Evaluation of Information Theory in Analyzing DNA Sequences

Abdullah Nazib, A.H.M Amimul Ahsan, and Dr Mahbubur Rahman

Abstract—In this paper, we explore the application of information theory in DNA sequence analysis. The main focus of this research work is to observe the performance of information theory concepts in differentiating coding and non-coding sequences. Mutual Information (MI) and Block Entropy measure are used to differentiate coding and non-coding sequences. The simulated performance measure illustrates that mutual information has better classification ability than entropy measure.

Index Terms—Mutual Information, Genome, Exon, Intron, Entropy, DNA.

I. INTRODUCTION

DNA is the source of information that carries genetic information of all living organism. The information contained in a DNA plays vital role in controlling organism specific characteristics. So a genome sequence is considered as a biological data warehouse of an organism. This biological warehouse contains large number of repetitive patterns that are not regarded as biologically significant and they are termed as intron/non-coding sequences. On the other hand, only 2-3% area of a genome is responsible for protein synthesis, mRNA generation etc and these sequences are termed as Exon/Coding sequences. The basic problem of analyzing DNA is the recognition of coding and non-coding sequences. A number of methods have been proposed to distinguish coding and non-coding sequences from a large genome, but most of them require a prior data set of specific organism. This training set dependence limits the applicability of coding measure techniques [1], [2], [3], [4]. In search of a method without requiring training on specific data set, information theory is considered in many researches. Some of them considered entropy as differentiating criteria and others has proposed mutual information. This paper focuses on the differentiating capability of two information theoretic concepts. In order to measure the performance between Entropy and Mutual Information in separating coding and non-coding regions, this paper considers difference of averages of Block entropy and mutual information as performance criteria.

The rest of the paper is organized as follows. Section II describes basics of Entropy and block entropy with mathematical details. Mutual information is verified in section III with experimental details. The mathematical details and computational findings of distance measurements based on the previous two sections are explained in section IV and finally section V concludes the paper.

II. ENTROPY ANALYSIS

Unlike the entropy of thermodynamics, the entropy in information theory represents average information content in a given probability distribution. Since a DNA sequence consists of four characters \{a, c, g, t\}, the probability of occurrence of each base is not same. The probability of a particular type of nucleotide is estimated by relative frequency measure as follows:

\[
P_{\text{base}} = \frac{N_{\text{base}}}{N_{\text{total}}} \quad (1)
\]

Where \( P_{\text{base}} \) denotes probability of a base, \( N_{\text{base}} \) is the frequency of that base and \( N_{\text{total}} \) total number of nucleotides in the given sequence. Then the entropy of each base is given by the Shanon entropy formula as

\[
H(S) = -\sum_{i} P_{i} \log_{2} \left( \frac{1}{P_{i}} \right) \quad (2)
\]

Since DNA bases are not statistically independent, direct application of Eq.2 cannot provide satisfactory result to distinguish coding and non-coding sequences.

In considering statistical dependence between DNA bases, block entropy measurement of varying block length is studied [5], [6]. The block entropy of a DNA sequence of length \( N \) has \( 'n' \) substring of specified length \( L \) then the relation between DNA length and its substring length is given by \( L=N/n \). Let’s consider the probability of the specified length substring is \( p(S) \), where \( p(S)=n_{S}/N \). Then the entropy measure of the genetic sequence is

\[
H(S)= -\sum_{i} P_{i} \log_{2} P_{i} \quad (3)
\]
As the size of alphabet increase, the block entropy increases accordingly by following equation

\[
H(S) = \frac{H(S)}{L}
\]

Where \( L = \text{block length} \).

Using this block entropy measure, we find that the average amount of information decreases as the size increases.

III. MUTUAL INFORMATION ANALYSIS

The concept of mutual information comes from entropy. From the block entropy definition we know that the block entropy of string of length 1 is given by \( H_1 \) where the subscript notation defines the string length. The conditional entropy \( h_i \) can be defined as

\[
h_i = H_2(k) - H_1
\]  

Where \( H_1 \) and \( H_2 \) can be defined as

\[
h_1 = -\sum p(i) \log_2(p(i))
\]

And

\[
h_2(k) = -\sum p_{ij} \log_2(p_{ij}(k))
\]

Where \( p_i \) is the probability of a symbol \( A_i \) and \( p_{ij} \) is the joint probability of finding two symbol \( A_i \) and \( A_j \). The difference between two symbols gives us the information about second symbol by getting to know the first one. Then the mutual information between two symbols apart from distance \( k \) is given by

\[
I(k) = H_1 - h_1
\]

Now applying Eq.4, Eq.5 and Eq.6 we can get mutual information as

\[
I(k) = 2H_1 - H_2(k)
\]

\[
I(k) = \sum p_{ij}(k) \log_2 \left( \frac{p_{ij}(k)}{p_i p_j} \right)
\]

This is the generalized equation for mutual information.

