2D Graphical Representation of DNA Sequences Based on Variant Map

RUOXUE WU, WENJIA LIU, YUYUAN MAO, AND JEFFREY ZHENG
School of Software, Yunnan University, Kunming 650000, China
Corresponding author: Jeffrey Zheng (conjugatelogic@yahoo.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61362014, and in part by the Key Project of Yunnan Science and Technology under Grant 2018ZI002.

ABSTRACT Mining and analyzing the information and structure of DNA sequences is an important method of exploring and explaining the mysteries of biology. In addition to traditional biological detection methods for DNA sequences, technological advances have allowed researchers to utilize computers to analyze DNA sequences, such as in DNA visualization. However, DNA visualization techniques have disadvantages, such as difficult operation and complex calculation. Therefore, in this paper, we propose an algorithm called VARCH based on a variant map that is a part of variant logic construction. VARCH is a high-efficiency and concise visualization algorithm for DNA sequences. The principle of this approach is that the number of base combinations in each DNA subsequence is counted, and a matrix is constructed based on the full arrangement of different base combinations. Finally, the matrix is mapped to the 2D coordinate plane, obtaining the image of DNA sequences. The experimental results demonstrate that VARCH can effectively display features of DNA sequences.

INDEX TERMS Data visualization, DNA sequence, variant logic construction, variant map, visualization model.

I. INTRODUCTION
In Earth’s long history, a number of species have existed, some of which have disappeared and some of which have survived to this day. It is intriguing to question the difference between species and how species have evolved. As highly intelligent creatures, human beings are committed to answering these questions, solving the mysteries of nature, and exploring the origin of all life, including our own. From the early observation and description of different species, including Darwin’s theory of species in On the origin of species in the 19th century, our understanding of species has advanced. In the 20th century, with the development of technology, scientists used microscopes to study molecular biotechnology. This allowed them to determine that cells are the basic unit of organisms and to study the construction of cells. Later, scientists began to use the molecular double helix model of DNA [1] to reveal the structure of DNA, and used the genetic central dogma [2] to explain the relationship between DNA, RNA, and protein. In the 21st century, scientists completed the DNA sequencing of different species, which revealed the composition of DNA (including nucleic acids, represented by letters A, T, G, and C) and the reasons for differences between different species and different individuals within the same species.

At present, because the DNA of different species is being sequenced, the number of DNA sequences in DNA databases is increasing at an explosive rate [3], [4]. It is thus important to handle and analyze these data quickly and effectively. To better understand the meaning of DNA, a graphical representation can enable the visual inspection of data and provide an effective method for analyzing and comparing among DNA sequences [5]. Up to now, for direct observation, there have been many different graphical representation methods proposed to numerically characterize DNA sequences on the basis of different multi-dimensional spaces. Since Hamori and Ruskin [6] initially proposed using the H-curve to graph DNA sequences in 1983, an increasing number of visualization methods have been proposed [7]–[9]. These methods include 2D space [10], 3D space [11], and higher-dimensional spaces, such as 4D [12], 5D [13], and 6D spaces [14].

However, these methods have several drawbacks. For instance, with some methods that use higher-dimensional space, the results are difficult to read and it is difficult to...
determine the specific coordinates of a certain point in the high-dimensional space. Most methods also face the following three critical challenges. First, the calculation process is complex. Second, it is difficult to observe long DNA sequences using the existing methods. Third, due to the enormous amount of information in a DNA sequence, it is difficult to obtain useful information from a large amount of redundant information.

Therefore, in this study we propose a new method for the visual analysis of DNA sequences based on variant logic construction called VARCH [15]–[17]. This method can help people observe the DNA sequence more intuitively, especially for non-biological researchers, and the feature of the DNA sequence becomes more easily found.

Our main contributions are as follows.

- In VARCH, we use variant logic construction variables to define DNA’s four bases, and use different combinations of variables to represent the DNA sequence. The calculation process of VARCH is thus simpler than that of other methods. In addition, the proposed method can handle long or entire DNA sequences.
- We then use a variant map to reveal the structure of the DNA sequence in 2D space. Our results are easier to read and more intuitive than those of other methods.

