Entropy based analysis of genetic information

Paramita Saha\textsuperscript{a} and Bimal Kumar Sarkar\textsuperscript{b}

\textsuperscript{a}Department of Physics, School of Science, Adamas University, Kolkata, India, paramitasaha.ac89@gmail.com
\textsuperscript{b}Department of Physics, School of Science, Adamas University, Kolkata, India, bks@physics.org.in (corresponding author)

Abstract. In this work, we have described the analysis of digitized sequences of genetic information by means of the notions of entropy. The occurrence of a particular pattern in the genetic sequence is paid a special attention. The occurrence of genetic word is expressed in a density manner. The occurrence frequency of the q-gram genetic word of interest is determined with the help of finite impulse response (FIR) type filter along the sequence. It is in turn, used for the determination of horizontal correlations, i.e., correlations between the word along the sequence. We use the probability distribution of the genetic word occurrence as the input for the calculation of entropy in the sequence. The sequence entropy is further used for principal component analysis (PCA) to determine the similarity / dissimilarity between the biological sequences. The technique is verified by using 48 HEV genotypes.

Key word: Sequences, genetic information, FIR filter, entropy, PCA, HEV.

1. Introduction

Human DNA contains about 3 billion nucleotides and conclusion of genome sequencing of several model organisms has further thriven the genomic databanks. For good understanding, the biological information in a DNA sequence is an overwhelming task. The development of fast, efficient, and cost-effective computational techniques for the same is a big challenge. Generous scenarios for DNA sequence suggest that the number of sequences that could have existed throughout global history is far less than the number of possible sequences \cite{1,2}. For a small DNA sequence that is 100 bp long, there are $4^{100}$ possible sequences. This difference suggests that only very small regions of "sequence space" have been explored. This implies that the requirement for generating a functional sequence is not stringent. DNA's genetic code can be represented as an alphabetic sequence composed of the four letters A, C, G, and T, which represent the four types of nucleotides--adenylic, cytidylic, guanylic, and thymidylic acid, of which DNA is composed. Now that these sequences have been identified for many genes and are available in computer-readable form, scientists can analyze these data and search for patterns in an attempt to learn more about the regulatory functions of the gene. One area of study is that of the frequency of occurrence of specific nucleotide subsequences (e.g., ACAC) within part or all of a nucleotide sequence.

A large number of gene finding algorithms use probabilistic models like Hidden Markov Models (HMM) \cite{3}. They depend on gene models or on an a priori knowledge of coding statistics of the genome of the organism under consideration. They utilize coding measures like codon usage measures, oligonucleotide counts, aminoacid usage, codon preference, hexamer usage etc. Papers \cite{4,5} contain reviews and assessments. On the contrary, digital processing techniques of gene finding are model independent. They utilize the general features of coding rather than the organism dependent measures. Furthermore, digital input can be considered for probabilistic inference on the biological sequencing exploration. It is worthy to be said that the probabilistic information is sound enough to
find the stability features in the sequence. For example, if some pattern appears in a periodic way or sub-sequences occur repeatedly, the sequence is more ordered. On the other hand, when all symbols occur at equal probability, the entropy becomes maximal [6]. If a pattern is less probable, there always a non-equilibrium state will be formed and the state will be more order. If the system is open, it will show a decrease in entropy [7]. It is available in literature that entropy and information ideas find an extensive application in DNA inquiries [8-10].

This paper focuses on the digital signal processing technique of gene finding and suggests a method by which coding statistics can be incorporated into it. Taking the digital input, we introduce the method to calculate the entropy of a well-defined biological sequence, irrespective of any predefined sequence alignment. We have calculated the probability of the genetic word in a single-stranded DNA sequence. The genetic word is made from four alphabet A, C, G, T with different length, viz, \{AA, ----TT\}, \{AAA, ----, TTT\} etc. having length 2,3, ---etc.in the sequence. We call a q-nucleotide word as q-mer or q-grams. As there are only four letters in the DNA alphabet, viz., \{A, C, G, T\} the number of all combinations of q-grams in a DNA sequence is 4q. Given a sequence, we count the frequency of occurrence (or frequency) of each q-mer using an overlapping sliding window of width W and sliding one [11].

