Chapter 7

Graphical Representation of Biological Sequences

Satoshi Mizuta

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74795

Abstract

Sequence comparison is one of the most fundamental tasks in bioinformatics. For biological sequence comparison, alignment is the most profitable method when the sequence lengths are not so large. However, as the time complexity of the alignment is the square order of the sequence length, the alignment requires a large amount of computational time for comparison of sequences of large size. Therefore, so-called alignment-free sequence comparison methods are needed for comparison between such as whole genome sequences in practical time. In this chapter, we reviewed the graphical representation of biological sequences, which is one of the major alignment-free sequence comparison methods. The notable effects of weighting during the course of the graphical representation introduced first by the author and co-workers were also mentioned.

Keywords: alignment-free, amino acid sequence, binary image, DNA sequence, mitochondria, phylogeny

1. Introduction

Comparison between biological sequences is one of the most fundamental tasks in the area of bioinformatics. For relatively short sequences, such as nucleotide sequences of genes or amino acid sequences of proteins, *alignment* is the most profitable method for the sequence comparison. However, as the dependency of the computational time of the alignment on the sequence length N is $O(N^2)$, the alignment is hard to be applied to comparison between sequences of large size, such as whole genome sequences. Therefore, the development of *alignment-free* methods is required to analyze the similarities between the sequences of large size in practical time. One of the most actively studied methods of the alignment-free sequence comparison is *graphical representation* [1, 2]. In addition to overcoming the time-consuming

problem mentioned above, the graphical representation has the advantage that the similarities between sequences can be easily noticed visually.

Since the seminal paper by Hamori and Ruskin [3] was published, various kinds of sequence comparison methods based on the graphical representation have been proposed by many researchers. The basic procedure of the graphical representation is outlined as follows: first, each character in a biological sequence, which is expressed by the four-letter alphabet for nucleotide sequences and the 20-letter alphabet for amino acid sequences, is expressed by individual vectors in a certain dimensional space; next, the vectors are connected successively in a head-to-tail fashion, drawing a curve, or a *graph*, in the expression space; and last, if necessary, the distances between the graphs are calculated based on the predefined distance measures.

In this chapter, we briefly review the graphical representation methods for biological sequence comparison. In addition, we introduce our work recently published, in which weighting during the course of the graphical representation shows the notable effects.

2. Variations of graphical representations

The graphical representation methods are classified into some classes according to the target sequences and the dimension of the representation space. The target sequences of the graphical representation are amino acid sequences of proteins and nucleotide sequences of DNA (or RNA), including specific genes, mitochondrial genomes, and others. **Table 1** summarizes the classification of the graphical representation methods published so far.

2.1. Graphical representation of DNA sequences

Biological sequences stored in data archives are expressed by the four-letter alphabet for nucleotide sequences of DNA and the 20-letter alphabet for amino acid sequences of proteins.

Target sequence	Dimension of expression space	Work
DNA sequences		
Specific genes	2D	[4–22]
	3D ≤	[23–36]
Mitochondrial genomes	2D	[37–41]
	3D ≤	[42]
Others	3D ≤	[3, 43]
Proteins		
	2D	[44–49]
	3D ≤	[50–53]

Table 1. Classification of graphical representation methods.

To represent the biological sequences by graphs, it is necessary to express each character composing the sequences in numerical form.

The most popular strategy for the numerical expression is assigning vectors to respective characters in the alphabet. As for nucleotide sequences of DNA, the individual vectors of two, three, or higher dimension are assigned to four types of bases, A, T, G, and C.

2.1.1. Two-dimensional representation

Figure 1 is the two-dimensional vector assignment utilized by Gates [4]. Although many variations of the assignments are given according to the layout of the four bases, the number of the independent assignments is reduced to 3!/2 = 3, when the assignments that are transformed to each other by rotation on the *xy*-plane or inversion with respect to the *x*- or *y*-axis are assumed to be equivalent. The assignments of this type including the variations with some modifications are utilized in Refs. [5, 6, 10, 16, 20, 21, 40, 41].

By connecting the vectors successively in a head-to-tail fashion according to each base appearing in a nucleotide sequence, a graphical representation is generated. **Figure 2** shows, as an example, the graphical representation of sequence "TGAGTTC" generated by Gates' assignment.

