Comparison-specialized Visualization Model for Whole Genome Sequences

Da-Young Lee Kyung-Rim Kim Taeyong Kim Hwan-Gue Cho*
Dept. of Electrical and Computer Engineering
Pusan National University, South Korea
{schematique, alflskfl, ktyong22, hgcho}@pusan.ac.kr

ABSTRACT

Analyzing and visualizing the whole genome sequence is very important to finding genetic evolution. Many researchers have used 2D or 3D DNA random walk plots to study short DNA sequences. However, visualizing a whole genome sequence is difficult because of overlapping, self-intersection, and biases. In this paper, we propose a 3D graphical representation of a whole-genome sequence based on a random walk plot. Our 3D graphical representation can reduce the overlaps or biases that can occur during the visualization of large sequences by using the 2D or 3D DNA walk plot algorithm. We visualized and compared data on the whole genomes of 10 species, including humans and anthropoid apes. In our experiment, the 3D graphical representation showed similarities between humans and apes and differences between other species.

Keywords

DNA visualization, random walk, DNA similarity

1 MOTIVATION

By the late 1990s, genetic maps were developed by using genome projects based on Sanger's method of gene analysis. A great deal of research was carried out. Since then, next-generation sequence (NGS) techniques such as Roche 454 and Illumina have been developed, and mass gene data have become available. Now, gene data consisting of billions of base points can be obtained in only a few hours. Obtaining gene data has become more easier. Gene analysis plays an important role in understanding biological features such as genetic expression and diversity and in the medical diagnosis, prevention, and treatment of genetic illnesses.

The Bio information in genetic material is decided by order of base composition, geneticists analyze what bio information is retained on particular sequence of particular position by comparing the order of encoded gene information. The most popular analysis method for gene information stored as data is the alignment algorithm. After the reference sequences are read, they are compared with the query sequence that the user wants to understand. Thus, not only the similarity score but also similar points between the two sequences and indel mutations can be confirmed. The alignment algorithm provides the advantage of detailed information about the characteristics or mutation of a gene. However, it requires many computational operations running at $O(n^2)$ times and a large amount of memory space. This makes it difficult to analyze considerably long sequences like a whole genome.

In contrast, analysis methods based on static components are used for obtaining rough gene information. These methods use statistical features such as the proportion and distribution of A, G, T, and C. Statistical results can be obtained from a single pass of reading the gene information. However, as the amount of gene data increases, the overall statistical proportion of the gene information converges to one point. Consequently, it is difficult to obtain feature points of long sequences and confirm details of the sequence.

We propose a visualization method of the geometric space that uses random walk with base sequences consisting of character strings. The computed DNA sequence information by preprocessing can be rapidly visualized, and the visualization results make it easy for the user to check the structure of the DNA base sequence.

2 RELATED WORK

2.1 Visualization for Sequence Data

Sequence data are a set of ordered data that change with time, such as economic indicators or weather, and order- dependent data such as documents and DNA

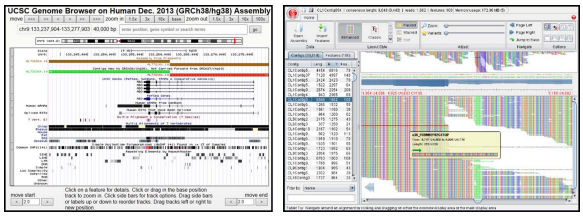
base sequences. The most common type of sequence data is the time series, which is utilized to study the current state or predict the future state. Research in this field is based on correlation analysis of collected information with an associative relationship but unknown rules, such as between the stock price and interest rate or between the temperature, humidity, and weather. Most sequence data processing involves collecting and analyzing a great deal of data. The given data are mostly stored and managed in a text or compressed format. Examining such huge amounts of data directly has a high computational cost. Developing a more efficient method for data visualization is an important topic of research.

To determine the correlation between data more effectively, Alencar et al. [1] and Krstajic et al. [2] proposed methods for comparing different types of data at a time by compressing the memory space. In addition, Graells and Jaimes [3] presented a method for visually grasping the data progress by compressing data that are not chosen by the user as a significant part with a helical form. Thakur and Hanson [4] suggested a visualization tool based on location information to compare the data trends among regions.

2.2 Genomic Analysis

As previously stated, the DNA base sequence is bioinformation encrypted in the form of bases. It is typical sequence data because biological features are determined by the order of bases. By analyzing the function of the encrypted base sequence, the characteristics can be obtained through a comparison with other well-known base sequences.

A typical analytical comparison of the base sequences involves using a tool based on the alignment algorithm, such as BLAST [5], Bowtie [6], or BWA[7] to grasp the similarity between the two sequences. Such tools eventually determine the similarity by using the alignment algorithm, which was described by Smith and Waterman [8]. However, the algorithm has a space complexity of $O(n^2)$. Thus, the longer the base sequence length, the greater the required computing time and memory space.



(a) Genome Browser (b) Tablet

Figure 1: The web based visualization tool (a) Genome Browser and also visualization tool but provides mapping result (b) Tablet. Using these visualization tool, we can check the detail information about genome, but grasp of the whole structure of DNA base sequence is difficult.

Research has also been carried out on analyzing base sequence information by visualizing the results of base sequence analysis is also carried out. Genome Browser [9, 10] visualizes various genome analysis results that can immediately be obtained from a website. The Tablet tool [11] provides the mapping result based on Bowtie or BWA in visual form for the user. These visualization tools give very detailed analysis results but cannot provide information about the overall features of the genome.