In the present work, we have calculated the Shannon entropy at each position of the sequence possessing genetic word probability distribution and in turn it has been used for principal component analysis (PCA). We applied PCA algorithm using the HEV (Hepatitis E virus) genome sequences. The HEV is a small, non-enveloped virus positive-sense, single-stranded RNA virus and belongs to Hepevirus genus under the family of Hepeviridea [12-15], whose genome is formed by a single RNA strand with positive polarity and a length of approximately 7.5 Kb. It has features typical of a eukaryotic mRNA, including a 7-methylguanine cap at its 5’ end and a poly(A) stretch at the 3’ end.

By PCA algorithm, 48 HEV genomes were distinctly clustered into four genotypes identical to the traditional classification.

2. Methodology

Our method is based on the observation through a sliding “counter “of width W over DNA sequence. A certain number of q-grams called as bins are set in the counter. As there are only four letters in the DNA alphabet, viz., \{A, C, G, T\} the number of all combinations of q-grams in a DNA sequence is 4q.

**Definition 1. q-gram of Sequence**

Given a sequence ‘seq’, when a window of length q slides over the characters of ‘seq’, its q-grams are formed. For a sequence 'seq, there are \(|seq|-(q-1)\) q-grams.

The number of all possible q-grams or called as “bin” is \(4^q\). Bins can be arranged in lexicographic order, and \(b_i\) is used to denote the \(i^{th}\) bin in this order. All the possible bins are denoted as:

\[
B_q = \{b_1, b_2, \ldots, b_{4^q}\}
\]

**Example 1.** One-gram bins are \(B_1 = \{A, C, G, T\}\), consisting 4 bins. Two-gram bins are \(B_2 = \{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}\), consisting 16 bins.
**Definition 2. Bin Signature**

For a sequence, the q-gram bin signature, $S_q$ is a mapping with the bin $b_j \in B_q$ where $i^{th}$ bit in $S_q$ is corresponding to the presence or absence of $b_j$. For a sequence ‘seq’, there are $|seq| - (|b_j| - 1)$ bits in $S_q$.

**Example 2.** Consider a sequence, $S = “AACTCG”$. Its two-grams ($q=2$) signature in the sequence is $S_2 = [0 \ 1 \ 0 \ 0 \ 0 \ 0]$.

**Definition 3. Filter**

A sequence $x[n]$ is filtered through mapping of the sequence into output sequence $y[n]$ via a weighted window $b$ by means of the convolution summation as

$$y[n] = \sum_{i=0}^{k} b_i \cdot x[n-i]$$

(1)

$b$ is independent of $x[n]$ and $y[n]$, where $n$ is the time index. $y[n]$ is the response of the filter to input signal $x[n]$. The filter is finite impulse response (FIR) digital filter. The term digital filter arises because it operates on discrete-time signals. Finite impulse response arises because the filter output is computed as a weighted, finite term sum, of past and present (Figure 1).

![Figure 1: Block diagram of finite impulse response (FIR) digital filter](image)

**Example 3:** Weighted filter output of $S_A$, with the weighted window $\beta = [0.2 \ 0.1 \ 0.3 \ 0.4]$ is as follows: $S_A = [1 \ 1 \ 0 \ 0 \ 0 \ 0]$.

$$y_A[n] = \sum_{k=0}^{3} \beta_k S_A[n-k]$$

With $\beta_0 = 0.2$, $\beta_1 = 0.1$, $\beta_2 = 0.3$, $\beta_3 = 0.4$.

$$y_A[n] = \beta_0 S_A[n] + \beta_1 S_A[n-1] + \beta_2 S_A[n-2] + \beta_3 S_A[n-3]$$

$y_A = [0.2 \ 0.3 \ 0.4 \ 0.7 \ 0.4 \ 0]$; Similarly for other nucleotide viz., C, G, T, the output is obtained as, $y_C = [0.2 \ 0.0 \ 0.2 \ 0.1 \ 0.5 \ 0.5]$; $y_G = [0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.0 \ 0.2]$; $y_T = [0.0 \ 0.0 \ 0.0 \ 0.2 \ 0.1 \ 0.3]$.
For nucleotide density calculation, evenly distributed window of unit value is considered. As explained, the output of the convolution summation represents the nucleotide density along the sequence. The detail algorithms for bin construction, bin signature, filter operation is displayed in Table-1, Table-2, and Table-3 respectively.

