Analysis of DNA Sequence
Information and Complexity

Sihai Tang\textsuperscript{1}, Rajan Alex\textsuperscript{2}, Yi Zhou\textsuperscript{3}, Shijun Tang\textsuperscript{2}

\textsuperscript{1}Department of Computer Science
The University of Texas at Austin
Austin, TX USA 78712

\textsuperscript{2}American Molecular Laboratories
Vernon Hills, IL USA 60061

\textsuperscript{3}School of Engineering, Computer Science and Mathematics
West Texas A&M University
Canyon, TX USA 79016

Abstract --- Lempel-Ziv (LZ) complexity of a DNA sequence can be used in modern structural analysis of complete genomes. The low complexity is corresponding to the strong inequality in nucleotide content (e.g. tandem or dispersed repeats or palindrome-hairpin structures, etc). Shannon’s information entropy is simply the “amount of information” in a source of data, which is used to study the amount of information in a sequence (e.g. DNA). In this paper, we utilize two different methods to analyze the same DNA sequence. The special structure patterns of the studied DNA would have a great significance for the clinic diagnosis and treatment of some diseases such as the Alzheimer’s disease.

Keywords --- LZ Complexity, Shannon’s Information Entropy, DNA

1. INTRODUCTION

Many diseases are caused by a genetic mutation, or permanent change in one or more specific genes. If a person inherits from a parent a genetic mutation that causes a certain disease, then sadly he or she will most likely get the disease. For example, Alzheimer’s disease is an inherited genetic disorder.

Identifying genetic variants can help researchers find the most effective ways to treat Alzheimer’s patients or even prevent the disease. This approach is called precision medicine in which genetic variants may vary with individual variability in genes, environment, and lifestyle for each person. So, analysis of genomic sequences is very important to the diagnosis and treatment of some diseases.

The general approach of estimating the complexity of symbolic sequences (texts) was suggested by A. N. Kolmogorov [1]. Kolmogorov complexity is the length of the shortest code generating a given sequence. Kolmogorov complexity is also used for DNA analysis [2][3][4]. In this paper, we use two different approaches to analyze the genomic sequences. One is Lempel-Ziv (LZ) complexity [5] which is used to evaluate structure of a DNA sequence with textual complexity. The other one is Shannon information entropy which shows the probabilities T, C, G, A in the certain window. Also, the maximum, minimum and [maximum - minimum] values of LZ complexities of sequences will be obtained and discussed in this paper.

2. METHODOLOGY

In this section, we display the proposed methods for the complexity and information analysis of a DNA sequence. The method was executed in Matlab (Version R2017a The Mathworks Inc.).

2.1 Data

The related DNA data (SNCA) in this paper is downloaded from the nucleotide database www.ncbi.nlm.nih.gov. SNCA (Homo sapiens synuclein alpha) may serve to integrate presynaptic signaling and membrane trafficking. Defects in SNCA(Homo sapiens synuclein alpha) have been implicated in the pathogenesis of Parkinson disease. SNCA peptides are a major component of amyloid plaques in the brains of patients with Alzheimer's disease[6].

2.2 Kolmogorov Complexity

A DNA sequence is composed of four nucleotides A, C, G, T, where the letters represent the four nucleotide bases: adenine, guanine, cytosine, and thymine. Comparison of DNA primary sequences should be considered not only the string structures but also their chemical properties.

DNA sequences have many unique characteristics, such as heterogeneity (not random) and the long-range correlation, etc. The complexity of DNA regions is very important to
modern structural analysis of complete genomes since the low complexity may be preconditioned by strong inequality in nucleotide content (biased composition)[7], by tandem or dispersed repeats or by palindrome-hairpin structures, as well as by a combination of all these factors[7].

Kolmogorov complexity is a branch of information theory that deals with the complexity contained in a single object or string. It is different from Shannon’s way of measuring entropy.

The Kolmogorov complexity of any string \( x \) is defined as:

\[
K(x) := \min \{ \ell(p) \mid U(p) = x \}
\]

(1)

where \( p \) is the computer program; the length of the shortest program \( \ell(p) \) which produces the output string \( x \), \( U(p) = x \) implies that computer “U” with input “p” produces the output string “x”. That is, we can use the shortest length of computer program (i.e., the number of the characters in computer program) to describe the complexity of a string.

2.3 Lempel-Ziv complexity

In contrast to the Kolmogorov complexity, Lempel-Ziv (LZ) complexity is easily computable and is also a universal depiction of sequence complexity[7][8]. Our approach may be stated as the following:

[1] Assume that S and Q represent two strings, respectively. SQ is the concatenation of S and Q, while SQ_d=SQ(deleted last character in the sequence ). Let V(SQP) be all subset of SQ_d. At the initial state, \( c(n) = 1 \) where \( n \) is the length of the sequence, \( S=s_1 \) and \( Q=s_2 \), So \( SQ_d=s_1 \).

