DNA SEQUENCES

A Thesis
Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

in

The Department of Electrical and Computer Engineering

By
Mohammed Riyazuddin
Bachelor of Engineering, Osmania University, 2003
December 2006

Dedicated to

my parents

ACKNOWLEDGEMENTS

I would like to express my gratitude towards Dr. Subhash Kak, my graduate advisor at LSU, for his kind support and encouragement throughout my thesis work. His constant motivation, discussions and suggestions were instrumental in allowing me to compile this work in its current form. I would like to thank Dr. Hsiao-Chun Wu and Dr. Xue-Bin Liang for sparing time to be a part of my thesis advisory committee.

I would like to thank my brother Mr. Mohammed Naseeruddin for his tips on C scripting and moral support throughout. My thanks are also due to all my friends who have made my stay at LSU a pleasant one.

Finally, I would like to thank my parents for their continual kindness and friendship throughout my Masters study at LSU.

TABLE OF CONTENTS

ABSTRACT

The problem of differentiating the informational content of coding (exons) and non-coding (introns) regions of a DNA sequence is one of the central problems of genomics. The introns are estimated to be nearly 95% of the DNA and since they do not seem to participate in the process of transcription of amino-acids, they have been termed "junk DNA." Although it is believed that the non-coding regions in genomes have no role in cell growth and evolution, demonstration that these regions carry useful information would tend to falsify this belief. In this thesis, we consider entropy as a measure of information by modifying the entropy expression to take into account the varying length of these sequences. Exons are usually much shorter in length than introns; therefore the comparison of the entropy values needs to be normalized. A length correction strategy was employed using randomly generated nucleonic base strings built out of the alphabet $\{A,T,G,C\}$ of the same size as the exons under question. The distance between exons and introns is calculated based on their probability distributions. We found that Zipf's distribution was not followed by the n-tuples in DNA sequences, and a newly modified power distribution derived from the Zipf's distribution was found by trial and error that closely modeled the codon frequencies. Correlation and divergence tests were performed. Our analysis shows that introns carry nearly as much of information as exons, disproving the notion that they do not carry any information. The entropy findings of this thesis are likely to be of use in further study of other challenging works like the analysis of symmetry models of the genetic code.

CHAPTER 1

INTRODUCTION

This thesis uses several methods to analyze the randomness of genomic sequences to measure the information content of exons and introns, which are the substrings that are the coding and the non-coding regions of DNA. We use finite character sequence theory for this problem, which has applications in other fields of science, such as spins in one-dimensional magnets, texts written in formal/informal languages also [1].

We will begin with introducing important definitions, the alphabet and terms commonly encountered in dealing with biological sequences and further express the motivation behind using DNA sequences for the research. With the advent of internet and enhancements in research technologies in bioinformatics, a vast number of DNA sequences are quickly and easily available in the form of strings of characters. Dedicated DNA sequencing centers render sequenced genetic data available online for free download and research purposes. In this work the NCBI database was used to collect DNA sequences of various organisms for the analysis.

1.1. Basics of DNA and Entropy

1.1.1. About DNA

The DNA (Deoxyribonucleic acid) molecule residing in the cell nucleus encodes information conventionally represented as a symbolic string over the alphabet $\{A,T,G,C\}$ [2]. The DNA molecule has a complex double helical structure (figure 1) which is formed as a result of folding between single strands of DNA. A single strand of DNA is a chain of nucleotides each of which consist of a base, sugar (S) and a phosphate (P) group. The letters of the alphabet above are derived from names of the four bases: A (Adenine),

Thymine (T), C (Cytosine) and G (Guanine). The combination between single strands of DNA takes place according to "Watson-Crick complementarity" that says that the only permissible combinations between bases are A-T or T-A and C-G or G-C hence one strand can easily be used to predict the other in a double stranded chain.

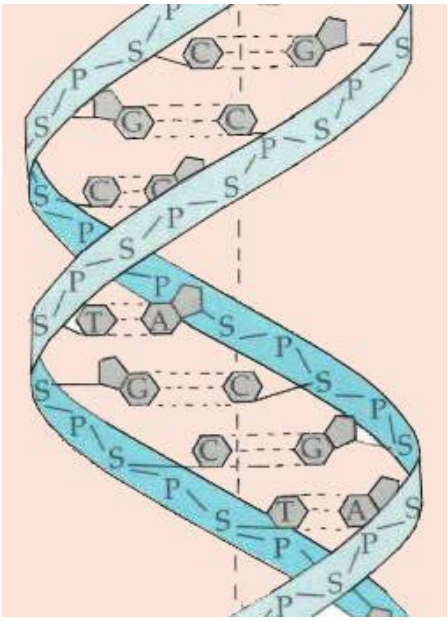


Figure 1. Double Helical structure of DNA

Central dogma of a cell and genetic code

The process of conversion of DNA to proteins involves the key stages: Transcription and Translation, according to Crick's Dogma of cell biology – figure 2.

Figure 2. Central Dogma of Cell Biology (Crick)

At the transcription stage, coding (exons) and non-coding (introns) DNA regions are separated and the Thymine (T) base is replaced by Urasil (U) to yield an intermediate polymer called "mRNA (messenger RNA)". During Translation, exons from different

positions in the genome are all concatenated and protein sequences or amino acid chains are then generated according to the genetic code (figure 4).

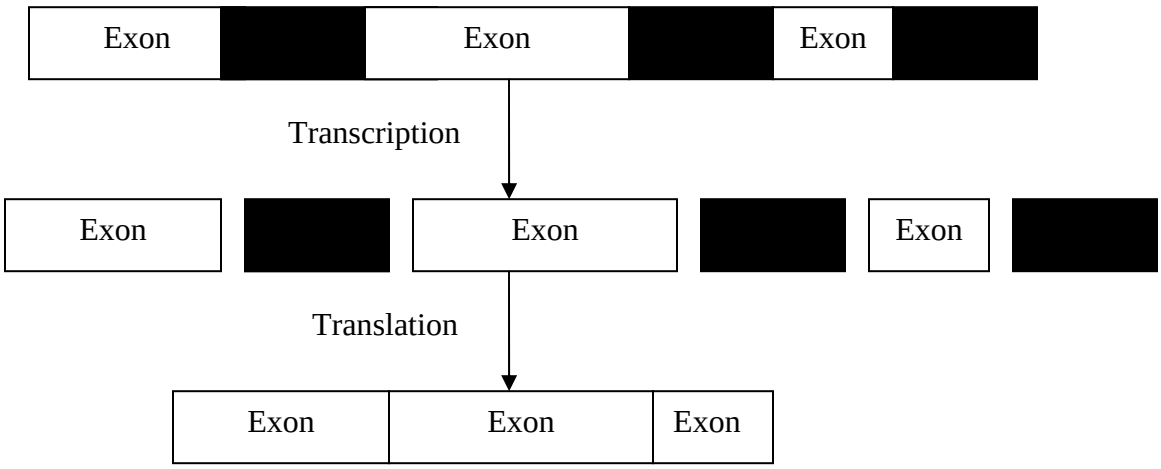


Figure 3. DNA Sequence on Translation

As illustrated above, all non-coding regions of DNA are cut out during translation. This fact that large portions of the RNA are removed before further translation is regarded as one of the most unexpected findings in molecular biology [cited in 3]. During the translation of DNA to proteins, one or more of the codons map to one of the 20 amino acids according to the "Universal Genetic code" of the organism, resulting in a sequence of amino acids as shown in the example below.

DNA Alphabet: ***{A,T,G,C}***

DNA Sequence: ***ATGCCGCCCAAAACCCCCCGAA……..***

Translated Protein Sequence: ***MPPKTPR****...*

The cell dogma sums up the translation of DNA to proteins which in reality includes a sequence of processes before the amino acids are generated. These include generation of intermediate sequences like mRNA and tRNA which themselves play specific roles in the translation. Mark White expresses the different levels of processing during genetic

translation as analogous to a computer algorithm in his work titled “Rafiki genetics” [4]; “The first level of the process takes DNA as input and performs a function that translates it into mRNA. The output of this process is fed into another function that translates the Information into tRNA according to its own set of rules. Perhaps investigators have yet to give us enough experimental data at this level to track the string Info accurately through a string made of tRNA molecules, but we can establish parameters at each level and take broad measurements of information entropy. These entropy-tracking protocols can then be used to query the process and broadly investigate the form and flow of information within and between various organic programs provided by nature.”

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 4. Universal Genetic Code of an organism

Figures 5 and 6 visually illustrate the analogy of genetic translation with computer processing and programming logic where DNA sequences may be considered as the data, the genetic code as the processor and the proteins denote results displayed on the monitor.

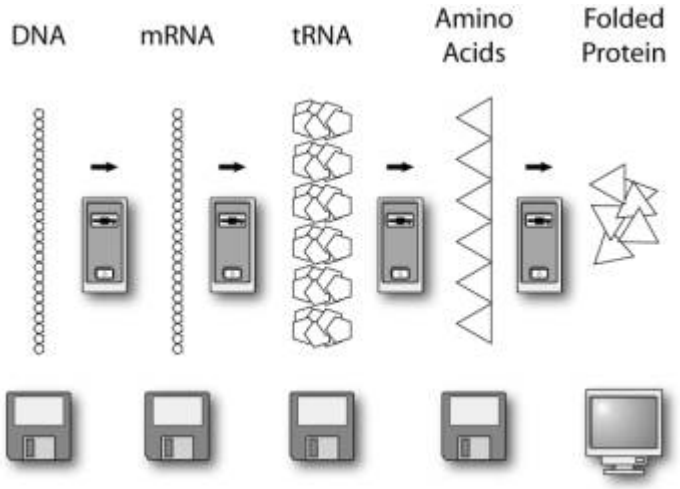


Figure 5. Levels of genetic translation (Rafiki Genetics, Mark White [4])

```
Do
  Function DNA (Info)
    Function mRNA (Info)
      Function tRNA (Info)
        Function Protein (Info)
          Function Cell (Info)
            Function Organism (Info)
              Function Species (Info)
                IF NOT SURVIVE (Info) THEN END
              End Function
            End Function
          End Function
        End Function
      End Function
    End Function
  End Function
Loop
```

Figure 6. Genetic translation algorithm (Rafiki Genetics by Mark White [4])

The length of a DNA sequence expressed in terms of base pairs (bp) varies from few thousands to several million bp. Although information in DNA sequence is normally analyzed using classical information theory, some quantum approaches have also been presented to better account for the structure. For quantum information and the logic behind its use in biological systems, see [5,6,7]; it has been suggested that codon symmetries can be better captured using quantum approach [8].

1.1.2. Entropy

Information Entropy was first introduced by Shannon [9]. Suppose X be a random variable that assumes the values $x \in X$, X being a finite set and the probability that X assumes the particular value x is denoted by $\Pr(x)$. Then the *Shannon entropy* of the random variable X is defined as,

$$H(X) = -\sum_{x \in X} \Pr(x) \log_2 \Pr(x) \tag{1.1}$$

The entropy $H(X)$ measures the average uncertainty in terms of bits of the outcome of the random variable X [10]. The Shannon entropy is a measure of the order and disorder in sequences [4]. The entropy of a finite character sequence of length *N* is defined as,

$$H(X) = \sum_i p_i \log(1/p_i) \tag{1.2}$$

where i extends over all symbols of the alphabet,

p_i is the probability of occurrence of symbol s_i at any position.

And $p_i \in [0,1]$ for all $i = 1...N$ and $p_1 + ...p_N = 1$

For a given context, entropy is a measure of the order or disorder in a sequence that can be regarded as information [11]. The estimate of sequence entropy depends on the probabilities of words in the sequence and a general form of the probabilities is written as,

$$P(A) = \frac{n_A + \beta}{N + \beta d} \tag{1.3}$$

where n_A is the frequency of event A among N total samples,

d is the cardinality of the alphabet,

β is a constant chosen as per case, β = 0 for normal maximum likelihood estimation and β =1 was proposed by Laplace.