**Table 1: Bin Construction**

| Input: q - length of bin | Output: set of bins |
|-------------------------|---------------------|
| \( B_q = \{ b_1, b_2, \ldots, b_{4^q} \} \) |

```plaintext
1: 0 \rightarrow \text{bincount};
2: 4^q \rightarrow \text{n};
3: \text{cell}(1,\text{n}) \rightarrow \text{bin};
4: \text{for} \text{first} = 1:4 \text{ do}
5: \text{for qth} = 1:4 \text{ do}
6: \text{convert integer to nucleotide character (first \ldots qth)} \rightarrow \text{binq};
7: \text{bincount} = \text{bincount} + 1;
8: \text{binq} \rightarrow \text{bin[bincount]};
9: \text{end}
10: \text{end}
11: \text{bin} \rightarrow B_q
```

**Table 2: Bin Signature**

| Input: Sequence (seq), bin (b) | Output: Bin Signature |
|-------------------------------|------------------------|
|                               | \( B_{eq} = \{ b_1, b_2, \ldots, b_{4^q} \} \) |

```plaintext
1: m \rightarrow \text{length (seq)};
2: nbin \rightarrow \text{length (b)};
3: \text{for} i = 1...m - (nbin - 1) \text{ do}
4: \text{if seq (i : i+nbin-1) = b then}
5: \text{signature (i) = 1}
6: \text{else}
7: \text{signature (i) = 0}
8: \text{end}
9: \text{signature} \rightarrow \text{Bin Signature}
```

**Algorithm 3 Filter**

| Input: BinSignature, window | Output: filter |
|-----------------------------|---------------|
| \( w \rightarrow \text{length (window)}; \) |

```plaintext
1: w \rightarrow \text{length (window)};
2: \text{window} = 1/w*\text{array of ones(1,w)};
3: 0 \rightarrow \text{sum}
4: \text{for} i = 1...\text{length (window)} \text{ do}
5: \text{make array of zeros with length of i -1} \rightarrow \text{zeros}
6: \text{sum} = \text{sum} + \text{window (i) * array[zeros BinSignature(1:(length(BinSignature)-(i-1)))]}
7: \text{end}
8: \text{filter} \rightarrow \text{sum}
```
3. Sequence analysis

The filter output is taken as a density distribution for DNA sequences. The density distribution is based on q-gram word density, which in turn is considered for the determination of Shannon Entropy as

\[ y_i = -\sum_{j=1}^{q} p_{ij} \log p_{ij} \]  

where \( p_{ij} \) is the probability of appearance of the \( j \)th genetic letter at \( i \)th position in the genetic sequence. Further we want to find a similarity/dissimilarity measure between two entropy distributions \( \rho_i = (y_{i1}, y_{i2}, \ldots, y_{in}) \) and \( \rho_j = (y_{j1}, y_{j2}, \ldots, y_{jn}) \). We construct the data matrix \( D \) comprising elements \( [\rho_1, \rho_2, \ldots, \rho_m]' \), where \( m \) is the number of sequences. Principal Component Analyses (PCA) is used to estimate scores between density distributions such that it reduces multidimensional data sets to lower dimensions with the consistent original data matrix \([16]\).

We determine the dissimilarity between two sequences from the scores in the first three principal components by computing the Euclidean distance between pairs of density distributions in the \( m \)-by-\( n \) data matrix \( D \). Rows of \( D \) correspond to sequence (observations) and columns correspond to position index in the sequence (variables). Thus, Euclidean distance \( X \) is a row vector of length \( m(m-1)/2 \), corresponding to pairs of observations in \( D \). The distances are arranged in the order \((2, 1), (3, 1), \ldots, (m, 1), (3, 2), \ldots, (m, 2), \ldots, (m, m-1)\). \( X \) is used as a dissimilarity matrix in clustering or multidimensional scaling. An unweighted pair group method with arithmetic mean (UPGMA) is employed on PC scores for the construction of a phylogenetic tree \([17]\). UPGMA uses a local objective function to construct a rooted bifurcating tree.