2.3 Random Walk Visualization Model

The random walk plot is a visualization technique that allocates states to features of the data and represents the change in states with the coordinates according to the priority so that users can easily visualize the data contents. Overall, a two-dimensional walk plot is visualized with four states. The characters A, G, T, C are allocated to the DNA base sequences of each state to confirm the overall form of the genome universally.

The process of DNA visualization by using a two-dimensional walk plot is as follows.

1. Read the base sequences from the DNA data.
2. Convert the base to unit vectors.
3. Visualize the sequence by using the unit vectors.

The unit vector can be represented with various form which depends on the setting of direction for ‘A’,‘G’,‘T’,‘C’, in case of WS-curve, the ith base of sequence is S_i, the unit vector $Unit^{WS}(i)$ is defined as follows:

$$Unit^{WS}(i) = \begin{cases} (-1,\ 0) & if\ \ S_i = A \\ (\ 0,+1) & if\ \ S_i = G \\ (+1,\ 0) & if\ \ S_i = T \\ (\ 0,-1) & if\ \ S_i = C \end{cases} \quad (1)$$

And then finally visualize the result of the sum of unit vector in order of base reading.

A two-dimensional random walk generally visualizes a sequence by mapping A, G, T, and C in each direction, as discussed earlier. This produces three main varieties: the WS-curve, RY-curve, and MK-curve. These depend on which bases are allocated in directly opposite directions of each other. The RY-curve allocates puRine(R = A, G) and pYrimidine(Y = C, T) as complementary relations. The MK-curve assigns aMino(M = A, C) and Keto (K = G, T) as complementary relations. The WS-curve allocates Weak (W = A, T) and Strong (S = G, C) as complementary relations. The two bases are combined in a double helix. The visualization results change depending on how we set the complementary axis, even if the same sequence is used.

Authors	Dimensions	Main characteristics
Kim et al. [12]	2	visualization in polygon form using K-Convex and comparison.
Liao and Ding [13]	3	Allocate the z axis with time and visualize the 1*1*1 space with a walk plot form and compute the entropy of those results to compare the similarity between genomes.
Bai et al.[14]	1	Read the base sequences, combine the complementary form into the genome as per the classifications of WS, RY, MK, visualize according to he order of sequences, and convert the results into entropy to compare the similarity between genomes.
Lo et al. [15]	3	Define the unit vector of A, G, C, T as a tetrahedron in a three-dimensional space and visualize the DNA.
Xie and Mo [16]	3	Add the axis representing the order of the sequence to the previous two-dimensional walk plot and visualize the DNA. With the additional axis, the beginning and end points of RY, MK, and WS provide the number of each base A, G, T, and C.

Table 1: Characteristics of DNA visualization and analysis research using a random walk plot

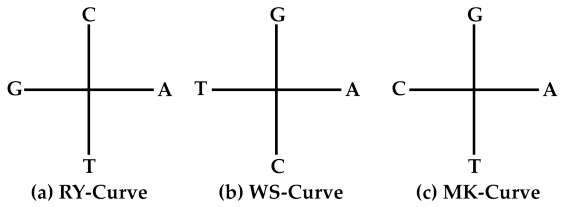


Figure 2: 2 dimensional walk plot.

Thus, we generally utilize all three curves to compare sequences. However, the genome bases A and T are generally much more common than G and C. If the sequence length is too long, the visualization result of the sequence tends to become lopsided.

The WS-curve, RY-curve, and MK-curve are each divided into four kinds depending on how the complementary axis is selected. For example, besides the above unit vector, the WS-curve is divided into four kinds depending on the directions of A, T, and G, C:

$$Unit^{SW}(i) = \begin{cases} (-1,\ 0) & if\ \ S_i = A \\ (\ 0,+1) & if\ \ S_i = G \\ (+1,\ 0) & if\ \ S_i = T \\ (\ 0,-1) & if\ \ S_i = C \end{cases} \quad (2)$$

These graphs are used to prevent different resulting values depending on the three curves of the axis when computing similarities or dissimilarities. Thus, the visualization results for the walk plot vary with the conversion method from each base to the unit vector. Many studies have examined methods for effective visualization. Table 1 introduces different studies on how to utilize the random walk plot.

There have been many studies on visualization with a three-dimensional random walk based on the concepts for a two-dimensional random walk. The biggest problem with the two-dimensional random walk is the loss of data by the two bases in the opposite directions. To prevent such a problem, many studies have allocated the z axis as a time stamp. Liao and Ding [13] represented the z axis as $1 - i/1$ so that the whole sequence appears in the space between $(0,0,0)$ and $(1,1,1)$ instead of showing just a simple cumulative sum of vectors about A, G, T, C. By representing the whole sequence in a given space, comparison results such as the similarity can be numerically expressed by $0 <= Similarity/Dissimilarity(x) <= 1$. However, the longer the sequence, the more difficult it is for the user to understand the visualization results. Xie and Mo [16] visualized the data by adding an axis to the basic two-dimensional visualization method and increasing the additional axis value for each base pair. Thus, just the location of the end point of the random walk provides the rate of each base (A, G, T, and C), and entropy can help with determining the similarity between different sequences. However, a longer sequence, causes the axis representing the order of base pairs to be out of proportion to the other axis, and the visualization results are too compressed for massive sequences.