If Y is another random variable that assumes values $y \in Y$, then the conditional entropy may be defined as,

$$H(X \mid Y) = \sum_{x \in X,\, y \in Y} \Pr(x, y) \log_2 \Pr(x \mid y) \tag{1.4}$$

This implies that Y carries information about X and the knowledge of Y reduces the average uncertainty about X. The mutual information between X and Y is defined as,

$$I[X;Y] = H[X] - H[X \mid Y] \tag{1.5}$$

We shall now consider a chain of random variables $S_1, S_2, S_3 \ldots$ that range over a finite set A. This chain may be viewed as 1-D spin system, a stationary time series of measurements, or an orbit of a symbolic dynamical system [10]. The Shannon entropy for this block S^L of variables may be defined as,

$$H(L) = -\sum_{s \in A} P(S_L) \log(S_L) \tag{1.6}$$

1.2. Motivation

The complexity and information carrying capacity of DNA data makes genomic sequence analysis an attractive research area today. More than 90% of the genome is known to be non-coding DNA (introns) and only 3-5 % of the sequence is the coding region (exons). It is well known that Richard Roberts and Phillip Sharp won the 1993 Nobel Prize in Physiology and Medicine for their discovery of introns. Although it is believed that the non-coding regions in genomes have no role in cell growth and evolution, demonstration that these regions carry useful information would tend to falsify this belief.

DNA sequence analysis presents challenges in applying finite sequence theory and provides opportunity to explore for improvement on existing techniques. Intron sequences have been regarded by some researchers as once active genes that were involved in the

evolution process but do not have any useful function now, much like the vestigial organs in the human body that are remains of our evolutional history [11].

Increasing availability of DNA data on the internet makes it possible to implement statistical and other techniques on sequences of different organisms. The advancement in technology over the past decade or so has created greater interest in studying genes, cell replication and the complexity of DNA. It seems more likely that the introns have an unknown function, although evidence is needed to indicate the certainty that they have structure and carry useful information.

1.3. Outline of Thesis

This work uses novel techniques in information theory to study the structure of exons and introns and to allow a reasonable comparison. Shannon's entropy of a finite sequence was primarily used as an analysis tool and to implement a benchmarking technique. Other techniques explored include Autocorrelation, Divergence and Kak's randomness test. MATLAB and C programming tools were used in the implementation of these novel methods.

The DNA character strings need to be converted to numerical values to apply mathematical tools on them. The bases were numericalized by substitution for performing the autocorrelation tests. It has been suggested in previous work that the coefficients of the Walsh Transform may be used to study the degree of randomness of a sequence [14]. This technique, called as Kak's randomness test was implemented for the total coding region in a genome and an intron sequence of a similar length.

The Shannon's entropy of exon and intron sequences was calculated by breaking up the sequence into sub-sequences of length L. An entropy plot is obtained by varying L and

calculating the probabilities of the sub-strings each time. The entropy thus obtained is termed as the "block entropy" which we shall denote as $H(E_i)$. The entropy per character (base) of the sequence is then obtained by normalizing the block entropy with the respective L i.e.

For $L = 3, H(E_3)$ is normalized as,

$$H(E_3)' = H(E_3)/3 \tag{1.7}$$

The entropy plots of various intron and exon sequences show that the entropy converges on increasing the search strings length. It follows that the entropy of sequences of a fixed length is a function of the finiteness of the sequence. There is hence a need to normalize the entropy in order to make a generic comparison of entropy patterns for sequences of different lengths. For this purpose, we have used randomly generated sequences from the alphabet $\{A,T,G,C\}$ of length equal to the DNA sequence under analysis, to obtain a proportionate correction factor for benchmarking the entropy values. An ensemble of random sequences having the same length was used to obtain optimum values of the benchmarking entropy values.

Finally, an approximate power distribution derived from Zipf's distribution was calculated that closely model word frequency distributions of a set of exon and intron sequences. The distribution was observed to represent the codon frequencies of exons with an error of the order of 0.1% and introns with an order of 1%. Kullback-Leiber distance measure or divergence of exon and intron sequences was calculated using the individual codon probabilities. The non-commutative property of divergence is well known and this has been utilized to derive an approximate measure of similarity between exons and introns.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we will review some of entropy estimation techniques employed to character sequences so far. We will also briefly introduce and discuss issues like convergence of entropy, accuracy of the estimate, structure in character sequences, finiteness, and the need for robust length correction. Finally, we will present the different distance measures employed in practice and their relevance to DNA character sequences.

2.1. Importance of Entropy Estimation

The Shannon Entropy of a data sequence is used to describe the complexity, compressibility, amount of information, weight of noise component, and so on. The DNA character strings are formed of the 4-letter alphabet $\{A,T,G,C\}$. Techniques have previously been used to apply the information theoretic notion of entropy to estimate the entropy of DNA sequences. It makes intuitive sense that the entropy of exons and introns differ since they are subject to different random processes [9]. Based on novel entropy estimation methods, issues like intron/exon boundary problem, the entropic difference of exons and introns, and the structure and information content of these sequences may be addressed. It has been demonstrated with tests on various genetic sequences that a significant difference exists between intron and exon entropies obtained using a match length entropy estimator [9]. This fast converging estimator was used to address the exon/intron boundary (splicing) problem extending the concept of indicators that represent the start and end of exon sequences. It was proved that a meaningful signal may be extracted from portions of a DNA sequence using this estimator. Another key result of the estimator was that the entropy of the gene sequences that actually code for proteins is higher compared to other DNA segments.

This is in contrast to the biological theory prevalent at that time which explained that introns are capable of tolerating random sequences to a higher degree than exons.

2.2. Entropy Estimation Techniques

There are several methods for estimating the entropy of a random process. The most straightforward would be to find a direct computation of the expected log of the empirical distribution function. An entropy estimate thus obtained might only be as accurate as the estimate of the probability of n-tuples where n may be large. The entropy estimation is made difficult because of the shortness of the DNA sequences that code for proteins since the amount of data is practically insufficient to achieve a good estimate of all but the marginal or first order distribution and perhaps the distribution of pairs [9]. Data compression techniques like Lempel-Ziv (LZ) algorithm is another popular choice for entropy estimation. It is however known to have a slow rate of convergence for this purpose. Most of the techniques involve string matching and pattern frequency as part of the calculation. A match length entropy estimator has been proved to have a fast convergence rate relative to other techniques [9].

2.2.1. Lempel Ziv Algorithm

Consider a binary data sequence for example $\{10001010110011\}$ or a character sequence say $\{AATAGAACGAA\}$. This sequence is parsed into unique phrases separated by a delimiter (comma) after each contiguous substring completes a new pattern. Every such pattern forms a phrase and is automatically a part of the "dictionary" of patterns, with a new phrase formed by searching left to right down the sequence to find the shortest contiguous substring that isn't already found in the dictionary. The above example sequences will be parsed into $\{1,0,00,10,11,001,000\}$ and $\{A, AT, AG, AA, C, G, ...\}$ respectively. Suppose there are

C_n delimiters in the dictionary formed in a LZ parse for a sequence of length n, then the Lempel and Ziv formula [cited in 9] shows that,

$$\frac{C_n \log C_n}{n} \to H \tag{2.1}$$

where H is the Entropy estimate.

This inequality makes the method suitable for an entropy estimate. This technique is easy to implement and universally applicable. The string matching involved here is intuitively appealing as a measure of complexity since it quantitatively captures repetitive structure [9]. A slow rate of convergence is an obvious drawback of this scheme since many observations are needed to build the dictionary of patterns.

Lempel-Ziv Algorithm Using a Fixed Database

This method employs a fixed database LZ algorithm and is identical to versions practically used. It is assumed that we have a database D_n of n observations X_{-n+1}^{0}. Then the longest match of the input sequence $X_1, X_2, X_3, \ldots$ in the database D_n is represented as,

$$L = \inf\{k : X_1^{k+1} \not\subset D_n\} \tag{2.2}$$

where $\not\subset$ means as a contiguous substring.

For example, if $D_n = \{AATAGAACGATAGACCA\}$ with $n = 18$ and the input sequence was $X_1, X_2, X_3, \ldots = \{ATAATAGA\ldots\}$ then $L = 6$ since the pattern $\{AATAGA\} \subseteq D_n$. A group of theorems cited in [9] demonstrate how L can be used in entropy estimation.

THEOREM 1. If $\{X_k\}$ is a Uniform i.i.d, then for any positive integers l and n,

$$\Pr\{L < 1 + \log n\} \approx \exp(-2^l) \tag{2.3}$$

THEOREM 2. Suppose $\{X_k\}$ is a stationary, ergodic source with finite memory and if it is not a uniform i.i.d. sequence. Then,

$$\Pr\{X_k = x_k \mid X_{-\infty}^{k-1} = x_{-\infty}^{k-1}\} = \Pr\{X_k = x_k \mid X_{k-M}^{k-1} = x_{k-1}^{k-M}\} \tag{2.4}$$

THEOREM 3. As $n \to \infty$, $\left| E[L] - \frac{\log n}{H} \right| = O(1)$ (2.5)

The entropy may be estimated based on the length of repeated patterns in light of the above theorems. This scheme is potentially better than the earlier one since it applies the pattern search for every letter instead of every phrase.

2.2.2. A Sliding Window Technique

Suppose N_w is a chosen positive integer to denote the window size of observations that will serve as our database into which the input data needs to be referenced for finding the longest match. If the input sequence is $X_1, X_2, X_3, \ldots$ then for all indices i,

$$L_i = \min\{k : X_{i+1}^{i+k+1} \not\subset X_{i-Nw+1}^{i}\} \tag{2.6}$$

This is a sequence of random variables $\{L_i\}$ and theorem 3 is used to calculate the entropy estimate as,

$$\hat{H} = \frac{\log_2 N_w}{\bar{L}} \tag{2.7}$$

where $\bar{L}$ is the average of L_i

The two sources of error here are; (i) the standard error of $\bar{L}$ for a fixed N_w and (ii) the bias term $O(1/\log N_w)$ from Theorem 3. In a sequence of length n, there is roughly n/N_w

number of independent match lengths which means that $\log N_w$ can be typically made equal to $\bar{L}$. The following assumptions were made in this analysis:

- The entropy measure only approximates an entropy measure since the actual sequence or process is longer than the chosen N_w.
- DNA is not stationary and this entropy estimator is robust to weak conditions and non-stationary processes cannot be characterized by entropy.
- DNA is not a random process hence the math is perhaps more of a guide towards a meaningful statistic and one cannot claim to have characterized the entropy of DNA.

The difference in the entropy estimates obtained using the match length estimator are more reliable compared to the equivalent values using the LZ algorithm because the string matching is done for each letter of a given search length n in the former and for each phrase in the latter technique. A signed rank test was performed between the exon and intron entropies under the hypothesis that the two entropies are identical or statistically equivalent. The test was performed on paired comparisons of adjacent exon/intron sequences and it was found that out of 303 comparisons, about 73% of the showed the average match length to be higher for the intron thus negating the equality hypothesis [9]. As a further verification, tests were run with a randomly generated test sequence with equal probability of the 4 characters from $\{A,T,G,C\}$ and they showed no significant entropic difference between the two groups.

2.2.3. Effect of Long Range Correlations on Entropy Estimation

Werner and Thorsten and others have explored that genetic sequences seem to have long range correlations [1]. And another view says that "Repeated Patterns lead to lower Entropy". This implies that the correlations in DNA sequences will have an effect on the

entropy estimate to an extent depending on the algorithm used. The presence of structures in character sequences was reasoned in [1] to be perhaps due to the following:

- Predictability: Briefly, predictability refers to the inherent quality of texts like books, files or programs that helps us to know about the later portions just by reading the first few paragraphs, lines or pages depending on the length of the sequences.
- Syntactical Limitations: Another reason for expecting correlations is the exponential increase in the number of possible sub words with increase in length of uncorrelated strings. The no. of subwords is shown to increases by $N(n) = \lambda^n = \exp(\ln \lambda . n)$ for different subwords of length n. Furthermore, several texts were observed to show have growths obeying a power law like: $N^*(n) \approx n^\alpha$ or an exponential growth: $N^*(n) \approx \exp(Cn^\alpha)$
- Evolution: Finally, evolution is generally known to happen at regions that essentially have long range correlations.

The results in [1] where also reported to illustrate block entropies and mutual information as appropriate measures of the correlations and the degree of order in strings.

2.2.4. Estimation Using Conditional Probability and Hamming Distance

DNA can be understood as a highly ordered molecule with purpose and structure, and hence we can expect that the statistical models of its string representation may yield lower entropy estimates than a random string over and $\{A,T,G,C\}$ for which 2 bits/nucleotide is the optimal code [2]. As a striking fact, most natural DNA sequences including parts of the human genome yield counter intuitive results and conventional techniques have shown entropy values like 1.90 and 1.95.