The remainder of this paper is organized as follows. In Section II, we review related research. In Section III, we provide background information and describe the principle of variant logic construction. We present the method of this research in Section IV. In Section V, we present our experimental results and a detailed performance evaluation. We provide our conclusions and ideas for future work in Section VI.

II. RELATED WORK

In this section, we briefly review related research on 2D graphical representation methods for bioinformatics visualization. In the 21st century, with the fourth industrial revolution, computer technology has experienced rapid developments that have promoted bioinformatics visualization. Scientists use graphics and geometric models to represent the molecular information of protein [18], DNA [19], RNA [11], and so on. For example, Li [20] proposed the UC-Curve to display the general composition of features of protein sequences and the differences among sequences of protein using 2D graphics. Huang et al. [21] used classification, dual vectors, and Euclidean distance to draw the HR-Curve. Hu et al. [22] proposed a novel method to represent protein sequences by obtaining a 2D discrete point set for protein sequences. Kerpedjiev et al. [23] presented a simple and effective web-based tool for representing the secondary structure of RNA. Then, Elias and Hoksza [24] optimized existing visualization tools based on the corresponding tree that is converted by targets and template structures. Thereafter, to overcome the drawback that some methods cannot always produce intersection-free drawing [25], Wiegreffe proposed an optimization algorithm called RNApuzzler.

For a more intuitive observation of DNA sequences, some researchers have used methods developed in other fields. For example, Liu et al. [26] used the Lempel-Ziv algorithm to display and compare DNA sequences. In addition, Li mapped four nucleotides to a unit circle and constructed a 2D graphic of a DNA sequence called the DUC-Curve. This method was inspired by the DV-Curve [19]. Also based on the DV-Curve, Gong and Fan [27] presented DNA sequences by a zigzag curve. Using different methods, Zou et al. [10] transformed a DNA sequence into a plot set by weight for 64 triplets, and Xie [28] used trigonometric functions to display DNA sequences in 2D space. However, these methods cannot represent all DNA information. Therefore, for displaying detailed DNA data, Halladjian et al. [29] proposed ScaleTrotter, which is based on a scale-dependent camera model. However, due to the complexity of the algorithm, these methods require a large amount of calculation. Our proposed method, VARCH, is easier in that respect. In addition, it is easier to understand and more efficient in extracting information from the graphical representation of DNA sequences.

III. VARIANT LOGIC CONSTRUCTION

Variant logic construction (VLC) is mainly composed of variant logic, variant measurement and variant maps [15]. It is based on classical logic and the Boolean function, adding new operations like replacement and complementarity. Hence, VLC can expand the size of n-variable logical functions from $2^{2n}$ to $2^{2n} \times 2^n$ [16].

A. PRINCIPLE OF VLC

From the definition of a Boolean function [30], we know that if $n$ is a positive integer, $F_2$ is a finite field formed by 0 and 1, and $F_2^n$ represents the vector of N-dimensional space on $F_2$, then we can obtain a mapping, which is an n-variable Boolean function from $F_2^n$ to $F_2$. This mapping can be demonstrated as follows:

$$f : F_2^n \rightarrow F_2, \quad F_2 = \{0, 1\} \quad (1)$$

Due to the number of $F_2^n$ vectors, that is, $2^n$, the number of n-variable Boolean function elements is $2^{2n}$.

Based on this principle, for any logical variant, the corresponding position relationship between $x$ and $y$ can be obtained after the function is determined. In other words, we can utilize the function to establish the corresponding relationship between input and output. According to [30], there are four relationships: $0 \rightarrow 0, 0 \rightarrow 1, 1 \rightarrow 0, 1 \rightarrow 1$, which are called the four primitive types of change patterns.