To avert the loss of base information, Lo et al. [15] mapped each base to each vertex of a tetrahedron in three dimensions instead of to the axes x and y. This method has the advantage of effectively representing the sequence characteristics, but the features of a base cannot be clearly shown in each direction. In addition, information is lost by the other three bases.

3 NEW VISUALIZATION MODEL

3.1 Usefulness of Three-dimensional Walk Plot Visualization

Random walk visualization is useful for visualizing DNA data in string form to make the results easier to understand. However, previous research focused on short DNA data lengths; their methods present drawbacks for visualizing long sequences. For example, table 2 presents the rate of the 2-mer base of human chromosome 1.

The ratio of the base pairs AT and GC was about 19.23%. The visualization process based on the WS-Curve incurred heavy data losses. To prevent this prob-

2mer	bp	ratio	2mer	bp	ratio
AA	21,411,841	9.50%	CC	12,264,137	5.44%
GG	12.274.467	5.45%	TT	21,464,952	9.53%
AC(CA)	29,523,084	13.11%	AT(TA)	31,095,833	13.80%
AG(GA)	27,699,140	12.30%	CG(GC)	12,235,412	5.43%
CT(TC)	27,746,387	12.32%	GT(TG)	29,565,347	13.12%

Table 2: The rate of two pair of base in Human chromosome

Figure 3: The transition process of sequence. Read the sequence with overlapped form per 2-mer unit. Through this process, we can find the complementary 'AT(TA)', 'GC(CG)' in sequence.

lem, other curves such as the MK-curve and RY-curve are used to check the sequence. However, A and T are generally more common than G and C, so a longer sequence makes the graph form more lopsided. Thus, this approach is inappropriate for visualizing long sequences. Three-dimensional visualization tries to reduce this problem by using a time axis, but a long sequence makes the time axis too big. It is difficult to grasp the visualization results at a glance. In this paper, we propose a new three-dimensional visualization method to improve the visualization results for a long DNA sequence. To reduce the loss of data, which is the repeated sections of AT and GC, we can represent that repetition through the z axis if the sequences AT and GC are repeated. Thus, we can get the same results of existing two-dimensional visualizations by using orthogonal projection for the XY plane with our new method, and we can check for loss of information caused by repetition of the AT and GC bases by viewing the results from another angle.

3.2 Proposed Walk Plot Procedure

To determine the AT and GC parts in DNA sequence data more easily, we reset the relation between the unit vector and each base, as given in Table 3, based on existing definitions given in section 2.

2mer	Symbol	Vector	2mer	Symbol	Vector
AA	A	(2, 0, 0)	AG(GA)	U	(1, 1, 0)
AC(CA)	V	(1,-1, 0)	AT(TA)	W	(0, 0,-2)
CC	C	(0,-2, 0)	CG(GC)	X	(0, 0,+2)
CT(TC)	A	(-1,-1, 0)	GG	G	(0, 2, 0)
GT(TG)	Z	(-1, 1, 0)	TT	T	(-2, 0, 0)

Table 3: Reset the vector for visualization of 3 dimensional random walk.

As indicated in Table 3, to detect AT and GC in the sequence, we combine the base pairs and convert the sequence into coordinates. In order to prevent the wrong combination of 2-mer unit bases, the base pairs are read in overlapped form, as shown in Figure 3.

The base pairs AT and GC are represented on the z axis. The other base pairs are represented as the sum

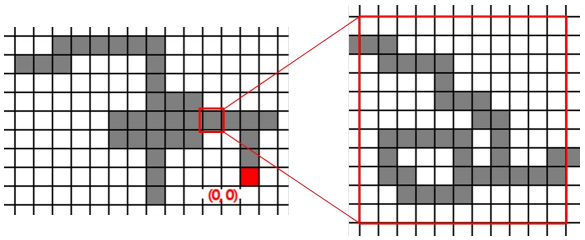


Figure 4: The phenomenon that simplification of widespread data of visualization result on limited screen. To visualize the whole data in limited space, the random walk such as (b) is simplified into (a).

of two unit vectors for each base, as given by the WS-curve method. After the vector transition for DNA genome data information, those vectors are visualized in three-dimensional space. The method of visualization is the same as that of two-dimensional visualization, where the sum of vector values is computed according to the order of sequences and the results are connected with a line to provide the final visualization result. For the random walk plot R,the beginning point is $R(0) = (X_0, Y_0, Z_0)$ $(X_0 = Y_0 = Z_0 = 0)$. $Unit^{3d}(i)$ is the converted value of the ith 2mer of the unit vector. The ith point $R(i) = (X_i, Y_i, Z_i)$ of the random walk plot is computed as follows:

$$R(i) = R(i-1) + Unit^{3d}(i) = \sum_{0}^{i} Unit^{3d}(i) \qquad (3)$$

3.3 Simplification and Normalization of the 3D Walk Plot

When trying to visualize a base sequence bigger than 100 Mbp like a chromosome with random walk, the screen size is more limited than the range of coordinates represented on the walk plot. Thus, it is impossible to represent the whole sequence. Usually, when representing these values with such a wide range in coordinates on a screen, the points in a certain range are generally simplified into one pixel, as shown in Figure 4.

However, reading these many points with little influence on the results and reflecting them in the visualization greatly wastes time and memory space during operation. Therefore, we fixed the screen size and developed a simplification preprocess for the vector transition time to reduce the unnecessary operation and processing time.