David Loewenstern and Peter N. Yianilos introduce a concept of inexact match information in their model and have achieved entropy values as low as 1.66 [2]. The algorithm introduces hamming distance as the number of positions where two equal length strings differ to measure the closeness of a repeated string in natural DNA. A random variable b represents the nucleotide to be predicted i.e. it takes on values $b = 1,2,3,$ or 4 corresponding to A, T, G and C. Let w be a positive integer random variable that denotes the length of target sequence, f denote the first hamming distance that assumes values from 0 to w, i.e. the minimum h that yields matches for w in a given string of length l and h be the hamming distance index that again assumes a value from 0 to w. Now, if $past$ refers to the DNA already predicted or reference DNA then a natural prediction of b formed by locating distance i matches in the past to the target window k is $\Pr(b \mid h = i, w = k, past)$, which is independent of f and can hence be written as $\Pr(b \mid h = i, f = j, w = k, past)$. Then the prediction $\Pr(b \mid past)$ can be achieved as,

$$\Pr(b \mid past) = \sum_{i,\, j,\, k} \Pr(b \mid h = i, f = j, w = k, past).\ \Pr(h = i, w = k, past) \qquad (2.8)$$

Where the 2nd term is again a product of conditional probabilities i.e.,

$$\Pr(h = i, f = j, w = k, past) = \Pr(h = i \mid f = j, w = k, past).\Pr(f = j \mid w = k, past).$$
$$\Pr(w = k \mid past).\Pr(past)$$

(2.9)

The probabilities $\Pr(past)$ and $\Pr(f = j \mid w = k, past)$ are equal to 1 when j is equal to the distance of the closest match to the window length k and zero otherwise. Another assumption in the model is that the above defined conditional probabilities $\Pr(h = i \mid f = j, w = k, past)$ and $\Pr(w = k \mid past)$ is independent of the past. If we introduce

this independence and $\Pr(f = j \mid w = k, past)$ as a Boolean function $f(j, k, past)$ then the above probability $\Pr(b \mid past)$ becomes,

$$\sum_{i,\, j,\, k} \Pr(b \mid h = i,\ f = j, w = k,\ past).\Pr(h = i \mid f = j, w = k).f(j, k,\ past).\Pr(w = k) \tag{2.10}$$

The paper presents achieved entropy of 1.70 which has an improvement of 0.25 bits over 1.95 bits from standard techniques. The experimental data for the above analysis was taken from GenBank and the sequences were selected according if they are long enough for the method of entropy estimation used and if they belong to various different species to emphasize the generality of the proposed technique.

2.3. Distance Measures

The distance between two probability distributions may be defined as a 'distance measure'. Distance measures have been used in statistics over a long time beginning with the work of Pearson (cited in [15]). The D^2-statistic of Mahalanobis and the linear discriminant function introduced by Fisher were the two most popular measures in statistics (cited in [15]). After the invention of Shannon's information theory in 1948, the "divergence" has gained popularity due to its similarity with the logarithmic entropy measure of Shannon although it was proposed even before that by Jeffreys (cited in [15]). This concept of distance measures may perhaps be useful in differentiating between exons and introns based on their sequence distributions. In order to model the word frequencies and probabilities of words occurring in such sequences, we refer to standard distributions in statistics like Binomial, Poisson, Zipf's etc. We later find that a modified version of Zipf's distribution closely models the exon/intron sequences.

CHAPTER 3

CORRELATION AND RANDOMNESS TESTS

Knowing and understanding the correlation between bases appearing in DNA sequences has been an interest for a long time now [16]. We begin with an introduction to the framework of correlation of random processes and later apply the same to DNA sequences.

3.1. Mathematical Framework for Correlation

3.1.1. Basic Definitions

Covariance: Covariance is a measure of how much two or more variables or processes match. Two process X and Y have a small covariance value if are not closely related and if they are similar then the covariance is large. Covariance is mathematically defined as,

$$\mathrm{cov}(X,Y) = \sigma_{xy} = E[(X - \bar{X})(Y - \bar{Y})] \tag{3.1}$$

Correlation: Correlation is the amount of dependency of one random variable on another. If we consider two random variables X and Y, then the correlation between X and Y is equal to the average of their product i.e.,

$$cor(X,Y) = E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \tag{3.2}$$

Autocorrelation Function: Autocorrelation is the expected value of the product of a random variable or signal realization with a time-shifted version of itself. Assuming that we have two instances of the same random variable X as $X_1 = X(t_1)$ and $X_2 = X(t_2)$, the autocorrelation of X is written as,

$$R_{XX}(t_1 t_2) = E[X_1 X_2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2)dx_2 dx_1 \tag{3.3}$$

This definition is valid for both stationary and non-stationary random process. It has been proved that the expected values for a stationary random process are depend on the time difference $\tau = t_1 - t_2$ and hence the representation,

$$R_{XX}(t, t+\tau) = R_{XX}(\tau) = E[X(t)X(t+\tau)] \tag{3.4}$$

We will be applying autocorrelation to a real DNA sequence and we would like to look at the discrete time case of autocorrelation.

$$R_{XX}(n, n+m) = \sum_{-\infty}^{\infty} x(n)x(n+m) \tag{3.5}$$

Properties of Autocorrelation

It might be particularly useful to look at the properties of autocorrelation here to allow a better understanding of the results obtained on applying the autocorrelation concept later on.

1) Autocorrelation is an even function of τ, $R_{XX}(\tau) = R_{XX}(-\tau)$
2) The mean-square value of the random variable may be calculated by the autocorrelation at $\tau = 0$ and this is the largest value of autocorrelation.

   $R_{XX}(0) = \bar{X}^2$; $R_{XX}(0) \geq | R_{XX}(\tau) |$
3) The autocorrelation function of period function is also period.

All definitions are from [17].

3.2. DNA Sequence as a Random Process

Under the above framework, we will introduce the parameters involved in terms of DNA character sequences. Lets assume the DNA character sequence under question i.e. the exon or intron sequence to be a sample space ‘S’ of randomly occurring character strings and let the occurrence of a base at a given position k in the genome be a discrete random variable ‘X’. We shall attempt to illustrate that there is underlying structure and patterns in DNA

sequences. A simple proof of structure is the unequal probabilities of the characters $\{A,T,G,C\}$ obtained for sample exon and intron sequences taken from HUMRETBLAS are as follows:

For an Exon sequence of length 137 bases,

P(A) = 0.18, P(T) 0.06, P(C) = 0.47 and P(G) = 0.29

For an Intron sequence of length 3227 bases,

P(A) = 0.29, P(T) = 0.30, P(C) = 0.19 and P(G) = 0.23

The calculation and results of correlation for real DNA sequences are presented in the next section.

3.3. Autocorrelation Plots of DNA Sequences

The Autocorrelation of a sequence is defined using Equation 3.5 as,

$$R_{XX}(n, n+m) = \sum_{-\infty}^{\infty} x(n)x(n+m)$$

where $x(n)$ is discrete sequence of length.

The DNA character sequence under analysis has to be converted to a numeric sequence to apply mathematical functions on it. We are primarily interested in the autocorrelation of the sequences. Since autocorrelation is related to the pattern of letters in the sequence, we can use substitution to generate a numeric sequence. The characters {A,T,G,C} in the given sequence are substituted with arbitrary numerical values {-.5,.5,-1.5,1.5}. The substitution is done suitably to satisfy Watson-Crick's property which says A is complimentary with T and C with G . Numerical values are chosen such that correlation plots are symmetric about zero and hence easy to visualize.

A normalized form of the autocorrelation function was implemented in MATLAB by fixing the number of lags m to 10. For a discrete sequence of length N, the function used is represented mathematically as,

$$R_{xy}(m) = \frac{1}{N} * R_{xy}(m) \tag{3.6}$$

where m is the number of lags and N the length of the sequence under test.

Homosapiens genome: (Genome Length = 16569 bp)

Gene sequence (L = 957 bp) Random sequence (L = 957)

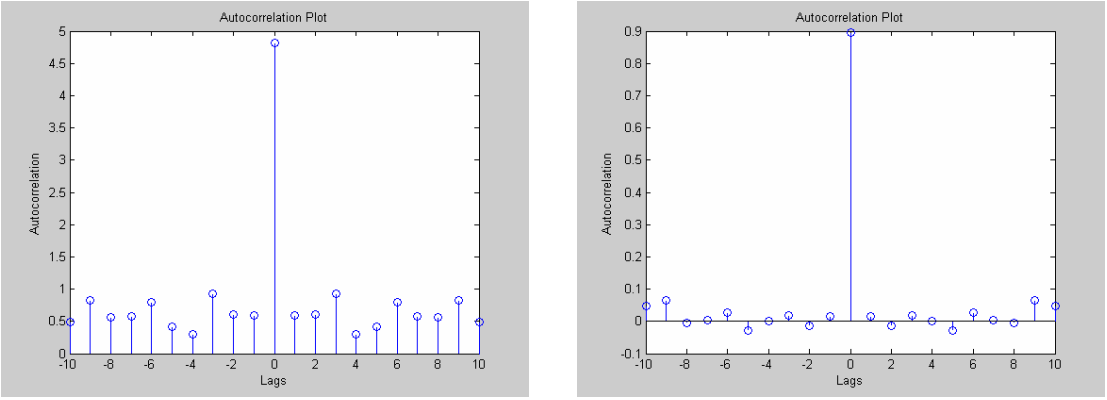


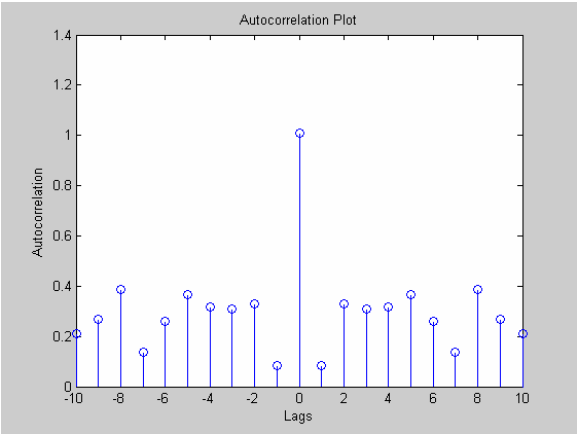
Figure 7. Autocorrelation of genetic sequence compared to a random sequence

It is evident from the above figure that a gene sequence shows structure compared to a random sequence of an equal length. The plot obtained is symmetric about the zero lag due to the symmetrical numerical values used in the sequence.

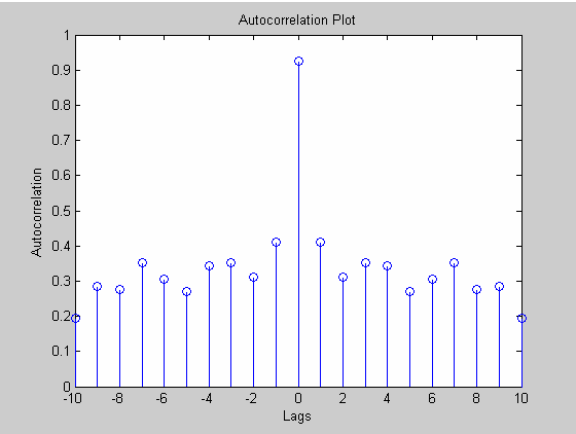
Exons and introns sometimes form part or whole of a gene sequence. We would be interested in studying the pattern and structure of the coding (exon) and non-coding regions (intron) in a DNA sequence. A comparison between their autocorrelation plots was our first test in this direction. Figure 8 shows autocorrelation plots for exon and intron sequences of similar length chosen from the HUMRETBLAS genome. A randomly generated sequence of the same length as the DNA sequences is used for comparison.