The assignment of **Figure 1** may draw circuits in the graphical representation, leading to the loss of information that the original biological sequence has. To get rid of the degeneracy, Yau [9] introduced the assignment shown in **Figure 3**, which makes no circuit in the graphical representation; because the *x*-components of the vectors have all positive values, no backward motion along the *x*-axis exists in the graphical representation. For comparison, **Figure 4** illustrates the graphical representations of the first exon of the human β-globin gene represented by Gates' vector assignment (**Figure 1**) and Yau's vector assignment (**Figure 3**). There are many circuits in the graph by Gates' assignment; on the other hand, there is no circuit in the graph by

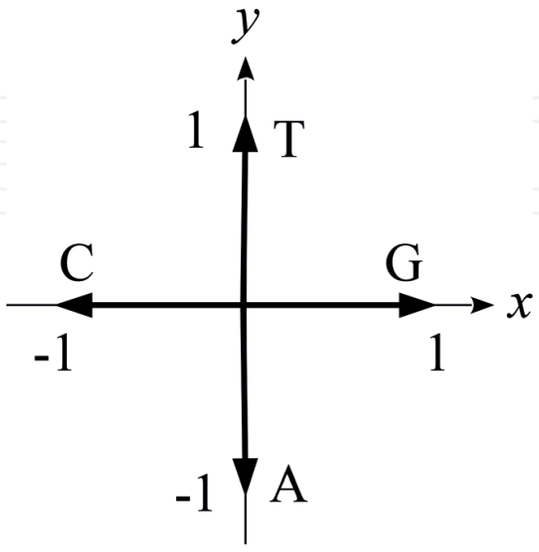

Figure 1. Two-dimensional vector assignment to bases utilized by Gates [4].

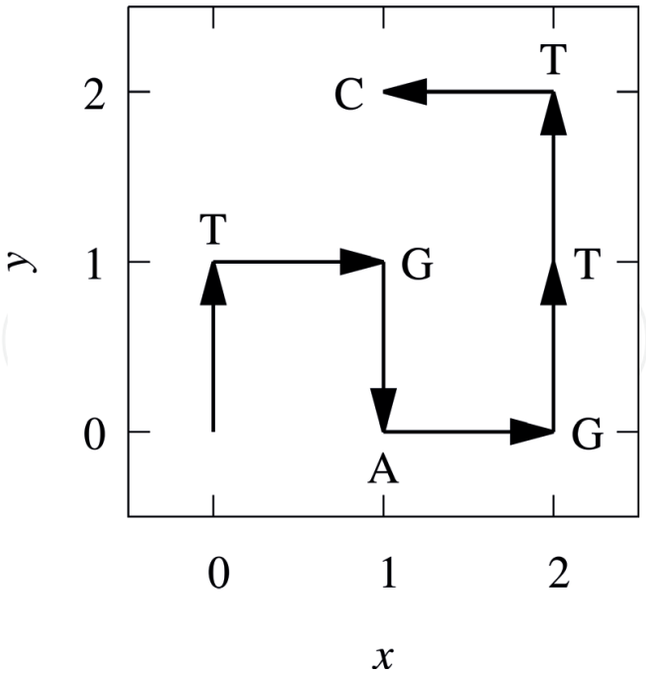

Figure 2. Graphical representation of sequence "TGAGTTC" generated by Gates' assignment (**Figure 1**).

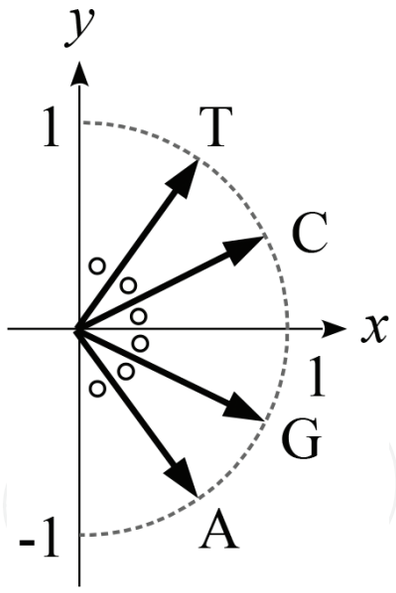

Figure 3. Two-dimensional vector assignment to bases utilized by Yau [9].