[2] Generally, \( S=s_1s_2…s_n \) and \( Q=s_{n+1} \), then, \( SQ_d=s_1s_2…s_n \); If \( Q \in V(SQ_d) \), then Q is a subsequence of \( SQ_d \), not a new sequence.

[3] Renew Q to be \( Q=S_{r+1}S_{r+2} \) and judge if \( Q \in V(SQ_d) \)

[4] Repeat steps(1-3) until \( Q \in V(SQ_d) \), So \( Q=s_1s_2…s_{n+1} \) is not the subsequence of \( SQ_d=s_1s_2…s_{n+1} \), thus, \( c(n)+1 \)

[5] Update \( S=s_1s_2…s_{n+1}S_{r+2}…s_{n+1} \) and \( Q=s_{r+1} \)

Repeat above steps until Q takes the final character.

\[
b(n) = \lim_{n \to \infty} \frac{c(n)}{\log_2(n)}
\]

(2)

where \( b(n) \) is the upper bound of \( c(n) \)

LZ complexity can be expressed as \( C(n) = \frac{c(n)}{b(n)} \)

(3)

Thus, LZ complexity of a random sequence is 1; LZ complexity of an order sequence is 0.

2.4 Information Entropy

The information entropy was provided by Ludwig Boltzmann when analyzing the statistical behaviour of system’s microscopic components in physics. The entropy was introduced by Claude Shannon to study the amount of information in a transmitted message in information theory. The estimating complexity is evaluated via Shannon Information Entropy:

\[
C_{entropy} = -\sum_{i=1}^{n} \left( \frac{n_i}{N} \right) \log_2 \left( \frac{n_i}{N} \right)
\]

(4)

where \( N \) is the window size, \( n_i \) is the number of symbols in a window and \( K \) is the alphabet size.

3. EXPERIMENTAL RESULTS AND ANALYSIS

Through applying the above methods and the corresponding MATLAB program, LZ complexity and Shannon Information Entropy can be obtained when taking the certain size of a sliding window on a DNA sequence. In order to compare LZ complexity value with Shannon information entropy, an offset value (\( Max_{Entropy} - Max_{LZ_C} \)) is added in the LZ complexity value.

In Fig. 1, the tendencies of LZ complexity values (green) seem similar to those of Shannon Information Entropy values when the sliding window size is 20bp. This is because the small window size limits the different structure patterns of DNA sequences.

![Figure 1. LZ Complexity and Shannon Information Entropy in size of sliding windows 20bp for genomic sequences. x-axis is sequence positions; y-axis is complexity value in the window 20bp.](image-url)

We are able to find detailed information in Fig. 2 which is part of Fig. 1 (position from 0–500 bp). There are some places with disconnection on the red line. That means, there are no entropy values at the points (see red circle). However, there are LZ complexity values corresponding to the special points. This implies that the structure pattern is made of one same bp and/or two/three same bps.
Figure 2. LZ Complexity and Shannon Information Entropy in size of sliding windows 20bp for genomic sequences. x-axis is sequence positions; y-axis is complexity value in the window 20bp. The red circle indicates the special region where unique patterns are made of one same bp and/or two/three same bps.

Figure 3. LZ Complexity and Shannon Information Entropy in size of sliding windows 150bp for genomic sequences. x-axis is sequence positions; y-axis is complexity value in the window 150bp. Except a few regions (e.g. maximum points), the tendencies of LZ complexity values (green) are different with those of Shannon information entropy values (red) when the sliding window size is 150bp in Fig. 3.

From the above results, we can find that [a] the Distance = |MAX-MIN| in LZ complexity calculation is obviously smaller than Shannon Information Entropy since the DNA of humans contain much information, but some sections just contain less information that may be compressed; that [b] information entropy exists the similar tendency of Distance = |MAX-MIN| with LZ complexity when the size of sliding window is small; that [c] the calculation value (whether LZ complexity or information entropy) is closely related with the window size; that [d] compared to the value of information entropy, with the increase of window size, the value of LZ complexity is becoming smaller, and the tendency of LZ complexity is completely different with that of Shannon information entropy.

4. CONCLUSION

In this paper, we have calculated LZ complexity and Shannon Information Entropy of a DNA sequence within different sliding window sizes (20 bp, 150bp and 260 bp). The LZ complexity reflects the various structural patterns inside the DNA sequence. Shannon Information Entropy just gives the probabilities of C, T, G, A in the selected windows. The minimum values for LZ complexity contains a direct tandem repeat. The maximum values for LZ complexity contains multiple different patterns. The MIN and MAX for LZ complexity are especially useful for real clinic value, which corresponds to the special structure and pattern of the DNA sequence. The special structure, pattern and information of the DNA may relate with the clinic diagnosis and treatment of some diseases such as the Alzheimer’s disease. The study of complexity and information entropy of DNA sequence reveals the regularities related to the structure and types of special gene repeats as well as their rate of occurrence in the studied regions of genomes. The results demonstrate that information entropy and LZ complexity values have a strong correlation when the size of sliding window is small.

REFERENCES