Humretblas genome: (Genome length = 180,388 bp)

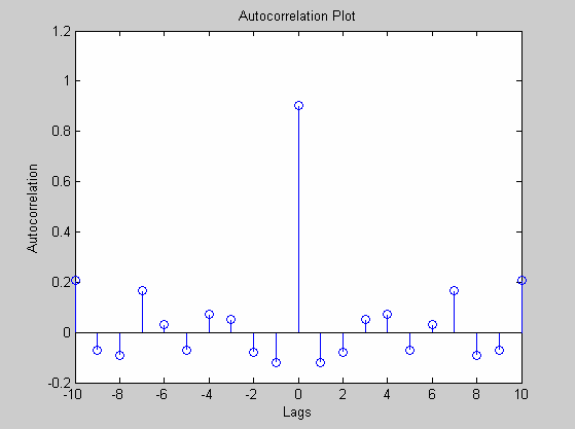
Exon (L = 77 bp)



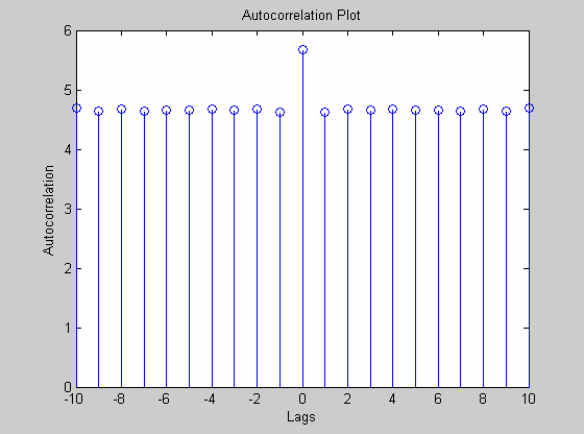
Intron (L = 77 bp)



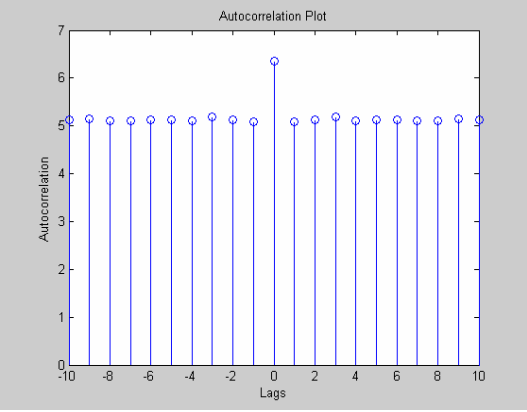
Random Sequence (L = 77 bp)



Intron (L = 2687 bp)



Humretblas coding regions (L = 2787)



Random Sequence (L = 2687)

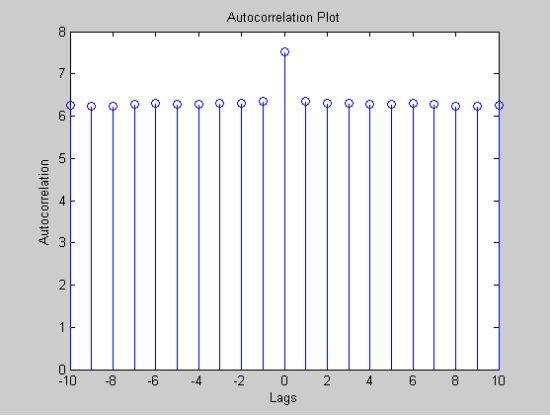


Figure 8. Autocorrelation plots of Introns and Exons

The amplitudes of autocorrelation at each lag point for exon and intron sequences of equal length were observed to differ only slightly and the pattern of the sequences seems identical. Intron sequences seem to carry meaningful patterns perhaps carrying information useful to the cell. The autocorrelation plots of introns and exons of similar length also suggest they may have structure comparable to exon sequences that encode critical information which is used in the translation of DNA to proteins.

3.4. Kak's Randomness Test

According to [14], "A sequence shall be said to have no pattern or be random if the number of independent amplitudes in the Walsh-Fourier transform is equal to the length of the sequence itself, i.e., 2^k." The measure of randomness r(s) shall therefore be defined in terms of Walsh Transform values in the frequency domain for the sequence under test i.e.,

$$r(s) = i(s) \big/ L(s) \tag{3.7}$$

where i(s) = no. of independent amplitudes W(s)

and L(S) =length of the sequence

In this, "The number of independent amplitudes of W(s) shall equal the number of its component Walsh waves."[14]

The Walsh amplitudes are calculated using the MATLAB function for Walsh Transform. Consider a sequence [1 2 1 1]. A sample output of the function is shown below:

walsh1D([1 2 1 1]) = 1.2500 0.2500 -0.2500 -0.2500

A simple C script was then used to count the number of independent amplitudes in this result.

Randomness Measure, r(s) = W(s)/L(s) = 3/4 = 0.75 [considering 0.25 and -0.25 as independent]

At this end, let us assume that the sequence is zero padded towards the end [1 2 1 1 0 0 0 0]

to increase the length of the final sequence to 8 characters i.e. the Walsh function is now,

walsh1D([1 2 1 1 0 0 0 0]) = 0.625 0.625 0.125 0.125 -0.125 -0.125 -0.125 -0.125

Randomness Measure**,** r(s) = W(s)/L(s) = 3/4 = 0.75 [considering 0.25 and -0.25 as different]

This observation clearly indicates that zero padding a sequence does not change the outcome of this randomness measure. This technique can be used to make the size of the DNA sequence equivalent to a 2^k.

Randomness Test Results

The Walsh function was applied to chosen intron and exon sequences and random sequences of the same length. The length of the sequence was adjusted to the nearest 2^k value either by truncating or zero padding the sequence. For example, the exon sequence of length 197 bp was zero padded to increase its length to 256. The following table shows results of this test.

Sequence	**Actual Length in bp**	**Adjusted to nearest 2^k**	**Kak's randomness coefficient R(s)/W(s)**
Humretblas	1,80,388 bp		
Exon 1 Random	197	256	41/256 45/256
Total Coding Region Random	2787	2048	146/2048 150/2048
Intron 1 Random	2687	2048	137/2048 133/2048

Figure 9. Results of Kak's Randomness Test

Due to the short length of exons, it's difficult to make a proper comparison of their structure to that of introns. For this purpose, we have used the total coding region that was generated by combining all exons from Humretblas and has a length comparable to the

introns. The total coding region and the non-coding (intron) sequence under analysis were truncated to the nearest value 2048 in order to apply the Walsh function. The results obtained indicate similar structure for the coding and non-coding region.

CHAPTER 4

BENCHMARKING FINITE SEQUENCE ENTROPY

4.1. Background of Finite Sequence Entropy

Entropy can be used to calculate the degree to which finite sequences can be compressed without any loss of information or to study structure of finite sequences. Although the existence of correlations in a sequence reduces the uncertainty of the symbols yet to be observed, it's important to locate them and account for them in our estimation. The most straightforward method of estimation would be based on the frequency of block strings up to a certain length and estimating their probabilities according to it.

$$\hat{p}(s_1, s_2, ..., s_n) = \frac{n_{s_1, s_2, ..., s_n}}{N} \tag{4.1}$$

where $n_{s1,s2,...sn}$ is the no. of occurrences of the word $s_1, s_2, .., s_n$

Then the entropy estimator may be written as,

$$\hat{h} = \lim_{n \to \infty} \hat{H}_n / n \tag{4.2}$$

Shannon's theory is based entirely on probabilistic concepts and deals with average code lengths but has a drawback that it doesn't take into account the information needed to describe the probability distribution itself.

4.2. Algorithm Used for Entropy Calculation

As we know from Shannon's Entropy, the mathematical formulation of entropy is,

$$H(X) = \sum_i \mathrm{p}(\mathrm{x_i}) \log \mathrm{p}(\mathrm{x_i}) \tag{4.3}$$

In the case of DNA character sequences comprised of letters from the 4-letter alphabet $\{A, T, G, C\}$, the entity X is represented as S and assumed to be either a coding

gene sequence or a non-coding sequence which we will refer interchangeably with exons and introns respectively. In order to compute the entropy using the Shannon's entropy shown in (4.3), the given sequence X of length N can be assumed to be comprised of a set of sub-strings of equal length L. Let us denote each of the sub-strings by g_i and the number of sub-strings that make the sequence of length N be n. In general, the number n is equal to N/L. Then the probability of each such sub-string $p(g_i)$ will be computed directly based on the frequency of occurrence of the string within the given DNA sequence i.e.

$$p(g_i) = \frac{n(g)}{N} \text{ Or } \frac{n(g)L}{N} \tag{4.4}$$

The entropy of a genetic sequence can then be represented as,

$$H(S) = -\sum_{i=1}^{N/L} p(g_i) \log p(g_i) \tag{4.5}$$

A C script was used to parse DNA character sequences and to calculate the frequencies of occurrence $n(g)$ for all sub-strings g_i. The script takes as input the genetic code and DNA sequence under question and implements a search algorithm to find the frequencies of occurrence of all the 64 possible codons from the genetic code in the input DNA sequence. The value of L is then varied from 3 to 9 and a range of values of $H(X)$ is obtained which are then used to illustrate the behavior of the sequence entropy with increase in search lengths. For every L, the genetic code table is updated to include all the possible L-tuples 4^L comprised of characters from the alphabet $\{A,T,G,C\}$ since the number of search strings increases with increase in L. The size of possible sub-string lengths in fact grows exponentially as follows:

No. of possible triplets: $4^3 = 64$,

No. of 4-tuples: $4^4 = 256$,

No. of 5-tuples: $4^5 = 1024$, and so on.

The entropy value obtained for different sub-string lengths is sometimes called as the block entropy and is mathematically represented by Eq. 1.6,

$$H(L) = -\sum_{s \in A} P(S_L) \log(S_L)$$

This value is normalized with the block length to calculate the entropy per base i.e.,

$$H(S) = \frac{H(L)}{L} \tag{4.6}$$

As mentioned earlier the convergence of the entropy estimator must be fast enough to accommodate the shortness of certain DNA sequences [9]. The number of possible substrings increases with increase in triplets from $\{A,T,G,C\}$ are $4^3 = 64$, the number of 4-tuples is $4^4 = 256$, the number of 5-tuples is $4^5 = 1024$, and so on. However, the length of sequence available is limited and most exon sequences are only about 200 bp long. If the DNA sequence were periodic to repeat itself, it would need a sequence length comparable to the set of possible sub-strings.

We have used the Humretblas genome that has a total length of 180388 bp from NCBI. The annotations provided by the database to indicate coding and non-coding regions were also used to obtain several exon and intron sequences for our analysis. Figure 10 below shows a sample of DNA data with the integers showing the position of bases in the sequences. The orientation or direction in which the sequence is read is indicated in the database by 'complement' and that needs to be observed before using the sequence.

The entropy results obtained are shown a tabular as well as a graphical form below. Figure 11 shows a tabular result of entropy values where as Figure 12 illustrates exon

entropies on a graph where y-axis indicates the entropy and sub-string lengths L = 3, 4,…9 are on the x-axis. Entropy convergence is rapid for exons when compared to some of the introns. As seen from the values, perhaps there is much similarity in the entropy convergence pattern of the coding sequence with that of an intron of a similar length.

```
ORIGIN
      1  gtaagtagtt cacagaatgt tatttttcac ttaaaaaaaa agatttttat ggaataatct
     61 caaacatctt gatagttagg gttagtttga tcgattatag caggctactt cataaattaa
    121 gcccatagat ttaagtcctg tgtagattat ttatcttctc acaaagaaaa tagtataaaa
    181 tacatgcctt gtactacaaa gaagaactaa taaggtggaa ttgattcagg acagcatatc
    241 accaactctg agaaaaatgc aacaaatgca aattcattga ctaaatcttt attgagggtc
    301 tgttacaggc actttattaa ctaataatca gcataatttc tgtgtgagaa taaatgtaaa
    361 aatctgtatt aaaatttcca aatgattatt ttaaatgtat aatgcatgct ctaacagtat
    421 gcccatgtag agctccagag ttttttcttg gaaacagaat gagtagtaca tgagattttc
    481 tgcctcattg gagtagtatt gaagataatt aatataaagg gaaattgtat atttactgat
    541 taattgatat caatctatta attccaacaa gtgaatgtct ctggaaagat tatcaaggca
    601 aagtgttaaa ttggcaaact aaagtcatcc aaaccttcat ttttctgctc acagtgttga
    661 taattaatca gaaaaaagag caaaaaatat taaggtaatt tgaaacaaag tatgttataa
    721 catactatgt tttttatata tttttatatt agaattgaaa tattcagtat ttcttttaca
    781 aaatttttct ttcaaaatgt atacttttt ttcttaattt tttttttgc agcttctcat
    841 ggtcaagaat gtatactatt ctgtgggcta aatatcatat cttagaatta taagacatag
    901 aaacattaaa tgaatagaga taaactcagg tgtaaattat gcaattaaaa tggactgcat
    961 tctattatgc atttaactaa ggtcattttt tttttaatgc acaaaaagaa acacccaaaa
   1021 gatatatctg gaaaactttc tttcagtgat acatttttcc tgttttttt ctgctttcta
   1081 tttgtttaat ag
//
```

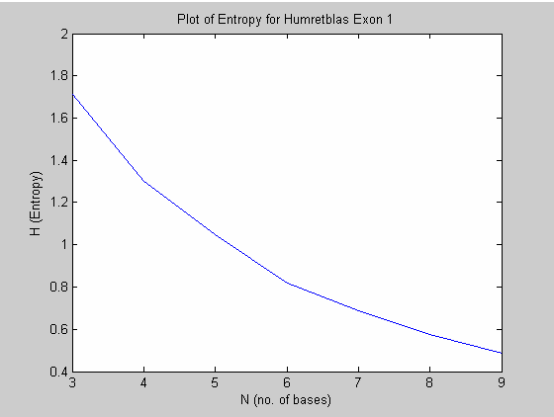
Figure 10. An Intron sequence from Humretblas Genome