To expand the logical space of Boolean functions, we introduced VLC [15]. In VLC, we define four meta operators of variant logic ($\perp, +, -, \top$) to represent the four meta change patterns. Table I presents the details of the corresponding relationship.
TABLE 1. The relationships in variant logic construction (VLC).

| Input | Output | Truth | False | Variant | Invariant Operator |
|-------|--------|-------|-------|---------|--------------------|
| 0     | 0      | 0     | 1     | 0       | ⊥                  |
| 0     | 1      | 1     | 0     | 0       | +                  |
| 1     | 0      | 0     | 1     | 0       | −                  |
| 1     | 1      | 1     | 0     | 0       | ⊤                  |

B. VARIANT MAP

According to [17], visualization is an important part of VLC. In VLC, we use a variant map to perform visualization. To visualize a sequence, we must count the number of variant values \{⊥, +, −, ⊤\} in the sequence based on normalization and non-normalization. Therefore, we can obtain 8 probability results of the variant value. The results are presented in Table 2.

TABLE 2. The calculation of normalization and non-normalization.

| Variant | Normalization | Non-normalization |
|---------|---------------|-------------------|
| \(N_⊥\) | \(N_⊥/N\)   | \(N_⊥/N_0\)      |
| \(N_+\) | \(N_+/N\)   | \(N_+/N_0\)      |
| \(N_-\) | \(N_-/N\)   | \(N_-/N_1\)      |
| \(N_⊤\) | \(N_⊤/N\)   | \(N_⊤/N_1\)      |

TABLE 3. The relationship between subsets and values.

| Number | Subset | Value |
|--------|--------|-------|
| 1      | \(\emptyset\) | \(T_0 = 0\) |
| 2      | \(\bot\) | \(T_1 = N_⊥\) |
| 3      | \(\bot\) | \(T_2 = N_+\) |
| 4      | \(\bot\) | \(T_3 = N_-\) |
| 5      | \(\bot\) | \(T_4 = N_⊤\) |
| 6      | \(\bot\) | \(T_5 = N_⊥ + N_+\) |
| 7      | \(\bot\) | \(T_6 = N_⊥ + N_-\) |
| 8      | \(\bot\) | \(T_7 = N_⊥ + N_⊤\) |
| 9      | \(\bot\) | \(T_8 = N_+ + N_-\) |
| 10     | \(\bot\) | \(T_9 = N_+ + N_⊤\) |
| 11     | \(\bot\) | \(T_{10} = N_- + N_⊤\) |
| 12     | \(\bot\) | \(T_{11} = N_⊥ + N_+ + N_-\) |
| 13     | \(\bot\) | \(T_{12} = N_⊥ + N_+ + N_⊤\) |
| 14     | \(\bot\) | \(T_{13} = N_⊥ + N_- + N_⊤\) |
| 15     | \(\bot\) | \(T_{14} = N_+ + N_- + N_⊤\) |
| 16     | \(\bot\) | \(T_{15} = N_⊥ + N_+ + N_- + N_⊤\) |

Then, we select two of combinations to form the 2D scattergram’s x- and y-axis coordinates. Thus, there are \(16 \times 16 = 256\) figures.

IV. PRINCIPLE OF VARCH

In this section, we present the motivation of VARCH by explaining how to visualize DNA sequences based on VLC.

A. VARCH OVERVIEW

In our proposed method, VARCH can be divided into three parts: the preprocess module, the measurement module, and the visualization module. Fig. 1 displays the structure of VARCH. Initially, we preprocess the data of the DNA sequence taken from a database, such as the US National Center for Biotechnology Information (NCBI), to convert the original data into variant logic data. Then, the measurement module handles the data based on the theory of VLC. Thereafter, we obtain 16 different subsets. The visualization module then utilizes these subsets and combines two of them. Later, we use the two subsets to construct the x- and y-axis coordinates. Finally, we obtain 2D graphical representations of DNA sequences.

B. PRE-PROCESS MODULE

The original data of the DNA coding sequences include biological and annotation information, such as the name of the chromosome and protein ID. Therefore, we must filter the original data and delete annotation information such as Sequence’s name, Source, Symbol’<’ and so forth. Moreover, in the DNA sequence there are some unidentified bases represented by the letter N, which is useless for us. So we also remove them. Therefore we only preserve the identified bases’ data. Finally, we obtain pure data that can meet the requirements.