Yau's assignment. The assignments of Yau's type (including the variations with some modifications) are utilized in Refs. [9, 12, 15, 18, 19, 37–39].

Some researchers used another approach; they directly mapped bases on the *xy*-plane without vector assignment. Randić et al. plotted the *i*th base of a DNA sequence on the *xy*-plane at (*i*,0), (*i*,1), (*i*,2), and (*i*,3) for bases C, G, T, and A, respectively [7]. By connecting the plots, a *zigzag curve* is given. **Figure 5** demonstrates the zigzag curve for sequence "ATGGTGCACC" given

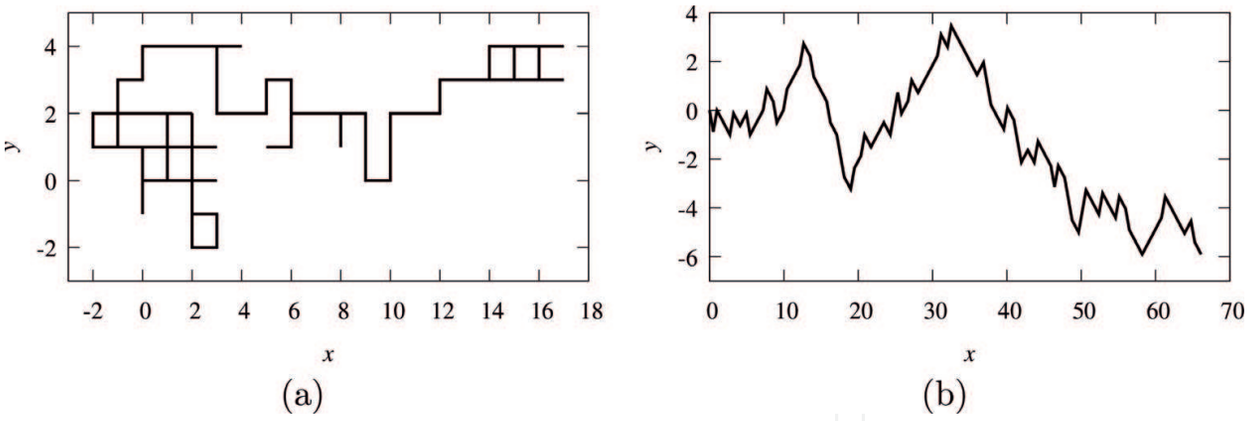

Figure 4. Graphical representations of the first exon of the human β-globin gene (GenBank: AF527577) represented by (a) Gates' vector assignment (**Figure 1**) and (b) Yau's vector assignment (**Figure 3**).

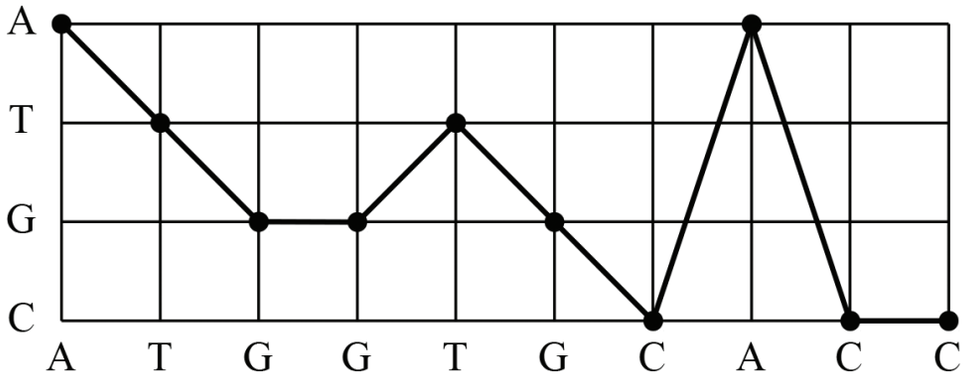

Figure 5. Zigzag curve for sequence "ATGGTGCACC" given by Randić's approach [7].

by Randić's approach. Similar to the graphical representation given by Yau's vector assignment (**Figure 4(b)**), the zigzag curve has no circuits. The approaches of this kind (including the variations with some modifications) are utilized in Refs. [8, 11, 13, 14, 17, 22].