The ith point of walk plot R is $R(i)$, and the screen range of visualization is $[-v, v]$. Then, the simplification results can be defined as follows:

1. Find the maximum value $max(R)$ from the absolute values of the x, y, z components for each point of R.

2. Determine the range of simplification $ran = max(R)/v$ by using the maximum value.

3. Convert all points of R into the simplified point $R'(i) = R(i)/ran$ by using the range of simplification ran.

4. Remove the continuous overlapping points among the converted points $R'(i)$.

The preprocess to simplify all of the points in the range ran into one point, as shown in Figure 4, reduces the number of points that need to be visualized. As explained before, however, in the case of DNA base sequences, A and T are generally more common than G and C. Thus, all curves excluding the WS-Curve have lopsided visualization results. This problem is also encountered for the z axis of the three-dimensional walk plot, which is determined by combining the high-frequency base AT (TA) and low-frequency base GC (CG). The rate of AT is 13.80%, which is much higher than that of GC (5.43%), as given in Table 2. This imbalance in the range of data slightly differs for each base sequence of the chromosome. However, similar trends can also be observed in the chromosomes of other species besides humans. In the case of short sequences, this does not greatly affect the result. However, with longer sequences, the range of the z axis is much bigger than that of the x and y axes. This makes the results difficult for the user to check. Table 4 presents the resulting maximum values of each axis for human chromosomes 1–5 based on the computed results for the walk plot R.

The z axis has a much larger value than the x and y axes. Because of this difference in the range of data, when the whole sequence is visualized, the changes in the x and y axes are too slight compared to that of the z axis. Thus, it is difficult to sense the changes in x and y. To solve this problem, when we simplify the data, we compute the ranges of simplification for x, y, and z independently. We defined this process as the normalization of the three-dimensional walk plot. The maximum values among the absolute values of the x and y components in the walk plot R are $max_{XY}(R)$. The maximum value of the z component is $max_Z(R)$, and the view size for visualization is v. Then, the results of normalizing each point can be defined as follows:

$$R_{regular}(i) = \left(X_i \frac{v}{max_{XY}(R)}, Y_i \frac{v}{max_{XY}(R)}, Z_i \frac{v}{max_Z(R)} \right) \quad (4)$$

In the preprocess, when we simplify a long base sequence by using the normalization method in advance, it is cumbersome to change the screen size or check the detailed information because the preprocess needs to be run again. However, if we visualize a chromosome of 100 Mbp in three-dimensional space in the range of $[-400, 400]$, the number of points is reduced to below $\frac{1}{100}$ (this slightly differs depending on the range of data and direction of progress), so the visualization results can be quickly obtained. Furthermore, if the screen size is fixed, the data can be visualized without computing each point by always using the preprocess results.

4 COMPARISON AND 3D VISUAL SHAPE

The walk plot results generated by simplification can be visualized on three-dimensional coordinates as a set of points. A simple method to grasp the similarity within a set of points is comparing the similarity of the polygon with the form containing all of the points. However, finding areas with a polygon form that contains all of the points and comparing these areas are not easy to do.

In this section, we compare orthographic projection planes such as XY, YZ, and XZ instead of comparing the visualization results in three-dimensional space to reduce the complicated computation and compare the results more easily. To effectively determine the area on the two-dimensional space and compare the area generated by each sequence, we assumed a "beta" shape. The similarity comparison method uses this beta shape.

Finding the area that contains the results of the walk plot is very important to comparing similarities based on the visualization result. In the case of a relatively dense area, if there are slight mutations in each sequence, they will be recognized as different. In contrast, if the area is too large, the results will not demonstrate the characteristics of the sequence. The simplest method for finding the area that contains all of the points on the two-dimensional plane is to use a bounding box or convex hull. The bounding box can define an area more easily if the maximum and minimum values of each coordinate are found. However, this tetragonal area is too rough to comprehend the characteristics of sequences.

The convex hull is the connection of the outermost points to make the smallest hull that includes all of the points. The convex hull provides a more concrete area than the bounding box but uses the outermost points; if there are many large empty spaces, it cannot consider the problem, as shown in Figure 5(a).

The alpha hull (alpha shape) [17, 18] was proposed in 1983 and is shown in Figure 5(b). This involves first selecting two points and drawing a circle with the radius of alpha. If there are no points in the circle, the two points are connected. This continues until the area that contains all points is found.

The alpha hull is robust against empty spaces that cannot be considered by the convex hull. The expected area can be determined by regulating the range with the α value.

However, the alpha hull requires a computation time of $O(n \log n)$ for n points. When there are many points, like in a long sequence, the computation time becomes

	Chr1	Chr2	Chr3	Chr4	Chr5
Max_X	233,197	121,923	215,564	126,126	218,859
Max_y	112,716	106,743	87,480	74,728	90,302
Max_Z	18,589,777	14,769,847	19,379,909	20,991,662	17,923,569

Table 4: The result of the maximum value of each axis for human chromosome 1-5 which is computed result for walk plot *R* The maximum values of the axis of x and y are similar to each other, but the z axis has dozens of time as big value as x, y.