4. Results and discussions

We have chosen some HEV sequences (HEV) from various countries to test our method. The details of the organism are presented in Table-4. Based on FIR filtering, firstly the nucleotide density distribution is generated. We have calculated the density distribution for one-, two-, three-, gram

![Figure-2: Entropy profile of five HEV sequence. Entropy is calculated based on single nucleotide distribution. Sequences are: B: Bur-82; C: Bur-86; D: C1; E: C2; F: C3 (see Table-4)](image-url)
nucleotide for different species. Secondly we have calculated entropy distributions \( \rho_i = (y_{i1}, y_{i2}, \ldots, y_{in}) \) and \( \rho_j = (y_{j1}, y_{j2}, \ldots, y_{jn}) \). Figure-2 displays the spatial variation of the entropy along the HEV

| Sequence No. | Strain  | Accession No. | Place            | Cluster |
|-------------|---------|---------------|------------------|---------|
| 1           | Bur-82  | M73218        | Burma            | I       |
| 2           | Bur-86  | D10330        | Burma            | I       |
| 3           | C1      | D11092        | China (Xinjiang) | I       |
| 4           | C2      | L25595        | China (Xinjiang) | I       |
| 5           | C3      | L08816        | China (Xinjiang) | I       |
| 6           | C4      | D11093        | China (Xinjiang) | I       |
| 7           | China Hebei | M94177   | China (Hebei)    | I       |
| 8           | FHF     | X98292        | India            | I       |
| 9           | I3      | AF076239      | India (Hyderabad) | I       |
| 10          | Mad-93  | X99441        | India (Madras)   | I       |
| 11          | Morocco | AY230202      | Morocco          | I       |
| 12          | Np1     | AF051830      | Nepal (Kathmandu)| I       |
| 13          | P2      | AF185822      | Pakistan (Abbottabad) | I |
| 14          | Sar-55  | M80581        | Pakistan (Sargodha)| I       |
| 15          | T3      | AY204877      | Chad             | I       |
| 16          | Yam-67  | AF459438      | India (Yamuna Nagar)| I       |
| 17          | M1      | M74506        | Mexico (Telixtac) | II      |
| 18          | HE-JA10 | AB089824      | Japan (Tokyo)    | III     |
| 19          | Arkell  | AY115488      | Canada (Ontario) | III     |
| 20          | JBOAR1-Hyo04 | AB189070   | Japan (Hyogo)    | III     |
| 21          | JDEER-Hyo03L | AB189071   | Japan (Hyogo)    | III     |
| 22          | JTI-KAN | AB091394      | Japan (Kanagawa) | III     |
| 23          | JKN-Sap | AB074918      | Japan (Sapporo)  | III     |
| 24          | JMO-Hyo03L | AB189072   | Japan (Hyogo)    | III     |
| 25          | JMY-HAW | AB074920      | Japan (Sapporo)  | III     |
| 26          | JRA1    | AP003430      | Japan (Tokyo)    | III     |
| 27          | JSO-Hyo03L | AB189073   | Japan (Tokyo)    | III     |
| 28          | JTH-Hyo03L | AB189074   | Japan (Tokyo)    | III     |
| 29          | JYO-Hyo03L | AB189075   | Japan (Tokyo)    | III     |
| 30          | Kyrgyz  | AF455784      | Kyrgyzstan       | III     |
| 31          | swJ570  | AB073912      | Japan (Tochigi)  | III     |
| 32          | swUS1   | AF082843      | USA              | III     |
| 33          | US1     | AF060668      | USA (Minnesota)  | III     |
| 34          | US2     | AF060669      | USA (Tennessee)  | III     |
| 35          | HE-JA1  | AB097812      | Japan (Hokkaido) | IV      |
| 36          | CCC220  | AB108537      | China (Changchun)| IV      |
| 37          | HE-JI4  | AB080575      | Japan (Tochigi)  | IV      |
| 38          | HE-JK4  | AB099347      | Japan (Tochigi)  | IV      |
| 39          | JAK-Sai | AB074915      | Japan (Saitama)  | IV      |
| 40          | JKK-Sap | AB074917      | Japan (Sapporo)  | IV      |
| 41          | JSM-Sap95 | AB161717    | Japan (Hokkaido) | IV      |
| 42          | JSN-Sap-FH | AB091395    | Japan (Hokkaido) | IV      |
| 43          | JSN-Sap-FH02C | AB200239 | Japan (Hokkaido) | IV      |
| 44          | JTS-Sap02 | AB161718    | Japan (Hokkaido) | IV      |
| 45          | JYV-Sap02 | AB161719    | Japan (Hokkaido) | IV      |
| 46          | swCH25  | AYS94199      | China (Uighur)   | IV      |
| 47          | swJ13-1 | AB097811      | Japan (Hokkaido) | IV      |
| 48          | T1      | AJ272108      | China (Beijing)  | IV      |
sequence for first 5 species. The spatial variation of the entropy along the HEV sequence for 6 to 10 species is shown in Figure-3.