2.1.2. Three-dimensional representation

Hamori and Ruskin [3] used a three-dimensional vector assignment to bases (**Figure 6**). Gates' approach (**Figure 1**) [4] is a simplified version of this assignment. However, unlike Gates' approach, Hamori's assignment does not make any circuit in the resultant curve (called *H-curve*), because the *z*-coordinate of the curve decreases monotonically with the positions of the bases in the original sequence. The assignments of this type (including the variations with some modifications) are utilized in Refs. [26, 27, 29, 31–36].

Zhang and Zhang [43] used another three-dimensional vector assignment shown in **Figure 7**. The resultant curve, called *Z-curve*, may have circuits therein like the curves generated by Gates' vector assignment (**Figure 1**). The assignments of this type (including the variations with some modifications) are utilized in Refs. [24, 42].

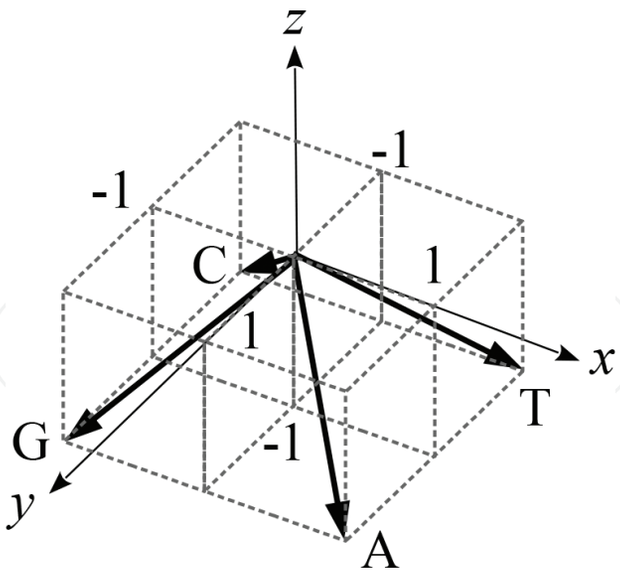

Figure 6. Three-dimensional vector assignment to four bases utilized by Hamori [3].

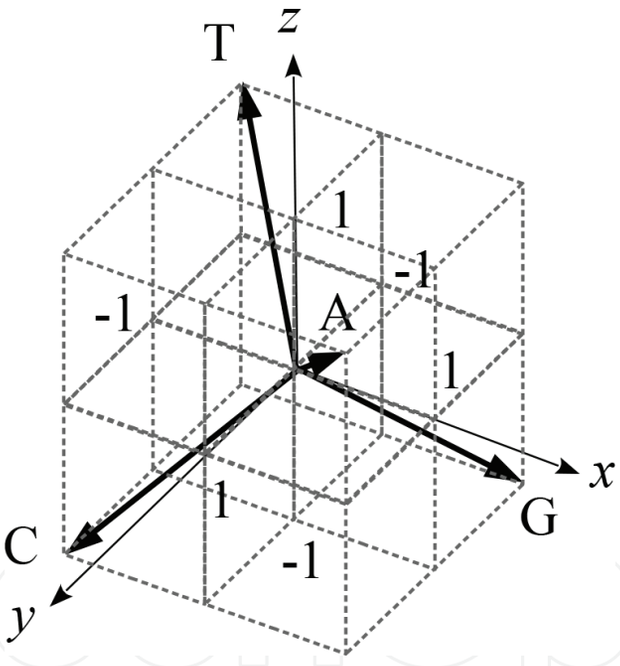

Figure 7. Three-dimensional vector assignment to four bases utilized by Zhang and Zhang [43].