The shortness of the average length of exons (114 bp) used from Humretblas as compared to that of the introns (2347 bp) is a fundamental limitation for this analysis. To make a reasonable comparison, we have used a total coding sequence by combining all exons

Sequence	Length (bp)	H3	H4	H5	H6	H7	H8	H9
HUMRETBLAS	180388							
Exon 1	197	1.71	1.30	1.05	0.82	0.6868	0.5731	0.4878
Exon 2	137	1.35	1.13	0.92	0.739	0.59	0.5	0.43
Exon 3	127	1.49	1.22	0.93	0.73	0.60	0.49	0.42
Exon 4	116	1.53	1.21	0.89	0.71	0.57	0.47	0.40
Exon 5	68	1.32	1.02	0.74	0.58	0.45	0.38	0.31
Exon 6	39	1.13	0.79	0.56	0.43	0.33	0.25	0.25
Intron 1	3227	1.94	1.89	1.71	1.48	1.26	1.08	0.95
Intron 2	2687	1.86	1.80	1.65	1.43	1.21	1.04	0.91
Intron 3	2622	1.85	1.79	1.63	1.41	1.20	1.04	0.90
Intron 4	2522	1.88	1.81	1.67	1.42	1.20	1.04	0.90
Intron 5	1936	1.88	1.78	1.59	1.35	1.15	0.98	0.86
Intron 6	1092	1.77	1.67	1.45	1.23	1.03	0.88	0.77
Total coding region	2787	1.87	1.851	1.68	1.44	1.22	1.05	0.92

Figure 11. Higher-order Exon and Intron Entropy Estimates for HUMRETBLAS

Length of Exon = 197 bp



Length of Exon = 137 bp

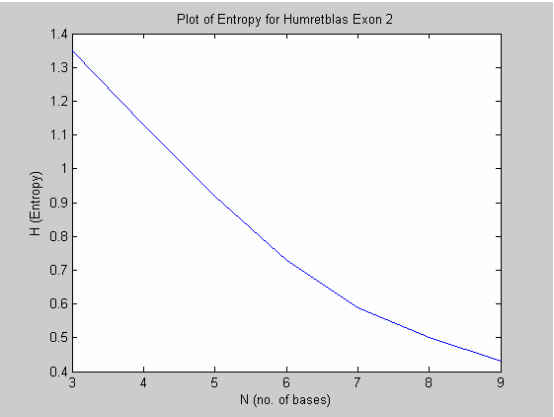
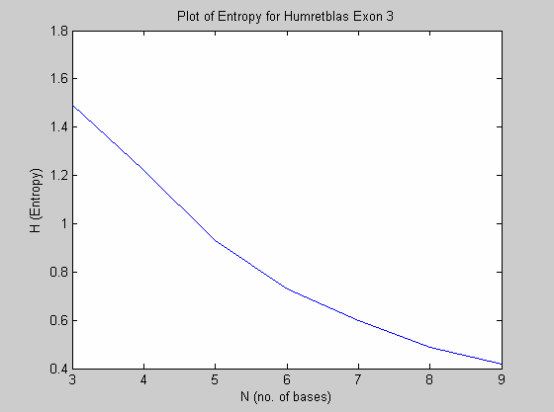
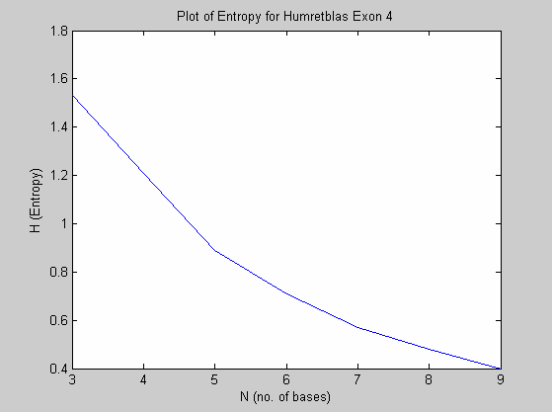


Figure 12. Exon Entropy Plots for Humretblas genome (contd.)

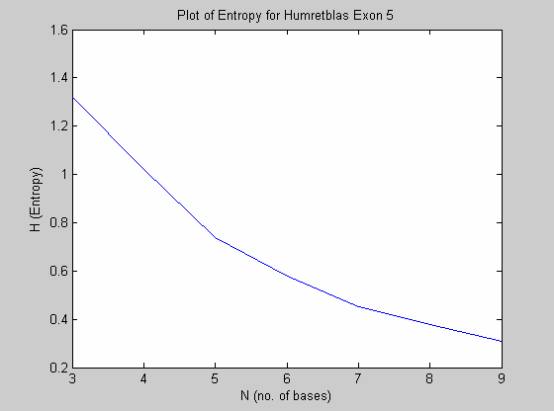
Length of Exon = 127 bp



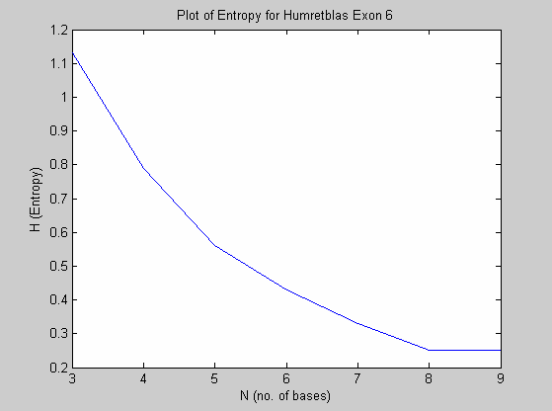
Length of Exon = 116 bp



Length of Exon = 68 bp



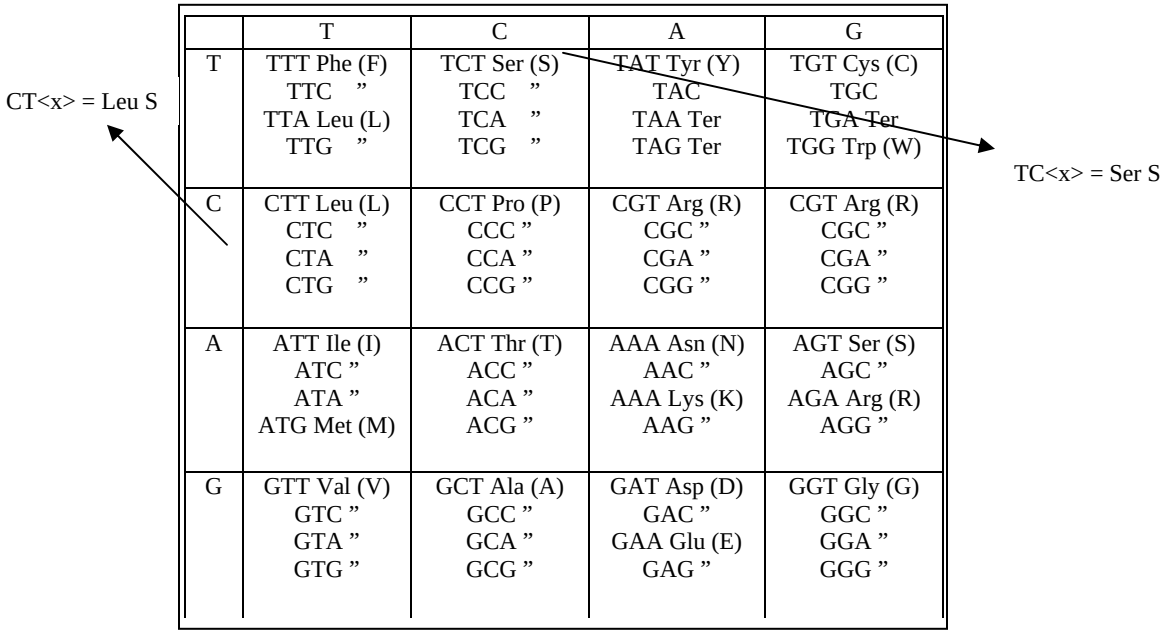
Length of Exon = 39 bp



4.2.1. Entropy Estimate with Nucleotide Pairs

The smallest sub-component of a DNA sequence was a codon based on its significance from the genetic code. The Universal genetic code indicates that every codon in the DNA sequence maps to a corresponding amino acid. This makes us believe that codons are the primary information carrying strings. There are two key motivations of carrying out this part of the work. The fact that the actual coding sequence is a combination of several exons (DNA Translation) makes it interesting to look at sub-strings other than the codons. Another motivation was derived

from an observation of the mapping in the genetic code. As we know, each of the 64 codons map to one of the 20 amino acids that comprise a protein sequence. The Universal genetic code of an organism is presented again along with the observation of codon positions as below.



	T	C	A	G
T	TTT Phe (F) TTC ” TTA Leu (L) TTG ”	TCT Ser (S) TCC ” TCA ” TCG ”	TAT Tyr (Y) TAC TAA Ter TAG Ter	TGT Cys (C) TGC TGA Ter TGG Trp (W)
C	CTT Leu (L) CTC ” CTA ” CTG ”	CCT Pro (P) CCC ” CCA ” CCG ”	CGT Arg (R) CGC ” CGA ” CGG ”	CGT Arg (R) CGC ” CGA ” CGG ”
A	ATT Ile (I) ATC ” ATA ” ATG Met (M)	ACT Thr (T) ACC ” ACA ” ACG ”	AAA Asn (N) AAC ” AAA Lys (K) AAG ”	AGT Ser (S) AGC ” AGA Arg (R) AGG ”
G	GTT Val (V) GTC ” GTA ” GTG ”	GCT Ala (A) GCC ” GCA ” GCG ”	GAT Asp (D) GAC ” GAA Glu (E) GAG ”	GGT Gly (G) GGC ” GGA ” GGG ”

Figure 13. Behavior of nucleotide pairs from Genetic Code

As see from the genetic code table, pairs of nucleotides are sometimes sufficient to encode for an amino acid which makes the third codon position seem redundant. For example, the codon CC<x> codes for Pro (P) irrespective of the base value that <x> assumes. Although there are exceptions to this in other parts of the genetic code, it holds for a majority of pairs. Intrigued by this behavior of pairs of nucleotides, it might be appropriate to also look at the DNA sequence as a sequence of pairs of nucleotides. There are 4^2 = 16 possible pairs of nucleotides for our alphabet {A,T,G,C} and the entropy of pairs was normalized by a factor 2 i.e., H(G) = H(G)’/2 to obtain the entropy per codon and hence achieving entropy plots with the first entropies obtained using pairs.