C. MEASUREMENT MODULE

We then utilize the pure data to produce the VLC data. As explained in Section 3, there are four relationships in VLC; therefore, we build a similar relationship between the four bases and variant operators, which is demonstrated in Table 4.

| Equally, based on Section 3.2, we can also obtain 16 combinations of different bases. Details are provided in Table 5. Due to the length of the coding sequence, we must divide the entire sequence into several subsegments. We can then
deal with these subsegments separately, thus simplifying the data processing. As we know, different species have different lengths of DNA sequences. Therefore, in our study, the lengths of subsegments differ based on the species. Moreover, to avoid the subsequence being too long or too short, which makes it impossible to display the features of the entire sequence in their entirety, a coding sequence can be separated over, to avoid the subsequence being too long or too short, the lengths of subsegments differ based on the species. Moreover, for the image to be more easily observed, the performance of the visualization image is considered.

D. VISUALIZATION MODULE

There is an enormous amount of information contained in DNA sequences. According to information theory, the information contained in DNA sequences is related to the probability and uncertainty of each base. To represent the feature of the DNA sequence, as described in Section 3.2, we select two configurations, where each configuration has m elements from all 16 combinations. We can thus obtain 16 × 16 = 256 results and can use the two configurations to build a matrix Mat(B), which is presented in (3).

\[
\text{Mat}(A) = \begin{bmatrix}
T_0^1 & T_1^1 & \ldots & T_j^1 & \ldots & T_{15}^1 \\
T_0^2 & T_1^2 & \ldots & T_j^2 & \ldots & T_{15}^2 \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
T_0^m & T_1^m & \ldots & T_j^m & \ldots & T_{15}^m 
\end{bmatrix}
\tag{2}
\]

\[
\text{Mat}(B) = \begin{bmatrix}
b_{11} & b_{12} & \ldots & b_{1h} & \ldots & b_{1m} \\
b_{21} & b_{22} & \ldots & b_{2h} & \ldots & b_{2m} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
b_{m1} & b_{m2} & \ldots & b_{mh} & \ldots & b_{mm} 
\end{bmatrix}
\tag{3}
\]

where \(b_{ij}\) represents the frequency value of the coordinate \((T_a^i, T_b^j)\). The coordinate \((T_a^i, T_b^j)\) signifies that the two configurations correspond to the elements in the matrix Mat(A) separately, as \((T_a^1, T_b^1), (T_a^2, T_b^2), \ldots, (T_a^m, T_b^m), 1 \leq i \leq m, 1 \leq h \leq m\). We can then use Mat(B) to generate a 2D visualization image. In VARCH, when \(m \in (20, 300)\), the performance of the visualization image is considered clear. Moreover, for the image to be more easily observed,
we reverse the y-axis direction. For example, Figure 3 shows the 2D visualization image of the Aotus nancymaae DNA sequence for different \( m \) values.

Due to the symmetry between the four bases and 16 configurations, of 256 images, we select 64 images sequentially that can display the features of all results, as illustrated in Fig. 4. We summarize the process of VARCH in Algorithm 1.

**Algorithm 1 VARCH Algorithm**

**Input:** the data of a certain DNA sequence after preprocessing, the length of DNA sequence \( N \), set subsegment length \( Len \)

**Output:** 2D images of a certain DNA sequence

According to Table 4, transforming bases from A,G,C,T to ⊥, +, −, ⊤

Calculate the number of subsegments \( m \) by

\[
m = \lfloor \frac{N}{Len} \rfloor
\]

Set \( i \) is the current number of the subsegment

while \( i <= m \) do

According to Table 5, calculate 16 configurations from \( T_0 \) to \( T_{15} \)

Record all results in a matrix \( \text{Mat}(A) \) like Equation (2)

end while

Build a \( m \times m \) matrix \( \text{Mat}(B) \) by choosing two configurations from \( \text{Mat}(A) \) as in Equation (3)

Use \( \text{Mat}(B) \) to generate a 2D visualization image

The program using the Python 3.0 programming language to process the data and implement the VARCH algorithm and built the software environment required for our experiment in PyCharm 2018.3.7. The DNA sequence data were taken from the NCBI database [31].

**B. EXPERIMENTAL EVALUATION**

In this section, we evaluate the performance of VARCH and compare the results among different species.