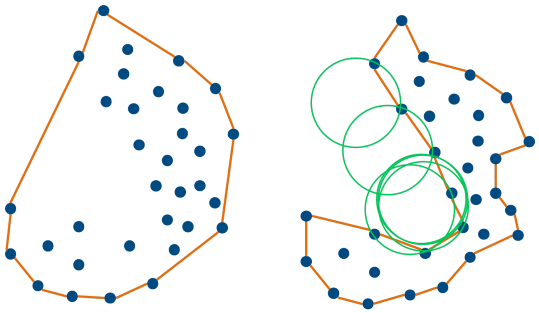


Figure 5: The example of Convex Hull (a) and Alpha-Hull(b). In the case of Convex-Hull (a),Because it uses outer-most points, if there are much big empty space, it can't consider this. Alpha-Hull, on the other hand, it can extract much detailed area using circle with a radius of alpha.

excessive. The alpha hull algorithm also prints edges that connect the two outermost points of the area. This result is not sorted, so additional computation is needed.

4.1 Beta Shape Similarity Algorithm

In this section, we propose a new algorithm for finding the area by comparing the visualization result with the above alpha hull. The alpha hull method generates the result in edge form by comparing the results located in α with each other. However, the beta shape searches the coordinate space and prints the dot-marked coordinate space in matrix form. The beta shape method takes a computation time of $O(n^2)$ to search $n*n$ space because the search space is the visualization of the vertex coordinates. As previously explained about simplifying the method for the walk plot, if there are many vertices and the range of coordinates is limited, the visualization time for the beta shape method is less than that for the alpha hull method. The process to obtain the beta shape is defined as follows:

1. Mark the visualization results with two-dimensional matrix coordinates as $mat[x][y] = 1$.
2. Search the whole coordinate space for the marked vertices.
3. Check if there are other marked vertices within the range of $(i-\beta, j)$ and $(i+\beta, j)$. The location of the marked vertex is found by searching (i, j).
4. If there are other marked vertices, fill the empty space by checking the range of (i, j).

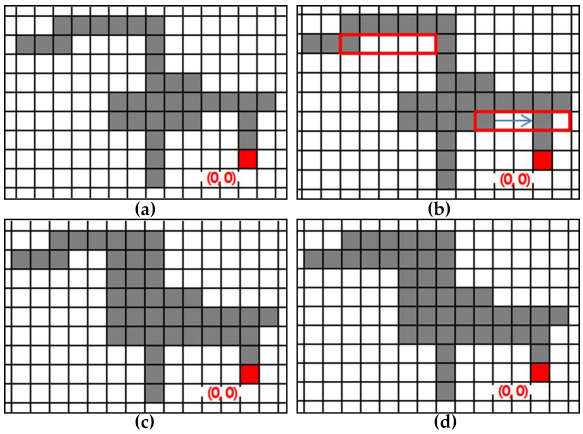


Figure 6: The example of generation process when the beta value is 5. Find the marked point (i, j) on the orthogonal projected walk plot as (a), and find the other marked point in the axis of x, y direction within the limit of β range. If there is another point in the β range, it fills the empty space with marked points as (c). Those process (b), (c) are repeated until there no space to be filled more. Figure (d) is final result for (a) using Beta-shape when $\beta = 5$.

Figure 6 presents the above process. The range of β is searched in the x and y axis directions, as shown in Figure 6(b). The empty space is filled, and coordinates of the area containing the vertices are stored to compute the whole area. Through this process, we can determine the area that includes the visualization result within a computation time of $O(n^2)$.

4.2 Comparing the Similarity of Walk Plots by Using the Beta Shape

Comparing the area on three-dimensional coordinates is a difficult problem, so we computed the similarity by using the results of orthogonal projection for each three walk plots on the two-dimensional planes XY, YZ, and XZ. If the two beta shapes $beta_a$, $beta_b$ are generated from the two base sequences a, b that are orthogonally projected onto a single two-dimensional plane k, we can get a broad outline of similarity by computing their overlapping area. The similarity of the areas $beta_a$ and $beta_a$ is defined in equation (5):

$$Sim_{ab}^{k} = \frac{\sum(beta_a \cap beta_b)}{\sum beta_b} \tag{5}$$

On the left-hand side, $\sum(beta_a \cap beta_b)$ is the size of the overlapping areas of $beta_a$ and $beta_b$. This equation can be used to determine how much of a overlaps with

b. If most of $beta_a$ overlaps with $beta_b$, this may be only a small part of $beta_b$, so the two areas may not be similar to each other. Therefore, to compare the similarity of two areas, we should consider Sim^k_{ab}, Sim^k_{ba}. After considering this point, we defined the similarity between $beta_a$ and $beta_b$ as follows:

$$Sim^k(a,b) = \sqrt{\frac{(Sim^k_{ab})^2 + (Sim^k_{ba})^2}{2}} \quad (6)$$

By using equation 6, we can compute the visualization results on a three-dimensional walk plot and get the similarity rate $Sim(a,b)$ between different two sequences. The similarity rate for each plane can be represented as a vector with three components $(Sim^{XY}(a,b), Sim^{YZ}(a,b), Sim^{XZ}(a,b))$ that are orthogonally projected on the XY, YZ, and XZ planes. By using the sizes of the vector values, we computed the similarity of two sequences with the rate $Sim(a,b)$. The computed results are given below.

$$Sim(A,B) = \sqrt{(Sim^{XY}(a,b))^2 + (Sim^{YZ}(a,b))^2 + (Sim^{XZ}(a,b))^2} \quad (7)$$

5 EXPERIMENT

5.1 Environment and Testing Data

The two main techniques were a visualization model for base sequences of 10 Mbp on a three-dimensional walk plot by using the simplification method and a comparison model based on the beta shape by using the visualization results. Using the simplification method, we could visualize the base sequences of 10 Mbp within a very short time on the three-dimensional walk plot and compare the visualization results of two sequences more intuitively. To represent the results as numerical values, we compared the result of two bases with the beta shape. To visualize the three-dimensional walk plot in real time, we developed a web service system based on Web-Gl Language.