2.1.3. Higher than three dimensions

The graphical representations in the space of higher than three dimensions cannot be visualized directly. Instead of direct visualization, they are expressed abstractly or projected on some spaces of lower dimensions. The approaches of this type (including the variations with some modifications) are utilized in Refs. [25, 30, 28].

http://dx.doi.org/10.5772/intechopen.74795

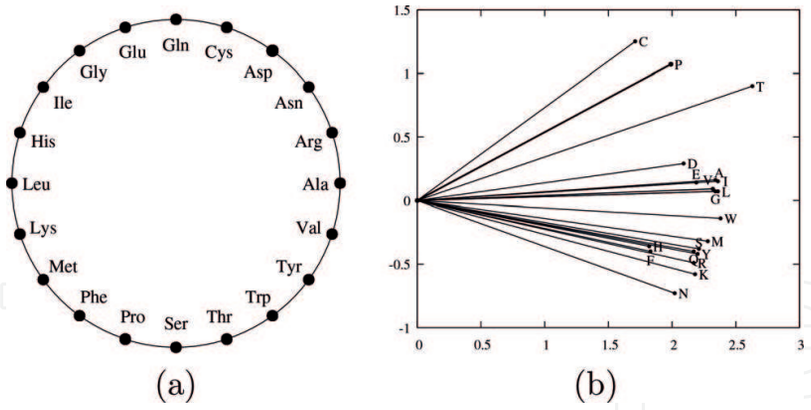

Figure 8. Two-dimensional vector assignments to 20 amino acids utilized by (a) Randić [44] and (b) Wen [45]. The 20 amino acids are indicated by three-letter codes and single-letter codes, respectively.

2.2. Graphical representation of proteins

A general strategy for graphical representation of protein sequences is common to that for DNA sequences, namely, numerical expression of characters followed by mapping on certain dimensional spaces, except for the fact that the number of character types is 20 instead of 4 for DNA sequences. A detailed review of graphical representation of protein sequences is given by Randić et al. [54]. Here, we briefly mention the variations of the graphical representation scheme of proteins.

Figure 8(a) and **(b)** presents two-dimensional vector assignments to 20 amino acids utilized by Randić et al. [44] and Wen and Zhang [45], respectively. In Randić's assignment, the 20 amino acids (indicated by three-letter codes) are arranged uniformly on a unit circle in alphabetical order. On the other hand, in Wen's assignment, the horizontal and the vertical coordinates of the vectors are given by pK_a values of COOH and NH_3^+ of the corresponding amino acid, respectively. The assignments of Randić's type and Wen's type (including the variations with some modifications) are utilized in Refs. [47] and [46, 48], respectively.

Yu and Huang [49] directly mapped 20 amino acids on a two-dimensional space and drew zigzag curves similar to the curve for the case of DNA sequences given by Randić's approach (**Figure 5**) [7].

He et al. [52] extended Randić's vector assignment (**Figure 8(a)**) to three dimensions by adding one extra coordinate corresponding to the position of the amino acid in the original sequence, with the modification of the arrangement of the 20 amino acids on the unit circle based on the 6-bit binary gray code assigned by the codon structure of the amino acids.

3. Numerical characterization of graphical representations

As well as the visual evaluation of the similarities between biological sequences through their graphical representations, the quantitative estimation of the similarities also can be done by the numerical characterization of the graphs. The general method of the quantitative

estimation is to construct *feature vector*s based on the various kinds of characteristics of the graphs and, then, to calculate the distances between the feature vectors based on some sort of distance measures.

For the numerical characterization, there are two kinds of methods: geometrical methods and graph-theoretical ones [2].

3.1. Geometrical characterization

The most simple method of the geometrical characterization was proposed by Raychaudhury and Nandy [55], in which the graphs are numerically characterized by their geometrical centers. Let (x_i, y_i) be the coordinate of the ith point of the graph, and then the geometrical center $\left(\overline{\mu}_x, \overline{\mu}_y\right)$ is computed by $\overline{\mu}_x = 1/N \sum_{i=1}^{N} x_i$ and $\overline{\mu}_y = 1/N \sum_{i=1}^{N} y_i$, where N is the total number of the points on the graph. The similarity/dissimilarity between the graphs of sequences, A and B, is measured by the Euclidean distance between their geometrical centers by

$$d_{AB} = \sqrt{\left(\overline{\mu}_x^A - \overline{\mu}_x^B\right)^2 + \left(\overline{\mu}_y^A - \overline{\mu}_y^B\right)^2}, \tag{1}$$

where A and B refer to the corresponding sequences.