The application of derived Eq.8 to analyze difference between coding and non-coding sequence, we have to measure mutual information between two base pairs at a distance \( k \). The joint probability distribution is calculated from a 12 positional nucleotide probabilities \( p_i \) which is the

12 positional probability vector and it can be represented by \([A_1, A_2, A_3, C_1, C_2, C_3, G_1, G_2, G_3, T_1, T_2, T_3] \)

For the simplicity of calculation we determine 4-by-3 matrix as follows:

Using this positional matrix the joint probability is measured by following mathematical formula for an arbitrary reading frame:

\[
R_{ij}G0 = \frac{1}{2} (p_{ij}p_{ij} + p_{ij}p_{ij} + p_{ij}p_{ij}) \text{ for } k = 3, 6, 9, ...
\]

Again the positional matrix can be used to generate marginal probabilities of each type of base as

\[
p = (p^{(1)} + p^{(2)} + p^{(3)})/3.
\]

The joint probability matrix obtained from Eq.9 in symbolic form is shown as:

From this joint probability matrix, we can determine the joint probability between two base pairs in a given distance \( k \). It becomes clear from above mentioned derivation that mutual information between two bases would be changed if the value of \( k \) is changed.

IV. EXPERIMENTAL RESULT

The implementation of two information theoretic model using ‘MatLab’ simulation shows how they differentiate coding and coding sequences. The data set of “Fickett and Tung” of Human genome is used in this experiment. Table I, shows the result of block entropy analysis over human intron and exon sequences. As shown in Table I, the block entropy of a sequence decreases as the length increases. According to Table I, the difference between intron and exon entropies are also calculated in the last row from which it becomes clear that as the length of the block is increased the differences between coding and non-coding sequences increases as well.

<table>
<thead>
<tr>
<th>Type</th>
<th>H(3)</th>
<th>H(4)</th>
<th>H(5)</th>
<th>H(6)</th>
<th>H(7)</th>
<th>H(8)</th>
<th>H(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron</td>
<td>1.3612</td>
<td>1.3564</td>
<td>1.35253</td>
<td>1.3483</td>
<td>1.34144</td>
<td>1.32826</td>
<td>1.29917</td>
</tr>
<tr>
<td>Exon</td>
<td>1.3636</td>
<td>1.3585</td>
<td>1.35423</td>
<td>1.3499</td>
<td>1.34496</td>
<td>1.3382</td>
<td>1.32805</td>
</tr>
</tbody>
</table>

url: www.ijcit.org or www.ijcit.uap.bd.edu
In the second phase of the experiment, the mutual information is studied. To analyze mutual information, 10 sequences of same length are selected from each type and then the mutual information at distance k=3 to 9 are calculated. This calculated result then averaged out which shown in Table II. In Table II the difference between coding and non-coding sequence clearly and more precisely classifies two types of sequence since the difference values are higher in comparison to block entropy differences.

**Table II. Obtained Result of Mutual Information**

<table>
<thead>
<tr>
<th>Type</th>
<th>MI(3)</th>
<th>MI(4)</th>
<th>MI(5)</th>
<th>MI(6)</th>
<th>MI(7)</th>
<th>MI(8)</th>
<th>MI(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron</td>
<td>3.23E-04</td>
<td>1.17E-04</td>
<td>7.97E-05</td>
<td>5.23E-05</td>
<td>1.11E-04</td>
<td>1.33E-04</td>
<td>3.35E-04</td>
</tr>
<tr>
<td>Exon</td>
<td>0.0003</td>
<td>1.23E-04</td>
<td>8.72E-05</td>
<td>5.38E-05</td>
<td>0.00011</td>
<td>1.44E-04</td>
<td>3.64E-04</td>
</tr>
<tr>
<td>Difference</td>
<td>0.2817</td>
<td>0.06</td>
<td>0.075</td>
<td>0.0149</td>
<td>0.0754</td>
<td>0.11</td>
<td>0.29</td>
</tr>
</tbody>
</table>

The plot graph of mutual information of intron vs exon in Fig.1 shows the MI between base pairs in both sequences. In exon sequences, a periodic peak value is obtained after 3-base but in intron sequences the amplitude of peaks are slightly greater than zero. In Fig.1 exon MI are shown by red lines and intron MIs are shown by blue lines.

**V. Conclusion**

This paper presents the implication of information theoretic models to differentiate coding and non-coding regions of a DNA. A raw DNA sequence can be analyzed with these theorems to determine coding and non-coding boundaries. Due to overlap and very short length of non-coding sequences in prokaryotic genome, information theory has lacking in this regard. This limitation can be solved using probabilistic models such as Hidden Markov Model (HMM) or Bayesian Network.

**References**