All of the data used for the experiment were obtained from UCSC [10] and NCBI [19]. We obtained the human chromosome data from the Human Genome Project; "homo sapiens gr 37" is offered free by UCSC [10]. The chromosome data of three apes (gorilla, chimpanzee, orangutan) and other species (milk cow, dog, green monkey, chicken, rat, pig) were obtained from the taxonomy in NCBI [19]. The average chromosome length was approximately 30–250 Mbp, and 261 chromosomes were used. The sex chromosomes were too short, so they were not selected for analysis. Table 5 presents the data for this experiment.

5.2 Visualization Results

5.2.1 *Verification of the DNA Visualization Result with the Proposed Method*

We checked if the three-dimensional walk plot visualization results could be used to distinguish the characteristics of real genomes. To judge the utility of the visualization result, we set three standard judgments based on other research about visualization-based walk plots [15]:

- Are the characteristics of the base sequences represented in the three-dimensional walk plot results?
- Is it possible to distinguish the different base sequences from each other by using the visualization results?
- Is it possible to check similar sequences against each other by using the visualization results?

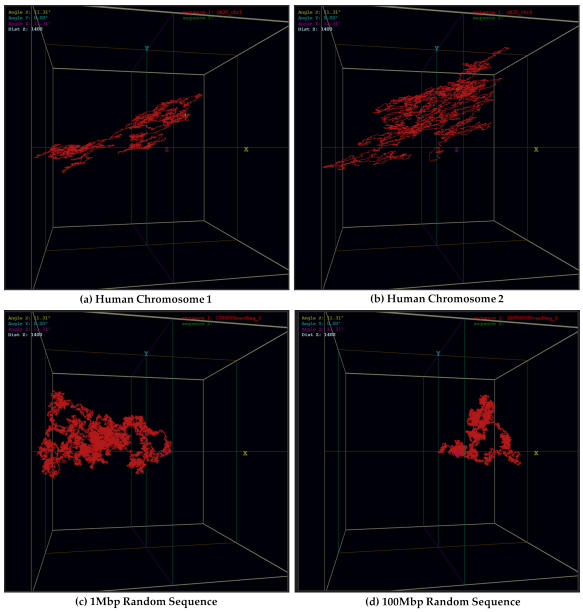


Figure 7: The visualization results of human chromosome(GR37) 1, 2 are (a), (b). And those of randomly generated sequences which has 1Mbp, 100Mbp are (c), (d). The walk plots of random sequences are lumpy generally, but those of human chromosome are scattered. We hereby can figure out that the chromosome data has certain pattern which is distinguished from random sequences.

To confirm the base sequences form the visualization results, we generated the human chromosome (GR37) 1, 2 and random sequences by using the visualization tool, as shown in Figure 7. The random sequences were provide by a random number generator in quantities of 1 and 100 Mbp. We then compared the results of both sequences. The total view size was limited to [-400, 400] in the preprocess for all experiments.

When we compared the human chromosomes 1 and 2, chromosome 1 was spread out densely, but chromosome 2 was distributed widely. Thus, they had different characteristics. For the random sequences created by the random number generator, the results were generally lumpy (Figures 7(c) and (d)), but the results of the human chromosome were not. Thus, we confirmed that the human chromosome has certain rules for genome

sequence ID	scientific name	name	Number of chromosome	notes
s01	Bos taurus	milk cow	38	-
s02	Canis lupus familiaris	dog	38	-
s03	Chlorocebus sabaeus	green monkey	29	-
s04	Gallus gallus	chicken	except 28(32)	29,30,31,32
s05	Gorilla gorilla gorilla	gorilla	23	chromosome 2A, 2B
s06	Homo sapiens	human	22	-
s07	Mus musculus	rat	19	-
s08	Pan troglodytes	chimpanzee	23	chromosome 2A, 2B
s09	Pongo abelii	orangutan	23	chromosome 2A, 2B
s10	Sus scrofa	pig	18	-

Table 5: The table shows the information of DNA Chromosome sequences with each sort which are used for experiment. The information for the mitochondria and X,Y Chromosome which is related with gender is excepted for this experiment.

construction, and those characteristics were revealed by the walk plot.