![Figure-3: Entropy profile of five HEV sequence. Entropy is calculated based on single nucleotide distribution. Sequences are: B: C4; C: China Hebei; D: FHF; E: I3; F: Mad-93 (see Table-4)](image)

Similarly, the variation of entropy with position for all other sequences are calculated, which are not displayed in this paper. Rather it is more comprehensive to show the entropy variation for all sequences (total 48) in a single panel. It has been shown in Figure-4.

The present work intends to assess the variability and complexity at each nucleotide site with the calculation of entropy for each position using the Shannon entropy formula, Eq. (2). The low entropy region around 5500 position is common to all 48 sequences. Entropy \( Y_i \) is an important parameter for the understanding of sequential stability. \( Y_i \) becomes maximal when all symbols occur at equal probability. On the other hand, \( Y_i \) becomes the least if one symbol occurs at probability 1 and in that case the other symbols will be forbidden. It means that lower the value of entropy the site is more stable without much complexity. Under this assumption, the zone around the site 5500 is most stable for all strain/species. It may find a good structural relationship between the regions of low entropy and the secondary structure of proteins which include \( \alpha \)-helix, \( \beta \)-sheets and loops regions. Similarly, low entropy zones are found around the site 500, 2300, 4950 for all strains (Figure 4).
Figure 4: Entropy profile of 48 HEV sequences. Entropy is calculated based on single nucleotide distribution. Sequences are represented as number starting from 1 to 48. (see Table. 4)

First 15 strains show the stability with lower entropy around 3090 site position. But this behaviour is not exhibited in case of rest of the strains. If one can go through strain number 35 onwards, as a whole, it is noticed that the entropy is increasing or in turn the complexity is more. It is an indication of evolutionary development among the HEV strains. Based on site entropy we prepared the dissimilarity matrix for 48 HEV sequences (Figure 5).

Figure 5: Dissimilarity Matrix for 48 HEV sequences
The dissimilarity matrix demonstrates the existence of 4 different clusters. One can see that the HEV sequences in a cluster shows very less dissimilarity among themselves. In other way to mean that the sequences have much similarity residing in a cluster [18-19]. The HEV sequences appearing in cluster I are typically from the countries Burma, China, India, Nepal and Pakistan. Remarkably the mentioned countries are neighbor to each other. It is worthy to be noted that the sequence no 17 (M1) appears very singly in unique cluster II. This sequence is quite different from any of the remaining 47 strains. M1 strain was found from the county Mexico. Strain 18 to 34 in cluster III are collected from Japan, Canada, Kyrgyzstan and USA [20-21]. Cluster IV comprising of strains 35 to 48 are mostly found from Japan and China.

Taking entropy distributions, we construct the data matrix X comprising elements $[p_1, p_2, ..., p_m]$, where m is the number of sequences. Principal Component Analyses (PCA) is used to estimate scores between entropy distributions such that it reduces multidimensional data sets to lower dimensions with the consistent to original data matrix. Figure 6 displays the scatter plot of first three PC values for 48 strains. The scatter plot of PC values demonstrates the cluster formation in different HEV strains. Cluster IV is less distributed but Cluster I and cluster III is dispersive in nature. Cluster II is independent with the others.