1) THE RESULT OF DIFFERENT PARAMETERS

As discussed in Section 4, we know that different parameters \( m \) affect the length of a subsequence, which can affect the details of the image presentation. Based on Section 4, we know that \( m \) belongs to (20, 300). Comparing Figs. 3(a) and (b), different parameters cause different visualization shapes. Therefore, it is necessary to determine the optimal parameter value so that the characteristics of the gene sequence can be observed more clearly. We utilized an Aotus nancymaae DNA sequence to test the performance of different \( m \) values. In addition, we used the \( T_8 \) (A+G) and \( T_5 \) (A+T) configuration to map on the Y- and X-axis coordinates, respectively. The results are presented in Fig. 5.

Fig. 5 indicates that the \( m \) value is related to the size of the image. In addition, comparing Figs. 5(a)-(c), we notice that when \( m < 50 \), the image shows symmetry. In addition, with the decrease of the \( m \) value, the symmetry becomes more notable. Moreover, as the \( m \) value increases, the image features of the DNA sequence become more obvious. For instance, in Fig. 5, we know that the shape of the Aotus nancymaae DNA sequence by visualization has a rabbit-like shape. In particular, from Figs. 5(c)-(j), we can easily observe that the images have an ear-like shape. Second, comparing Figs. 5(a)-(j), we see that when the \( m \) value is smaller, the image is closer to a circle. In addition, Figs. 5(g)-(j)

**V. EXPERIMENT**

**A. EXPERIMENT SETUP**

The program ran on a Macbook Pro laptop with an Intel Core i7 CPU at 2.5 GHz and 16 GB memory. We developed

![Image](attachment:image1.jpg)

**FIGURE 3.** 2D visualization of Aotus Nancymaea's DNA on \( m = 20 \) and \( m = 300 \).

![Image](attachment:image2.jpg)

**FIGURE 4.** The 64 example results.

VOLUME 8, 2020 173759
FIGURE 5. Visualization of the Aotus Nancymaeae's DNA sequence at different $m$ values.

demonstrate that with the increase of the $m$ value, the image displays a scatter trend.

We thus consider that when the $m$ value is approximately 80, the performance of the visualization image is optimal, as in Fig. 5(f). In this condition, the feature of the DNA sequence is the easiest to observe.

2) COMPARISON AMONG DIFFERENT SPECIES

To assess the performance of VARCH, we estimated the similarity of different DNA sequences by images and compared the results with existing data.

In this paper, the similarity of different images is calculated by the Euclidean distance. According to [32], a large Euclidean distance means that the similarity between images is low. In this paper, we calculate the Euclidean distance of each point’s coordinate for two figures by (4).

$$d = \sqrt{\sum_{i=1}^{N} (x_{i1} - x_{i2})^2}$$

(4)

where $d$ represents the result of the Euclidean distance, $N$ is the number of points in the image, and $x_{i1}$ and $x_{i2}$ represent the point’s coordinates in the different images, respectively.

To display the results more comprehensively, we compared the image results of DNA sequences on different chromosomes for the same species. For instance, we used VARCH to display the image of Homo sapiens (human) DNA sequences in 23 chromosomes, which is illustrated in Fig. 7 in the Appendix. In Fig. 7, we set the $m$ value to 78, and selected the $T_8$, $T_9$ configuration to map on the coordinates separately.

Then, we used (4) to calculate the similarity between all subfigures in Fig. 6, and part of the results are presented in Table 6.

To validate the accuracy of our method, we used BLAST algorithm to align the sequences. Table 7 presents the results of the same chromosomes. Combining Tables 6 and 7, we observe that the result of the Euclidean distance correctly reflects the similarity of different sequences. For instance, when the Euclidean distance between sequences on No.2 and 6 chromosomes is the lowest in Table 6, the corresponding result in Table 7 also demonstrates the highest similarity. In addition, Table 6 reveals that the sequence on the No.19 chromosome is quite different from other chromosomes, especially the X chromosome. Furthermore, in Table 6, most of the results are in [0, 70000], particularly in [30000, 50000], and the data of Table 7 verify these results.