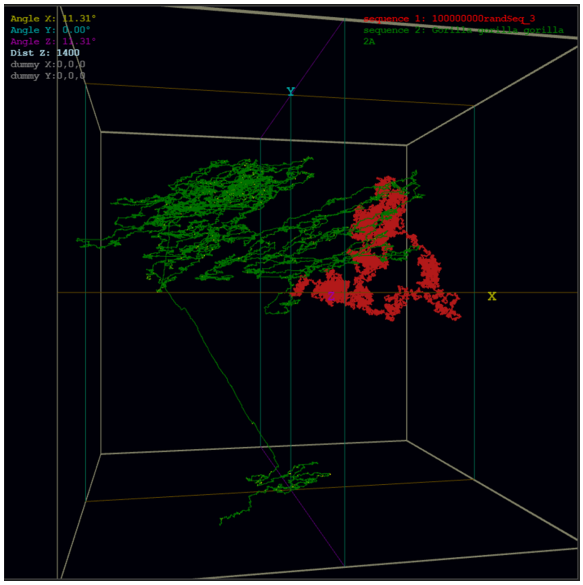
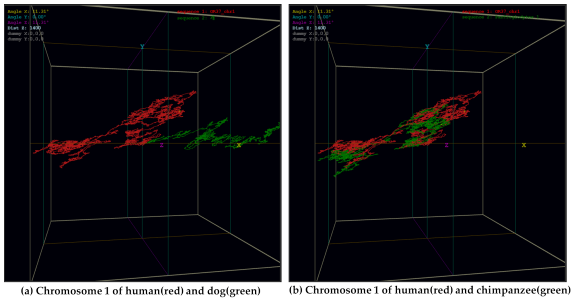


Figure 8: The comparison result of visualization for random sequence which has 100Mbp and gorilla. In case of random sequence, the visualization result is represented with overlapped form as (d). Also the visualization result of random sequence has more lumpy than the gorilla chromosome 2A which has similar size with random sequence. That way, we can confirm that the DNA sequence is not irregular data.

To get more precise results, when we compare the visualization results of two sequences on screen, we considered the scale of the gorilla chromosome 2A with a similar size to that of a random sequence, as shown in Figure 8. For the gorilla chromosome 2A, the whole sequence length was approximately 111 Mbp, which was different from the random sequence by 10%. However, the random sequence was expressed at a point on the coordinate plane in lumpy form, and there were no changes in the walk plot. On the other hand, the gorilla chromosome showed large changes within the data range.

Figure 9 compares the results of human chromosome 1 and chromosome 1 of the dog and chimpanzee. The two chromosomes clearly had different visual results, as shown in Figure 9(a). This result was true not only for the dog but also for all species excluding the apes. On the other hand, the chimpanzee chromosome 1, which is known to have similar genome information as human chromosome 1, had similar visualization results. The chromosomes of other apes such as the gorilla and orangutan also had generally similar results.



(a) Chromosome 1 of human(red) and dog(green) (b) Chromosome 1 of human(red) and chimpanzee(green)

Figure 9: The comparison result of the chromosome 1 of human and dog, chimpanzee. The chromosome of human and dog is quite different as (a), but the chromosome of human and chimpanzee, which is known that more than 90% of chromosome is similar, is also highly similar.

5.2.2 *Verification of the Similarity of Genome Information by Using the Beta Shape*

By using the above three-dimensional walk plot method, as shown in Figure 9, we compared the information of base sequences with different genome information according to the visualization results for validation. The chromosomes of the human and apes, which are known to be quite similar, were also checked and showed only small differences. We computed the similarity of the visualization results of two different sequences by using the beta shape and verified the comparison method for similarity based

on the results. The similarity comparison method was verified according to the following criterion:

- Is it possible to confirm the degree of similarity with numerical values by using the similarity comparison method?

To confirm this criterion, we performed a comparative experiment to check whether the results of the similarity comparison using the beta shape met this requirement. Table 6 presents the computed similarity for each chromosome 1 introduced in Table 5.

In the case of the human (S06), other apes were very similar in the order of S08, S05, and S09, as given in Table 6. In addition, the comparison results were better with the human than with the other species. Figure 10 shows the visualization results of chromosome 1 for the human and gorilla. Figure 10(a) shows a three-dimensional walk plot. Compared with the human, the front part of the gorilla sequence was weighted towards the y axis, but the general forms were very similar. The similarity between the two base sequences could be checked more concretely when their results were orthogonally projected on the (b) XY, (c) XZ, and (d) YZ planes. There was no difference between the two sequence in the orthogonal projection on the XZ plane. For the orthogonal projections in the XY and YZ planes, the random walk was slightly weighted towards the y axis, but the general forms were quite similar. However, the orangutan (S09) was similar to the other apes but also had no difference with the other species.

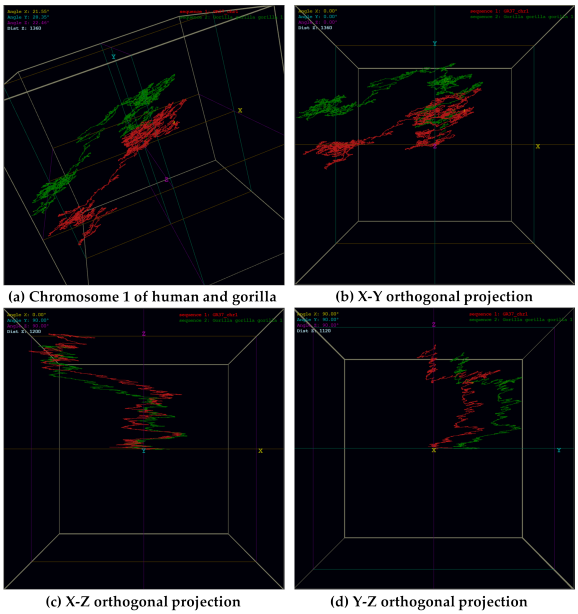


Figure 10: The comparison result of the chromosome 1 of human and gorilla (a). And the result of orthogonal projection on plane of XY(b), XZ(d), YZ(d). Compared with human, in the case of gorilla, the front part of sequence is weighted towards y axis but the general form is very similar to each other. Those characteristics are more revealed at the result of orthogonal projection on 2 dimensional plane.