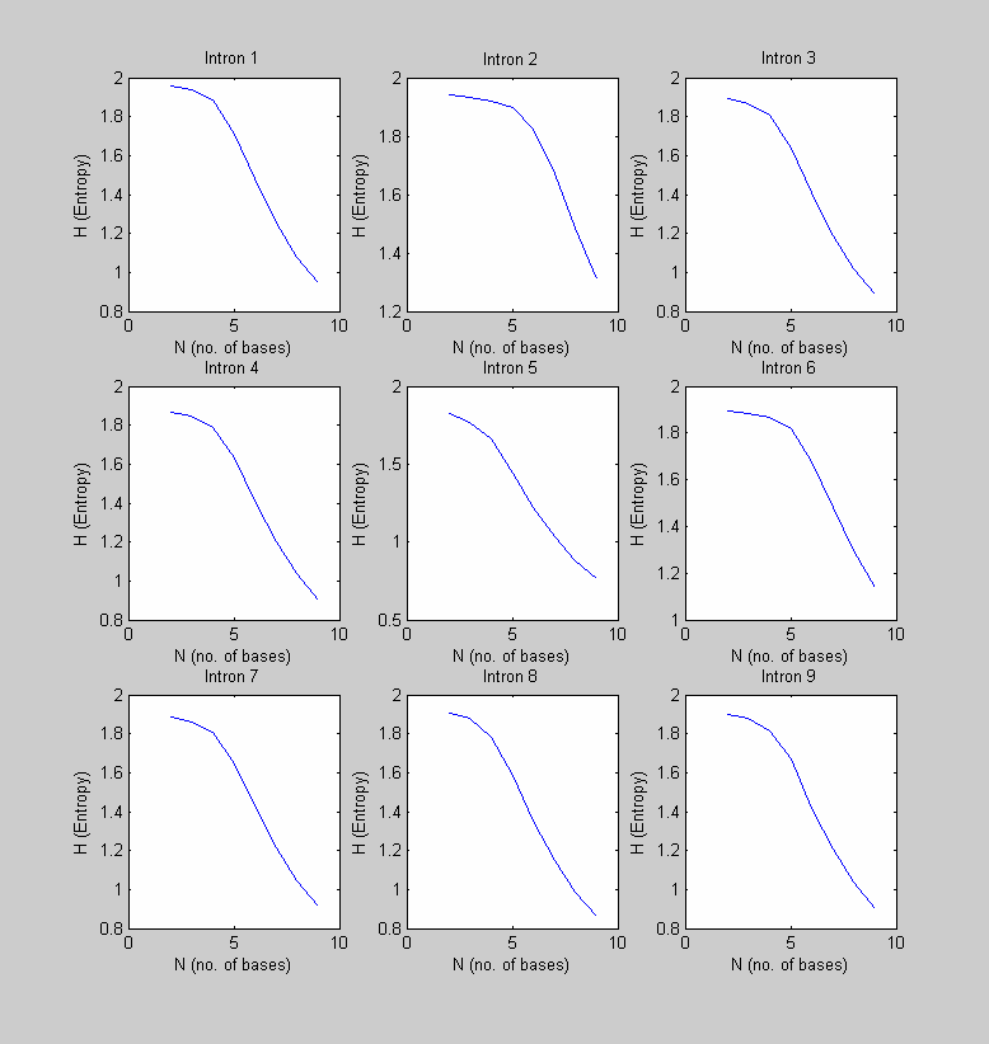


Figure 14. Intron entropy plots for Humretblas including entropies estimated using pairs The entropy values calculated using a sub-string length of 2 are almost equal to the entropy values found using triplets for the sequences tested. The entropy variation plots above include the entropy estimate using base pairs. The entropy due to pairs doesn't seem to affect the rate of entropy convergence seems to be the same for all sequences used. From one perspective, this perhaps further emphasizes a fast rate entropy convergence as a universal property of all intron and exon sequences using our estimator. The finiteness of the exons might have a significance effect on the entropy. This also calls for a technique to take the effect of finiteness into account and hence support our exploration of entropy variation.

4.3. Benchmarking of Entropy Using an Ensemble of Random Sequences

As the length of the sub-string L used in entropy calculation increases so does the set of possible L-tuples. A larger sequence of length exceeding the number of L-tuples is more likely give an accurate estimate. However, due to shortness of the exon sequences, a technique is needed to make a reasonable study of their entropy. Previously attempts have been made to derive the onset of finite sample effects on entropy estimation under the assumption that rank ordered distributions tend to follow Zipf's law [18]. We have discovered in this work that the codon frequencies for genomic sequences do not strictly follow the Zipf's law.

In order to address the problem of finiteness of the sequence, a proportionate change is made to entropy values of the sequence under study.

The same entropy estimation method was applied to random character sequences equal in length to each of the sequences observed and comprised of uniformly distributed characters from the alphabet $\{A,T,G,C\}$. By using simple MATLAB codes, random sequences from the character $\{A,T,G,C\}$ and length equal to the DNA character sequence under question are generated. The entropy values for an ensemble of random sequences of the same length show entropy convergence. This may well be attributed to the finiteness of the sequences.

Since the random sequences are also from the same alphabet, we can use the same entropy algorithm get the entropy values $H(R_i)$ using randomly generate sequences from the alphabet $\{A,T,G,C\}$. From the above results, entropy values decrease with increase in search string length one can attribute this in part to the finiteness of the character sequence. The proportion of the random sequence is calculated as,

$$\Delta_i = \frac{2L}{H(R_i)} \tag{4.6}$$

Using an ensemble of such random sequences, we calculate the average $H(R_i)$. The correction in the entropy value of sequence under test is done by a multiplying the entropy $H(G_i)$ with the corresponding proportion Δ_i to get the corrected entropy $\hat{H}(G_i)$ i.e.,

$$\hat{H}(G_i) = H(G_i) * \Delta_i \tag{4.7}$$

The benchmarked entropies for intron and exon sequences of the Humretblas genome are as illustrated in the table below.

Sequence	Length (bp)	H3	H4	H5	H6	H7	H8	H9
HUMRETBLAS	180388							
Exon 1	197	1.98	1.93	1.99	1.97	2	2	1.97
Exon 2	137	1.67	1.82	1.94	1.93	1.95	1.94	2
Exon 3	127	1.84	2.03	2.02	2	2	1.95	2
Exon 4	116	1.91	2.04	1.98	2	2	2	2
Exon 5	68	1.92	2	2	2	2	2	2
Exon 6	39	1.91	2	2	2	2	2	2
Intron 1	3227	1.95	1.94	1.95	1.99	1.94	1.99	1.93
Intron 2	2687	1.88	1.88	1.92	1.98	1.98	1.99	2
Intron 3	2622	1.81	1.84	1.92	1.97	1.98	1.98	1.99
Intron 4	2522	1.90	1.89	1.97	1.99	2.1	2.01	2
Intron 5	1936	1.90	1.88	1.92	1.96	1.99	1.99	2
Intron 6	1092	1.82	1.85	1.92	1.97	1.98	1.99	2
Total coding region	2787	1.89	1.92	1.94	1.98	1.98	2	2

Figure 15. Benchmarked Entropy Results for Humretblas

Entropy of majority of sequences that were used here show that the entropy at L = 4 is either a peak or equal to the codon entropy and followed by a steady fall in the slope. This implies that a string of 4 characters in a DNA character sequence carries an entropy value higher than that of triplets and all strings of a higher length. It makes an interesting

observation that strings of length four carry greater information in comparison to codons that play a major role in the transcription of mRNA to protein sequences as seen earlier. This result is perhaps a consequence of redundancy in certain codon positions. The role of the genetic code has been under study and it is believed that there may be more unknown information present within DNA sequences that plays a role in the conversion of DNA to proteins.

CHAPTER 5

SIGNIFICANCE OF THE WORK

5.1. About Intron Sequences

Considering the fact that introns are finite character sequences, we have an opportunity of exploring their structure, statistical behavior and patterns in comparison to other sequences. In this chapter, we will present the results in our attempt to study the behavior of intronic sequences. We would like to recall the earlier argument that it is necessary to account for the length of the DNA sequences, in addition, to their complexity, structure and sequence pattern, in order to apply information theoretic concepts for their analysis. It is a well known observation that exons tend to be around 200 characters long while introns can stretch to as many as tens of thousands of characters in length [9]. Although the observation may only be an approximation, it is in conformity with the well known fact that only about 3% of the entire genome actually codes for proteins and the remaining is introns and non-coding DNA. Just as the codons indicate the beginning and end of exons, the introns have mostly occurring start and stop indicators as being GT and AG respectively [9]. However many other locations can resemble such patterns and hence these indicators cannot solely suffice to recognize a splice junction. Researchers have attempted using information theoretic techniques like entropy to differentiate between the two finite character sequences – exons and introns. It has been said that entropy is a useful tool in the analysis of DNA sequences [9]. In order to present a rational point of view on the characteristic of such complex, high capacity information storage sequences as DNA sequences we would need a flawless framework.

5.2. Intronic Entropy Results

With the intent of analyzing the effectiveness of entropy as a measure, let us take a quick look back at how entropy of the sequences was calculated here. The entropy calculation is based on Shannon's definition,

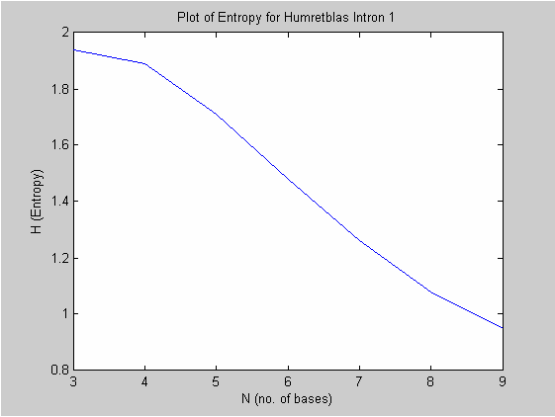
$$H(X) = -\sum_{i} p(x_i) \log p(x_i) \tag{5.1}$$

This can be represented for DNA sequences in terms of the constituent strings g_i as,

$$H(S) = -\sum_{i-1}^{N/L} p(g_i) \log p(g_i) \tag{5.2}$$

where S is the DNA sequence under question which in this case will be either introns or exons and g_i is a search sequence formed of characters from $\{A,T,G,C\}$ of length L and N is the length of the DNA sequence.

Length of Intron = 3227



Length of Intron = 2687

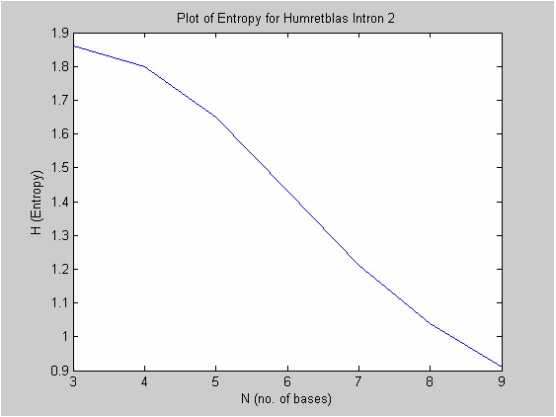
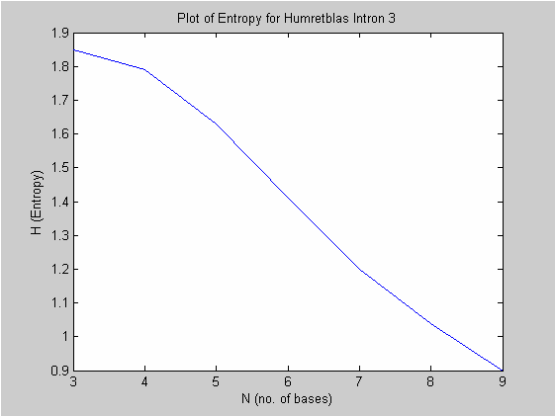
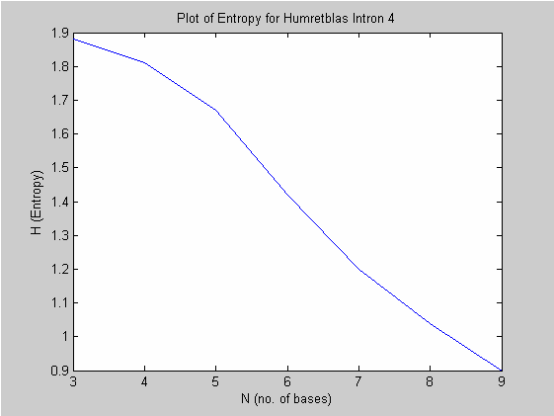


Figure 16. Entropy plots of Intron sequences (contd.)