Next, we compare images which are produced by different chromosomes in different species. In our experiments, we choose Mus musculus (house mouse), Pan troglodytes (chimpanzee) and Homo sapiens (human) to compare each other. We show images of these gene sequences in Figure 8 and Figure 9 in the Appendix. For a fair comparison, we set $m$ value also is 78 and use $T_8$, $T_9$ configuration in Figure 8 and 9. Then we also calculate the Euclidean distance between Homo sapiens and Mus musculus, Pan troglodyte respectively. Next, we show the result in Figure 6.

The results in Fig. 6 reveal that the Euclidean distance between Homo sapiens and Mus musculus is larger than the distance between Homo sapiens and Pan troglodyte, which signifies that the similarity between humans and chimpanzees
TABLE 6. The Euclidean distance among Homo sapiens’ (human) chromosomes.

| The chromosome’s number | 1  | 2  | 3  | 6  | 11 | 12 | 15 | 16 | 19 | 21 | X | Y  |
|-------------------------|----|----|----|----|----|----|----|----|----|----|---|----|
|                         | 1  | 0  | 27136.67 | 35165.12 | 33703.65 | 39426.37 | 43831.81 | 38373.92 | 53610.06 | 85464.63 | 58542.99 | 40981.73 | 32722.66 |
|                         | 2  | 0  | 13775.59 | 13247.69 | 38718.26 | 37335.44 | 43538.37 | 69162.53 | 96311.64 | 130923.29 | 52593.01 | 75234.26 | 56716.46 |
|                         | 3  | 0  | 13775.59 | 13247.69 | 38718.26 | 37335.44 | 43538.37 | 69162.53 | 96311.64 | 130923.29 | 52593.01 | 75234.26 | 56716.46 |
|                         | 6  | 0  | 27136.67 | 35165.12 | 33703.65 | 39426.37 | 43831.81 | 38373.92 | 53610.06 | 85464.63 | 58542.99 | 40981.73 | 32722.66 |
|                         | 11 | 0  | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 | 13046.89 |
|                         | 21 | 0  | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 |
|                         | X  | 0  | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 | 13550.03 |

FIGURE 6. The Euclidean distance among Homo sapiens’ (human) chromosomes.

TABLE 7. The similarity among Homo sapiens’ (human) chromosomes.

| The chromosome’s number | 1  | 2  | 3  | 6  | 11 | 12 | 15 | 16 | 19 | 21 | X  | Y  |
|-------------------------|----|----|----|----|----|----|----|----|----|----|----|----|
|                         | 1  | 100 | 40.76% | 35.83% | 36.29% | 32.58% | 30.16% | 33.62% | 26.68% | 25.75% | 26.35% | 32.01% | 37.22% |
|                         | 2  | 100 | 100%  | 44.22% | 44.75% | 33.22% | 34.64% | 30.46% | 26.13% | 23.66% | 27.44% | 40.65% | 39.79% |
|                         | 3  | 35.83% | 44.22% | 100%  | 43.95% | 32.31% | 39.08% | 28.43% | 25.99% | 25.51% | 29.12% | 41.97% | 39.21% |
|                         | 6  | 36.29% | 46.75% | 43.95% | 100%  | 25.84% | 31.80% | 27.75% | 25.88% | 25.30% | 26.94% | 40.47% | 38.23% |
|                         | 11 | 32.58% | 35.83% | 36.29% | 32.58% | 100%  | 43.43% | 42.63% | 28.14% | 25.81% | 36.95% | 31.24% | 31.67% |
|                         | 12 | 30.16% | 34.64% | 35.08% | 31.80% | 43.43% | 100%  | 41.74% | 26.46% | 25.74% | 41.06% | 31.76% | 31.15% |
|                         | 15 | 33.02% | 30.46% | 28.45% | 27.75% | 42.63% | 41.74% | 100%  | 32.90% | 25.91% | 34.06% | 27.77% | 29.59% |
|                         | 16 | 26.68% | 26.13% | 25.93% | 25.88% | 28.14% | 26.46% | 32.90% | 100%  | 27.04% | 26.27% | 25.85% | 26.17% |
|                         | 19 | 25.75% | 25.66% | 25.31% | 25.30% | 25.81% | 25.74% | 25.91% | 27.04% | 100%  | 25.80% | 25.09% | 25.67% |
|                         | 21 | 26.35% | 27.44% | 25.12% | 26.94% | 36.95% | 41.00% | 34.06% | 26.27% | 25.80% | 100%  | 27.48% | 26.36% |
|                         | X  | 32.01% | 40.63% | 41.97% | 40.47% | 31.24% | 31.76% | 27.77% | 25.85% | 25.09% | 27.48% | 100%  | 36.34% |