6 CONCLUSION

The base sequences, which contain the DNA information, are data in character string form. To check such data visually, related research based upon previous visualization methods of two- and three-dimensional walk plots is actively ongoing. Many methods of comparison between two difference sequences have been proposed based on the walk plot characteristics.

This visualization method using base sequences can visualize the sequence in a short amount of time, and the characteristics of the base sequences can be intuitively analyzed from the visualization result without the use of complicated algorithms such as the alignment method. Because previous research usually focused on short base sequences with special functions such as insulin and the β-globin base, these methods are inapplicable to long base sequences of 10 Mbp.

We proposed a new three-dimensional visualization method for long sequences of more than 10 Mbp. The advantages of our proposed visualization method are as follows:

- It can visualize quite long sequence data such as chromosomes.
- If there are no changes to the screen size or sequence data, it can visualize the sequence in a short time by reducing the number of points that are generated with each base point and that need to be visualized through a preprocess.
- It can reveal more specific characteristics than previous visualization methods by using the z axis for bases such as AT (TA) and GC (CG); such information is lost with the WS-curve method.
- It can compare massive amounts of base sequences like a chromosome in a short amount of time through our proposed beta shape comparison.

We expect that our proposed method can be used to grasp the relation between long sequences in a short amount of time without comparison-based alignment, which would reduce the computation cost and contribute to evolutionary research. Because DNA information necessarily contains mutation, deletion, and metastasis, the following requires further study:

- Searching for partial similarity in the visualization results of base sequences.
- Determining high levels of similarity by comparing visualization results at various scales.
- Separating parts with high and low levels of similarity when visually comparing two sequences.

	s01	s02	s03	s04	s05	s06	s07	s08	s09	s10
s01	1.000	0.135	0.187	0.024	0.051	0.130	0.236	0.104	0.164	0.180
s02		1.000	0.190	0.083	0.061	0.129	0.201	0.138	0.163	0.112
s03			1.000	0.032	0.127	0.179	0.220	0.159	0.207	0.162
s04				1.000	0.080	0.073	0.140	0.060	0.171	0.053
s05					1.000	0.345	0.143	0.299	0.192	0.174
s06						1.000	0.181	0.535	0.267	0.214
s07							1.000	0.206	0.198	0.200
s08								1.000	0.249	0.223
s09									1.000	0.166
s10										1.000

Table 6: The comparative result for chromosome 1 of each sort, when $\beta = 10$. The similarity between apes(s05, s08, s09) and human(s06) is relatively high.

7 ACKNOWLEDGEMENT

This research was supported by a grant from Marine Biotechnology Program(PJT200620, Genome Analysis of Marine Organisms and Development of Functional Applications) Funded by Ministry of Oceans and Fisheries, Korea.

8 REFERENCES

[1] A. B. Alencar, F. V. Paulovich, R. Minghim, M. G. Andrade, and M. C. F. Oliveira. Similarity-based visualization of time series collections. *12th Int. Conf. on Information Visualisation*, pages 280–286, 2008.

[2] M. Krstajic, E. Bertini, and D. A. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE Trans. Visualization and Computer Graphics*, 17(12):2432–2439, Dec 2011.

[3] E. Graells and A. Jaimes. Lin-spiration: using a mixture of spiral and linear visualization layouts to explore time series. *Proc. ACM Int. Conf. on Intelligent User Interfaces*, pages 237–240, 2012.

[4] S. Thakur and A. J. Hanson. A 3d visualization of multiple time series on maps. *Proc. IEEE 14th Information Visualization*, pages 336–340, 2010.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[6] B. Langmead, C. Trapnell, M. Pop, and SL Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[7] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[8] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[9] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome Res.*, 12(6):996–1006, 2002.

[10] UCSC. Genome browser. http://genome.ucsc.edu/.

[11] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–402, 2010.

[12] M. A. Kim, E.J. Lee, H. G. Cho, and K. J. Park. A visualization technique for dna walk plot using k-convex. *Proc. Fifth Int. Conf. on Central Europe ComputGraphVisualization*, pages 10–14, 1997.

[13] B. Liao and K. Ding. A 3d graphical representation of dna sequences and its application. *Theor. Comput. Sci.*, 358(1):56–64, 2006.

[14] F. Bai, Y. Liu, and T. Wang. A representation of dna primary sequences by random walk. *Math. Biosci.*, 207(1):282–291, 2007.

[15] N.W. Lo, H. T. Chang, S. W. Xiao, C. H. Li, and C. J. Kuo. Global visualization and comparison of dna sequences by use of three-dimensional trajectories. *J. Inf. Sci. Eng.*, 23(6):1723, 2007.

[16] G. Xie and Z. Mo. Three 3d graphical representations of dna primary sequences based on the classifications of dna bases and their applications. *J. Theor. Biol.*, 269(1):123–130, 2011.

[17] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inf. Theor.*, 29(4):551–559, Jul 1983.

[18] N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E. P. Mucke, , and C. Varela. Alpha shapes: definition and software. *Int. Comput. Geom. Softw. Workshop*, 1995.

[19] NCBI. National center for biotechnology information. http://www.ncbi.nlm.nih.gov/.