Figure 6: Scatter plot PC values for 48 HEV sequences. Cluster I is in blue colored elliptically encircled. Cluster II is single strain shown at top. Cluster III is in green colored elliptically encircled. Cluster IV is in yellow colored rectangle encircled

5. Conclusions
We propose a novel method based on Entropy Analysis to determine the similarity / dissimilarity between the biological sequences. The probability distribution of the occurrence of genetic word is taken as the input for the calculation of entropy. The sequence entropy is used for principal component analysis (PCA) to determine the similarity / dissimilarity between the HEV sequences, which belong to Hepevirus genus with sequence length of approximately 7.5 Kb. 48 HEV genomes were distinctly clustered into four genotypes. The genotype taken from the countries Burma, China, India, Nepal and Pakistan very notably shows neighbor relationship. The single genotype of sequence no 17 (M1) is very exclusive in nature differentiated from others. Other two clusters maintain a neighbor relationship. Entropy calculation demonstrate that there is a good structural relationship between the regions of low entropy and the secondary structure of proteins which include $\alpha$-helix, $\beta$-sheets and
loops regions. Thus, entropy calculation is very potential tool for the interpretation of the stability in a biological sequence.

6. References

[1] M. Eigen, “Steps Toward Life”, Oxford University Press, Oxford, England, 1992.  
[2] S. A. Kauffman, “The Origins of Order, Oxford University Press, New York, 1993.  
[3] Stigler, J.; Ziegler, F.; Gieseke, A.; Gebhardt, J. C. M.; Rief, M. Science. 334 (2011): 512–516.  
[4] S. Vinga, J. Almeida, 19 (2003) 513–23, 2003.  
[5] M. Domazet-Lošo, B. Haubold, 27 (2011) 1466–1472.  
[6] Armin O. Schmitt and Hanspeter Herzel, J. theor. Biol., 1888 (1997) 369-377.  
[7] L.P. Csernai, S.F. Spinnangr, S. Velle, Physica A 473 (2017) 363–376.  
[8] L. Gatlin, J. Theor. Biol. 10 (1966) 281.  
[9] G.W. Rowe, J. Theor. Biol. 101 (1983) 151.  
[10] D.J. Lipman, J. Maizel, Nucl. Acids Res. 10 (1982) 2723.  
[11] N. Jafarzadeh, A. Iranmanesh, Commun Math Comput Chem, 68 (2012) 611–620.  
[12] P. Ward, P. Müller, A. Letellier, S. Quessy, C. Simard, Y. L. Trottier, A. Houde, J. Brassard, Can J Vet Res, 72 (2008) 27–31.  
[13] G. G. Schlauder, I. K. Mushahwar, J Med Virol, 65 (2001) 282–292.  
[14] Y. Wang, R. Ling, J. C. Erker, H. Zhang, H. Li, S. Desai, I.K. Mushahwar, T.J. Harrison, J Gen Virol, 80 (1999) 169–177.  
[15] S. A. Tsarev, L. N. Binn, P. J. Gomatos, R. R. Arthur, M. K. Monier, H. van Cuyck-Gandre, C. F. Longer, B. L. Innis, J Med Virol, 57 (1999) 68–74.  
[16] Novembre, J., Stephens, M., Nat Genet 40 (2008) 646–649.  
[17] JF Yu, JH Wang, X. Sun, Commun Math Comput Chem, 63 (2010) 493–512.  
[18] L. Lu, C. Li, C.H. Hagedorn, Rev. Med. Virol, 16 (2006) 5–36.  
[19] S.U. Emerson, R. H. Purcell, Rev Med Virol, 3 (2003) 145–154.  
[20] X. J. Meng, R. H. Purcell, P. G. Halbur, J. R. Lehman, D. M. Webb, T. S. Tsareva, J. S. Haynes, B. J. Thacker, S. U. Emerson, Proc. Natl. Acad. Sci. USA, 94 (1997) 9860–9865.  
[21] P. Ward, P. Müller, A. Letellier, S. Quessy, C. Simard, Y. L. Trottier, A. Houde, J. Brassard, Can J Vet Res, 72 (2008) 27–31.

Acknowledgement

The two authors (P.S. and B.K.S.) are thankful to the Department of Physics, Admas University for the support. The authors are also thankful to the NCBI Gene Bank for sequence data usage.