is greater than between humans and a house mouse. This result is consistent with the theory of biological evolution. In addition, in Fig. 6, most of the Euclidean distance results between humans and chimpanzees are in [40000, 55000]. The largest result is the No.18 chromosome, which signifies that there is a large difference between human and chimpanzee DNA sequences in the No.18 chromosome. The smallest result of the Euclidean distance between a human and house mouse is approximately 44000 in the X chromosome.

We also investigate this method execution time in the different situation. The execution time of the VARCH method for one picture of about 240MB data like Homo sapiens’ No.1 chromosome is about 6s and for 50MB data, the time is about 1s. Although VARCH method is able to generate 256 images at once time, the total time overhead is not significantly increased with only 9s for Homo sapiens’ No.1 chromosome in our experimental environment.

VI. CONCLUSION AND FUTURE WORK

In this study, we propose a new 2D graphical representation of DNA sequences called VARCH, which is based on variant logic construction. The principle of VARCH is that the bases in the DNA sequence are replaced with four meta operators in the variant logic construction. We then calculate the number of different 16 configurations of the variant logic construction for a DNA sequence. Next, VARCH selects any two configurations to map onto the X- and Y-axis coordinates, respectively. Finally, VARCH displays the variant map of the DNA sequence in 2D space. In Section V, we investigate the effect of different parameters and compare different images for different DNA sequences. The results demonstrate that VARCH can effectively reveal the relationship and diversity among different DNA sequences through images.
However, this study can still be promoted in the future. First, we do not specifically study the relationship between each picture; that is, we do not evaluate whether the pictures are related. In addition, there are problems related to
degeneracy and loss of information. Therefore, in future work, VARCH must be optimized. Second, in our experiment, most results are based on subjective judgement. Thus, in the future, more objective and scientific methods should be used to describe the results. Third, we did not generate accurate mathematical models and relationships between...
the Euclidean distance results and similarity results, as shown in Table 6 and 7. Therefore, we will research how to build precise models in future work.

APPENDIX
See Figures. 7–9.

**ACKNOWLEDGMENT**
The authors thank Wukang Liu and Xiaoxuan Liang for their help in experiments.

**REFERENCES**
[1] J. D. Watson and F. H. C. Crick, “Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid,” Nature, vol. 171, no. 4356, pp. 737–738, Apr. 1953.
[2] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970.

[3] K. Kumar, “Introduction to bioinformatics,” *J. Roy. Stat. Soc., A, Statist. Soc.*, vol. 171, no. 3, pp. 761–772, Jun. 2008.

[4] A. S. Krawetz and D. D. Womble, *Introduction to Bioinformatics*. Totowa, NJ, USA: Humana Press, 2003.

[5] P. Hoffmann, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, “DNA visual and analytic data mining,” in *Proc. Vis.*, Oct. 1997, pp. 437–441.

[6] E. Hamori and J. Ruskin, “H curves, a novel method of representation of nucleotide series especially suited for long dna sequences,” *J. Biol. Chem.*, vol. 258, pp. 1318–1327, Jan. 1983.

[7] N. Alharbi, M. Alharbi, X. Martinez, M. Krone, A. Rose, M. Baaden, R. S. Laramee, and M. Chavent, “Molecular visualization of computational biology data: A survey of surveys,” in *Proc. Eurograph/IEEE VGTC Conf.*., *Vis., Short Papers, Eurograph. Assoc.*, 2017, pp. 133–137.

[8] B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege, “Visualization of biomolecular structures: State of the art revisited,” *Comput. Graph. Forum*, vol. 36, no. 8, pp. 178–204, Nov. 2016.