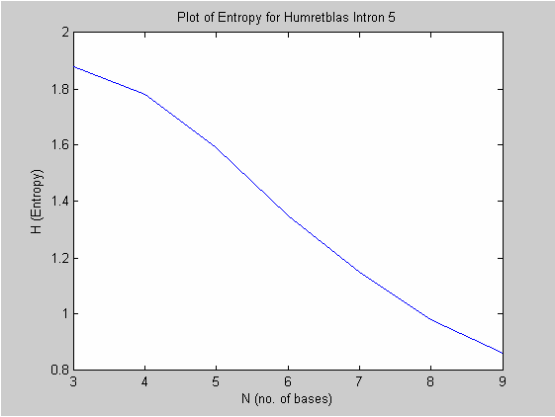
Length of Intron = 2622



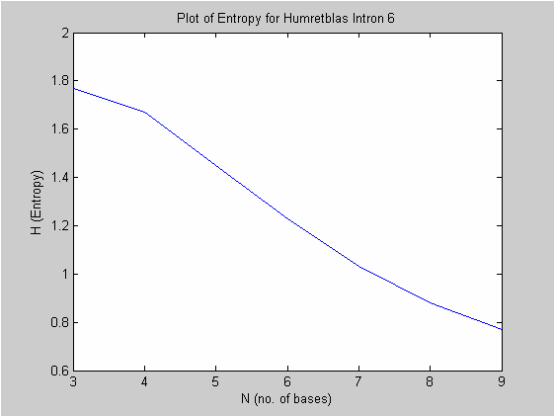
Length of Intron = 2522



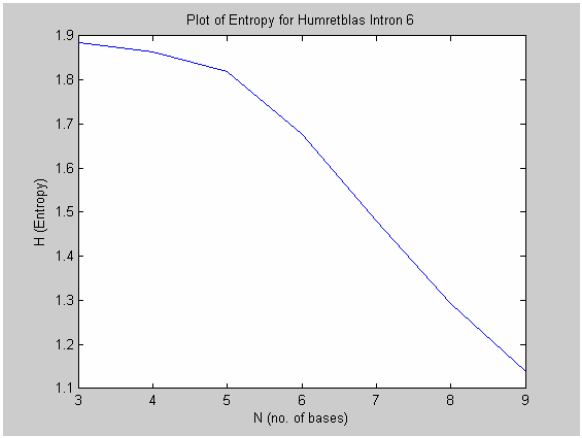
Length of Intron = 1936



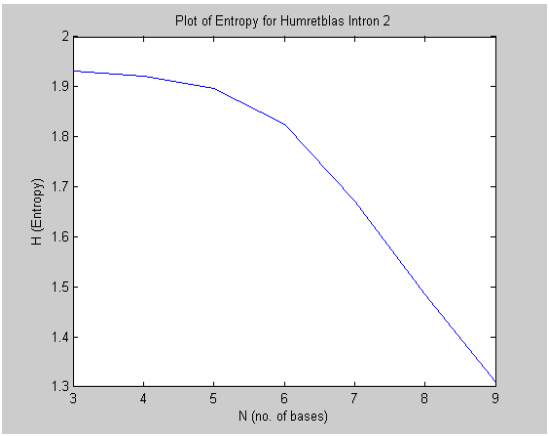
Length of Intron = 1092



Length of sequence = 10986 bases



Length of sequence = 33895 bases



The pattern of entropy convergence in intron sequences seems to match that of the exons in spite of the much higher lengths of introns than the exons. The limited amount of exons is expected due to the presence of only 5% of coding DNA in an organism. Introns have high entropy values as close to 1.97 and 1.98 using codons and in some cases are seen to fall below 1. This might indicate they have some kind of underlying structure. It has been observed earlier that entropy values of the total coding region are very comparable to the intron sequence of equivalent length. This result is likely to be useful in determining similarity between intron and exon sequences.

CHAPTER 6

EXONS VS INTRONS – A DISTANCE MEASURE

6.1. Background

We shall take another look at the algorithm that was used in obtaining the entropy plots for a genome. It follows intuitively that repetition of short strings within the sequence under question will affect the entropy of the sequence. The entropy was calculated using the Shannon's entropy formula by changing the length L of the short strings constituting the sequence from 3 to 9.

By observing the entropy plots for sequences of different lengths, we can infer that the convergence of entropy with increase in length of search strings L depends perhaps partially on the length of the actual exon/intron sequence N. In other words, a study of the pattern of convergence of the entropy plot by varying the sequence length N and the search string length L, can to an extent predict whether the sequence is very long or short. It is an established fact that the majority of exons are of a length $N \leq 200$ bases long [9]. The difference in the convergence time of entropy at $L = 9$ for the exon and intron entropy plots may partly be due to the lengths of the sequence. Hence there is limitation to the rationale used to differentiate introns from exons; we will further explore possibilities of calculating a differentiating measure between exons and introns. Utilizing the fact that most exons are short in length i.e. $\approx 200bp$ (base pairs), we employ the above test to predict the length of the sequence. Since the short length of exons often holds true, we will treat this as our first test towards finding if the sequence is an exon or an intron. We will have justified this later if the analysis of the structure of the sequence can support our choice and give us a computational proof.

6.2. Kullback-Leibler and Bhattacharya Measures

To find a measure of difference between sequences, it is important to first list down the conditions that it must satisfy. One such issue is that the difference measure has to differentiate between two sequences having equal length but that may carry different structure, patterns and information capacity, and another to use a similar kind of measure irrespective of the length of the two sequences. One of the commonly used methods in computer algorithms is the distance between sequences. The proportion of differences between exons and introns can be simply calculated as,

$$D = k/n \tag{6.1}$$

where k is the number of nucleotides and

n is the length of the sequence.

We will first introduce a framework for finding the distance between two sequences using their probability distributions. Considering p_0 and p_1 as two probability densities, the Kullback-Leibler distance [19] may be defined as,

$$D(p_0 \,||\, p_1) = \int p_1(x) \log \frac{p_1(x)}{p_o(x)} \tag{6.2}$$

where the log (.) is calculated to the base 2

This distance is one example of Ali-Silvey class of information-theoretic distance measures [cited in 19] which take a general form,

$$d\,(p_0, p_1) = f(\in_0 [c(\Lambda(X))]) \tag{6.3}$$

where $\Lambda(.)$ represents the ratio of probabilities $p_1(.)$

$p_0(.)$, c(.) is convex,

$\in_0 [.]$ is the expected value with respect to the probability distribution p_0

and f(.) is a non-decreasing function.

A well known characteristic of the K-L distance is,

$$D(p_0 \,||\, p_1) \neq D(p_1 \,||\, p_0) \tag{6.4}$$

Another important distance measure that can be derived from this class is the Chernoff distance [cited in 12] which is defined as

$$C(p_0,p_1) = \max_{0 \le t \le 1} -\log \mu(t) \text{ , where } \mu(t) = \int [p_0(x)]^{1-t} [p_1(x)]^t \, dx \tag{6.5}$$

A special case then has also been defined as "Bhattacharya distance" [19][20][15] which is simply $B(p_0,p_1) = -\log \mu(\frac{1}{2})$. These distances have been touted as important measures in information processing due to three special characteristics [19]. At the discovery of Shannon's information theory in 1948, a distance measure named "divergence" became popular in several applications [15]. Divergence, sometimes referred to as J-Divergence was first introduced by Jeffreys [15][21][22] and is defined in terms of the Likelihood ratio $L(x) = \frac{p_1(x)}{p_0(x)}$ as,

$$J = E_1[\ln L(x)] - E_2[\ln L(x)] \tag{6.6}$$

where $E_i[\ln L(x)] = \int [\ln L(x)] p_i(x) dx$, i = 1,2.

In other words, divergence is the difference of the Kullback-Leibler numbers with the difference that they K-L numbers are asymmetric and J-Divergence is symmetric [15]. With our focus being on how to apply the concept of distance measures to DNA character sequences, we shall leave further discussion of the characteristics of these measures as a reference. It is clear from the above discussion that we need to know the probability distributions of the sequence in order to measure its distance from another sequence. The

following section presents a framework of how we can possibly find the probability distribution for a DNA character sequence and the technique used to find distribution(s) that can closely model the word frequencies of DNA sequences.

6.3. Exon and Intron Word Frequencies and Probability Distribution

In this section, we present a word frequency analysis of genetic sequences. We foresee this to help us in finding a generic probability distribution that can model the probabilities used earlier in entropy calculations

Lexical statistics is an area that deals with the study of word frequency distributions. To carry out this study, we first need to define all the distinct words and then their instances for our character sequences. As we know the 64 codons form the 'types' or distinct words and every occurrence of a codon in the sequence under question may be defined as a 'token'. A mapping of the types to tokens helps us build the data that we can use for the analysis. The data in such a frequency list can be re-organized in two useful ways namely "rank/frequency profiles" and as "frequency spectra" [23]. These are two ways of representing mainly the same information and on of them can be derived from the other if needed. The frequency profile and spectra plots of exons and introns using codons as tokens were observed to be skewed. The frequency profile is a plot of the codon frequencies against the ranks and a frequency spectrum is a graph that shows the number of codons that occur with a certain frequency. The intron of length 3227 shows has 23 different codon frequencies where as the one with length 33895 bp has a total of 57 different codon frequencies. As evident from the plots below, there is a steady fall in the slope of codon frequencies of introns and the frequencies tends to fall less rapidly only for higher ranks or less frequent codons. The plots for exon sequences show a linear decrease in codon frequencies over the range of ranks that

were observed. The frequency spectra of introns illustrate that a majority of codons occur with the same frequency. This observation is strict for the lower frequencies for an Intron of length 3227 shown by a peak in the frequency spectrum and is uniform over a set of frequencies for an Intron length of 33985. The exon sequences used have the largest number of codons occurring a single time. This is observed by a peak at the extreme left of the frequency spectra of exons followed by a steady fall towards higher frequencies.

Analysis of Humretblas Genome

N= 3227 N=33895

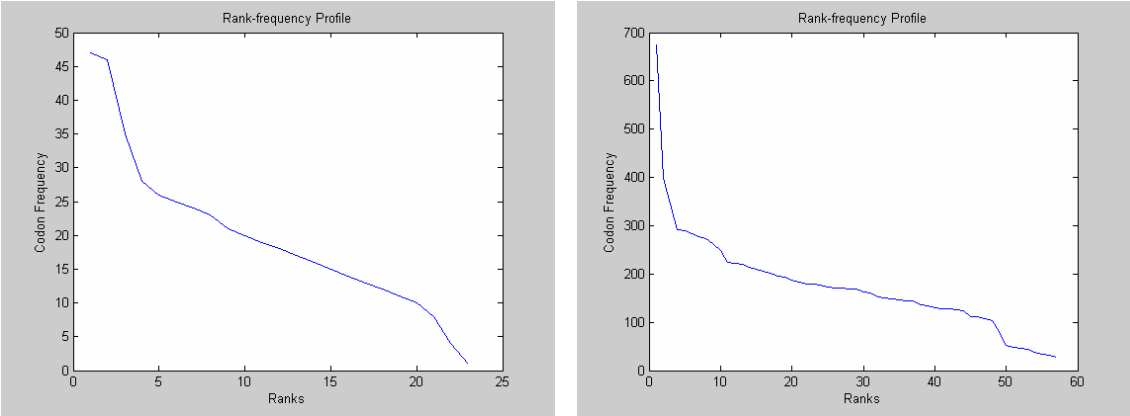


Figure 17. Rank/Frequency Profiles for Introns

N = 137 N = 127

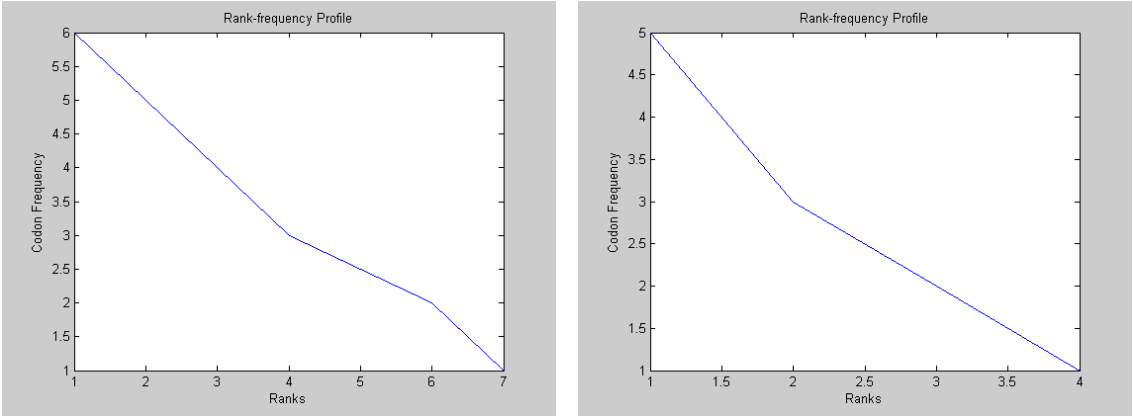


Figure 18. Rank/Frequency Profiles for Exons

N= 3227 N=33895

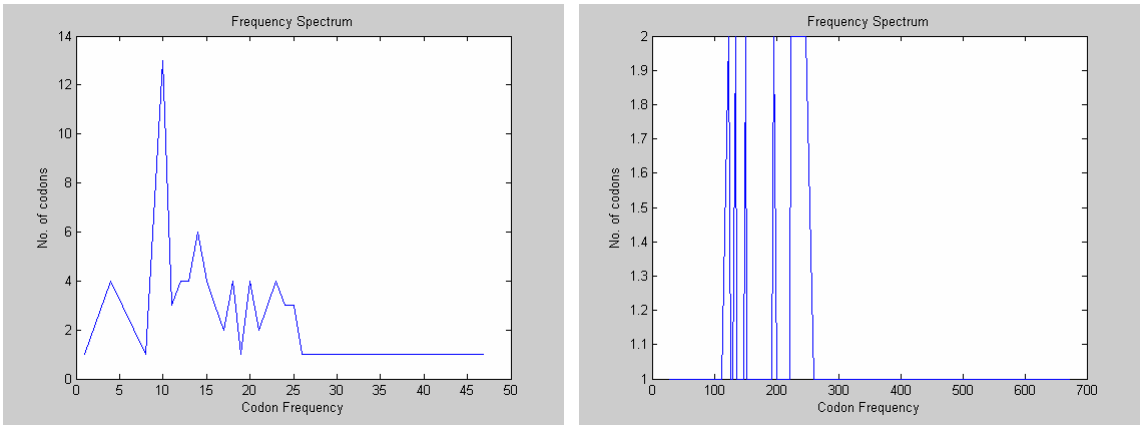


Figure 19. Frequency spectra for Introns

N = 137 N = 127

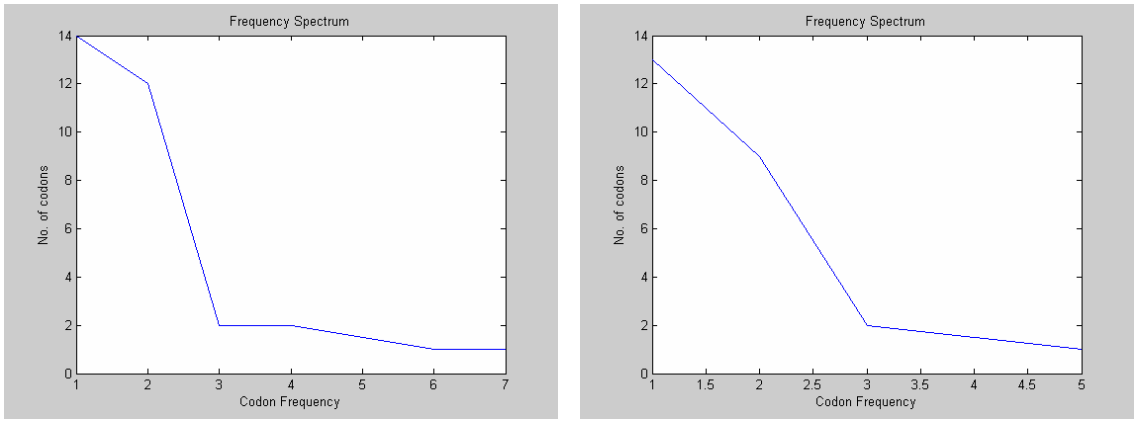


Figure 20. Frequency spectra for Exons

A famous linguist named Zipf who is well known as the father of lexical statistics had first studied the structure of word frequency distributions and found a law known as *Zipf's law* that sequences in varied fields prove as valid. A mathematical representation of Zipf's law is,:

$$F(w) = C/r(w)^a \qquad (6.7)$$

where F(w) and r(w) stand for frequency and rank of word w, respectively [23].

C and a are constants that take values corresponding to the data being modeled.

Zipf's law is one of the key examples of distributions that evolved as a result of the study of word frequency distributions. According to Zipf's law, the probability of occurrence of a word varies inversely as a power of the rank of the word, in a list of frequencies with the most frequent placed at the top. The steadily decreasing frequency spectrum of the Humretblas sequences we used seems to drift from the simple form of the Zipf's distribution. Based on repetitive iterations of trial and error, we were able to establish a closed formula for the probabilities of occurrence of each codon as:

$$F(w) = k.\ C/r(w)^{a} + n \qquad (6.8)$$

The frequency spectra and rank profiles of exons and introns illustrate structure in these sequences. Perhaps these character sequences carry an underlying distribution based on the occurrence of codons in them. With this motivation, we seek to model the intron/exon frequency spectra. As seen above, Zipf's law is a unique effort in defining the distribution of a sequence based on its frequencies. In order to find a reasonable model, we have worked with the following assumptions:

(i) Value of C is set to the probability of the codon that ranks 1 in the frequency list.

(ii) The constant k takes on integer values, usually of the order of a single digit 1,2,3, so on.

(iii)The value 'n' indicates the error or difference in the actual values to the generated probability.

With these set of assumptions, we tested the same intron and exon sequences from Humretblas genome for which the frequency spectra and rank profile were calculated. The introns give an error range of 0.00x with max n = 0.015 and the exons give the error of the

order of 0.0x with max n = 0.048. We can hence write the probability distributions for sequences in the Humretblas genome as,

$$P_1(w) = k.\ C/r(w)^a + 0.005 \text{ for introns} \quad (6.9)$$

$$P_o(w) = k.\ C/r(w)^a + 0.02 \text{ for exons} \quad (6.10)$$

6.4. Distance Measure for DNA Sequences

Using Equation 6.2, the Kullback-Liebler distance between the probability distributions may be defined as,

$$D(p_0 \parallel p_1) = \int p_1(x) \log \frac{p_1(x)}{p_o(x)} \quad (6.11)$$

where p_0 and p_1 as two probability densities.

For discrete probability distributions p = {p_0, p_1 …pn} and q = {q_0, q_1 …qn}, this is defined as,

$$D(p \parallel q) = \sum_i p_i \log_2 (p_i / q_i) \ , \ i = 1,2,\ldots n \quad (6.12)$$

In order to apply this distance measure to calculate the distance between exons and introns, we may use the discrete probabilities of components of these sequences. For example, let us consider codons as the basic components of the sequences. Then the discrete probability distribution is the set of 64 probabilities. Consider an exon and intron sequence of length 78 bases taken from the Humretblas genome. At the first instance, let p = {p_0, p_1, ……p_{64}} be the probability distribution for an intron sequence and q = {q_0, q_1, ……q_{64}} that of exon sequence (both equal in length), the value of K-L distance in this case may be represented as D_{IG} and swapping the distributions p and q for exons and introns, this value is say, D_{GI}. For the above example, values of D_{IG} = 0.5995 and D_{GI} = 0.2823 were obtained. It is understandable that these values satisfy Eq. (6.3) according to the definition of K-L

distance. This distance measure is sometimes referred to as 'Relative Entropy'. If the two distributions are not too dissimilar, the difference between D_{GI} and D_{IG} is small and is in turn related to the sample size. For the distance values obtained above,

$$\Delta = D_{IG} - D_{GI} = 0.3172 \tag{6.13}$$

This value shall be close to zero for two completely similar sequences. Contrarily, the value obtained here isn't large enough to tell that the exons and introns are completely different in structure. Further efforts in this direction might be useful to determine a precise measure of similarity between exons and introns.

CHAPTER 7

CONCLUSION

The entropy of a number of DNA coding and non-coding sequences collected from different genomes was estimated using a frequency based entropy estimation algorithm for finite sequences. The exon and intron entropy plots both converge in value with increase in length in a similar fashion. For bench-marking, the same entropy estimation method was applied to random character sequences equal in length to each of the sequences tested; the bench-marking sequence comprised of uniformly distributed characters from the alphabet $\{A,T,G,C\}$. In order to deal with the problem of finiteness of the sequence and to make a reasonable entropy comparison between intron and exon sequences that come in different lengths, a correction factor was obtained for every exon/intron sequence using an ensemble of random sequences of the same length. Entropy plots of some of the sequences show a peak at L = 4 followed by a steady fall in the slope. This implies that the bases in a string of 4 characters in a DNA sequence carry average information higher than that for triplets and all strings of a higher length. This is one of the significant findings of this thesis, indicating that least correlation occurs across adjacent codons and that there exist stronger correlation beyond. The relationship across codons was captured most clearly when the normalization using the benchmark sequence results was done, and it was found that both exons and introns have such long-range correlations. This finding is likely to be useful in further understanding of the nature of the genetic code.

The similarity in the entropy of exons and introns suggests that the introns are likely to be playing some hitherto unknown but useful role. Since entropy is directly related to the information content, the similarity of entropy patterns indicates that introns have hidden

information. It is not known if this information is useful in repair of exons, or for some independent function. We hope that future work would seek correlations of this information with cellular function.

A distance measure was applied to compare exon and intron information content and it was again found that they are structurally quite similar. An approximate distribution, modified from Zipf's law, was found for sub-strings of exons and introns for code word lengths of 3. As future work, further investigations into the sub-string distribution may be made.

REFERENCES

[1] Werner Ebeling, Thorsten Poschel, and Karl-Friedrich Albrecht, Entropy, Transinformation and Word Distribution of Information-Carrying Sequences, arXiv: cond-mat/0204045 v1, 2 Apr 2002

[2] David Loewenstern, Peter N. Yianilos and Department of Computer Science Princeton University, Princeton, New Jersey 08544, Significantly Lower Entropy Estimates for Natural DNA Sequences, Journal of Computational Biology, volume 6, number 1, 1997

[3] Martin Farach, Michiel Noordewier, Serap Savari, Larry Shepp, Abraham Wyner and Jacob Ziv, On the Entropy of DNA: Algorithms and Measurements based on Memory and Rapid Convergence, Nov 1 1994

[4] Mark White, Rafiki Genetics; http://www.codefun.com/Genetic_what1.htm

[5] S. Kak, Quantum information and entropy. International Journal of Theoretical Physics, vol 46, 2007; arXiv: quant-ph/0605096

[6] S. Kak, Information complexity of quantum gates. Int. J. Theoretical Physics, vol. 45, pp. 933-941, 2006; arXiv: quant-ph/0506013

[7] S. Kak, Artificial and biological intelligence. ACM Ubiquity, vol 6, pp. 1-20, 2005; arXiv: cs.AI/0601052

[8] J. Balakrishnan, A symmetry scheme for amino acid codons, 2003. arXiv: physics/0308091

[9] Martin Farach, Michiel Noordewier, Serap Savari, Larry Shepp, Abraham Wyner and Jacob Ziv, On the Entropy of DNA: Algorithms and Measurements based on Memory and Rapid Convergence, Nov 1 1994

[10] David P. Feldman and James P. Crutchfield, Structural Information in Two- Dimensional Patterns: Entropy Convergence and Excess Entropy, arxiv.org/abs/cond-mat/0212078, December 5, 2002

[11] Matthew J. Berryman, Mathematic principles underlying genetic structures, arXiv.q-bio.GN/0607039 v1 22 Jul 2006

[12] Jeremy J. Ramsden, "Bioinformatics: An Introduction", Kluwer Academic Publishers Dordrecht/Boston/London 2004

[13] H. Chernoff. Large-sample theory: Parametric case. Ann. Math. Stat., 27:1-22, 1956

[14] S. Kak, Classification of random binary sequences using Walsh-Fourier analysis. Proceedings of Applications of Walsh Functions. 74-77. Washington, D.C., 1971

[15] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. On Comm. Tech., COM-15(1): 52-60, 1967

[16] Wentian Li, The Study of Correlation Structures of DNA Sequences: A Critical Review, arXiv.adap-org/9704003 v1 9 Apr 1997

[17] Michael Haag, Autocorrelation of Random sequences; http://cnx.org/content/m10676/latest/

[18] T. Dudok de Wit, When do finite sample effects significantly affect entropy estimates?, European Physical Journal B 11, 513-516 ,1999

[19] Don H. Johnson and Sinan Sinanovic, Symmetrizing the Kullback-Leibler Distance, IEEE Trans. On Information Theory March 18, 2001

[20] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc., 35:99-109, 1943

[21] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. Roy. Soc. A., vol. 186, pp. 453-461, 1946

[22] H. Jeffreys, Theory of Probability. Oxford University Press, 1948

[23] Marco Baroni, 39 Distributions in text, June 30 2006 http://sslmit.unibo.it/~baroni/publications/hsk_39_dist_rev2.pdf

[24] All Hariri, Bruce Weber, and John Olmstead. On the validity of Shannon-information calculations for molecular biological sequence, Journal of Theoretical Biology, 147:235-254, 1990

[25] H. Herzel, A. O Schmitt,. & Ebeling, W. Finite Sample effects in Sequence Analysis, Chaos, Solitons & Fractals 4, 97-113. 1994

VITA

Riyazuddin Mohammed was born in Hyderabad, Andhra Pradesh, India on March 2, 1982. He earned his primary and secondary education in St.George's Grammar School, Hyderabad, Andhra Pradesh, India. He was awarded the Indian Certificate of Secondary Education (ICSE) on completing his 10^{th} grade.

He received his Bachelor of Engineering degree in Electronics and Communication Engineering from Osmania University, Hyderabad, Andhra Pradesh, India in Spring 2003. He worked with Pentagram Research Company during his undergraduate study to carry out a project titled "Genomic Signal Processing". This project was submitted to fulfill part of the requirement for award of his Bachelors degree. A paper based on the findings of this project was accepted for poster presentation at the International Signal Processing Conference, March 31-April 3, 2003 held in Dallas, Texas.

He came to U.S. to pursue his Master's degree. He then joined the graduate program with the Department of Electrical and Computer Engineering at LSU, Baton Rouge in Spring 2004. His major area of study was communications/systems. He is a candidate for the degree of Master of Science in Electrical Engineering to be awarded at the commencement of Fall, 2006.