[9] H. Miao, T. Klein, D. Kouli, P. Mindek, K. Schatz, M. E. Gröller, B. Kozlíková, T. Isenberg, and I. Viola, “Multiscale molecular visualization,” *J. Mol. Biol.*, vol. 431, no. 6, pp. 1049–1070, Mar. 2020.

[10] S. Zou, L. Wang, and J. Wang, “A 2D graphical representation of the sequences of DNA based on triplets and its application,” *EURASIP J. Bioinf. Syst. Biol.*, vol. 5, no. 9, pp. 1202–1209, May 2013.

[11] N. Lindow, D. Baum, M. Leborgeon, and H.-C. Hege, “Interactive visualization of RNA and DNA structures,” *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 967–976, Jan. 2019.

[12] X. Tang, P. Zhou, and W. Qiu, “On the similarity/dissimilarity of DNA sequences based on 4D graphical representation,” *Chin. Sci. Bull.*, vol. 55, no. 8, pp. 701–704, Mar. 2010.

[13] B. Liao, R. Li, W. Zhu, and X. Xiang, “On the similarity of DNA primary sequences based on 5-D representation,” *J. Math. Chem.*, vol. 42, no. 1, pp. 47–57, Apr. 2006.

[14] B. Liao and T.-M. Wang, “Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases,” *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 5, pp. 1666–1670, Sep. 2004.

[15] J. Z. J. Zheng and C. H. Zheng, “A framework to express variant and invariant functional spaces for binary logic,” Frontiers Electr. Electron. Eng. China, vol. 5, no. 2, pp. 163–172, Jun. 2010.

[16] J. Zheng, *Variant Construction from Theoretical Foundation to Applications*. Singapore: Springer, 2019, pp. 51–65.

[17] J. Zheng and C. Zheng, “Variant simulation system using quaternion structures,” *J. Mod. Opt.*, vol. 59, no. 5, pp. 484–492, Mar. 2012.

[18] M. Randic, J. Zupan, T. A. Balaban, D. Vukić-Topić, and D. Plavsic, “Efficient outerplanar drawing of RNA-secondary structures,” *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 654–664, Jan. 2020.

[19] R. O'Donnell, *Analysis Boolean Functions*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[20] Y. Li, Q. Liu, X. Zheng, and P.-A. He, “UC-curve: A highly compact 2D graphical representation of protein sequences,” *Int. J. Quantum Chem.*, vol. 114, no. 6, pp. 409–415, Nov. 2013.

[21] L. Huang, H. Tan, and B. Liao, “HR-curve: A novel 2D graphical representation of protein sequence and its multi-application,” *J. Comput. Theor. Nanoscience*, vol. 10, no. 1, pp. 257–264, Jan. 2013.

[22] H. Hu, Z. Li, H. Dong, and T. Zhou, “Graphical representation and similarity analysis of protein sequences based on fractal interpolation,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1, pp. 182–192, Nov. 2017.

[23] P. Kerpedjievec, S. Hammer, and I. L. Hofacker, “Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams,” *Bioinformatics*, vol. 31, no. 20, pp. 3377–3379, Jun. 2015.

[24] R. Elias and D. Hoksza, “TRAVeLer: A tool for template-based RNA secondary structure visualization,” *BMC Bioinf.*, vol. 18, no. 1, pp. 1–10, Nov. 2017.

[25] D. Wiegreffe, D. Alexander, P. F. Stadler, and D. Zeckzer, “RNApuzzler: Efficient outerplanar drawing of RNA-secondary structures,” *Bioinformatics*, vol. 35, no. 8, pp. 1342–1349, Sep. 2018.

[26] C. Liu, C. Li, F. Bai, Q. Zhao, and Y. Wang, “An optimization approach and its application to compare DNA sequences,” *J. Mol. Struct.*, vol. 1082, pp. 49–53, Feb. 2015.

[27] W. Gong and X.-Q. Fan, “A geometric characterization of DNA sequence,” *Phys. A, Stat. Mech. Appl.*, vol. 527, Aug. 2019, Art. no. 121429.