A more accomplished geometrical characterization was proposed by Liao et al. [37], in which they constructed a two-component feature vector based on the 2×2 covariance matrix CM calculated from the two-dimensional graph by

$$CM = \begin{pmatrix} 1/N \sum_{i=1}^{N} \left(x_i - \overline{\mu}_x\right)^2 & 1/N \sum_{i=1}^{N} \left(x_i - \overline{\mu}_x\right)\left(y_i - \overline{\mu}_y\right) \\ 1/N \sum_{i=1}^{N} \left(y_i - \overline{\mu}_y\right)\left(x_i - \overline{\mu}_x\right) & 1/N \sum_{i=1}^{N} \left(y_i - \overline{\mu}_y\right)^2 \end{pmatrix}. \tag{2}$$

The two-component vector is given by the two eigenvalues of CM, λ_1, and λ_2, as (λ_1, λ_2). The similarity/dissimilarity between the graphs is measured by the Euclidean distance between the end points of their feature vectors.

The approach proposed by Qi et al. [18] is another example of the geometrical characterization. They constructed an eight-component feature vector from the averages of the y-coordinates of the eight different patterns of the two-dimensional graphical representations. The similarity/dissimilarity between the graphs is measured by the Euclidean distance between the end points of their feature vectors.

3.2. Graph-theoretical characterization

The graph-theoretical characterization that is most widely used is the method based on the D/D (distance/distance) matrix [56]. The off-diagonal (i, j) elements of the D/D matrix are defined as the quotient of the Euclidean distance between the ith and the jth vertices of the graph and the

graph-theoretical distance between the two vertices. The D/D matrix is symmetric, and all the diagonal elements are zero by definition.

There are two variations of the D/D matrix. If the denominator (the graph-theoretical distance) is replaced by the sum of the geometrical lengths of the edges between the two vertices, the D/D matrix is denoted as the L/L matrix; if the denominator is replaced by the total number of the edges between the two vertices, the D/D matrix is denoted as the M/M matrix.

As an example, **Table 2** demonstrates the upper off-diagonal elements of the L/L matrix calculated for the graph of sequence "TGAGTTC" in **Figure 2**.

The feature vectors are constructed from the leading eigenvalues of the D/D matrix, which are the invariants of the matrix and can well describe the characteristics of the individual graphs. For example, Randić et al. [8] used 12-component vectors given by the first leading eigenvalues of the L/L matrices calculated from the 12 essentially different patterns of the graphical representations, and Liao and Wang [13] used three-component vectors constructed by the similar manner.

The similarity/dissimilarity between the sequences, A and B, is measured by the Euclidean distance between the end points of the corresponding feature vectors by

$$d_{AB} = \sqrt{\sum_{i=1}^{K} \left(\lambda_i^A - \lambda_i^B\right)^2}, \tag{3}$$

or the cosine of the angle between the feature vectors by

$$C_{AB} = \frac{\sum_{i=1}^{K} \lambda_i^A \cdot \lambda_i^B}{\sqrt{\sum_{i=1}^{K} \left(\lambda_i^A\right)^2 \cdot \sum_{i=1}^{K} \left(\lambda_i^B\right)^2}}, \tag{4}$$

where λ_i^A and λ_i^B are the ith components of the K-component feature vectors of the sequence A and B, respectively.

	G	**A**	**G**	**T**	**T**	**C**
T	1/1	$\sqrt{2}/2$	$\sqrt{5}/3$	2/4	$\sqrt{5}/5$	$\sqrt{2}/6$
G		1/1	$\sqrt{2}/2$	1/3	$\sqrt{2}/4$	1/5
A			1/1	$\sqrt{2}/2$	$\sqrt{5}/3$	2/4
G				1/1	2/2	$\sqrt{5}/3$
T					1/1	$\sqrt{2}/2$
T						1/1

Table 2. The upper off-diagonal elements of the L/L matrix for the graph in **Figure 2**.

4. Graphical representation based on binary images

The author and co-workers recently published the paper about a novel two-dimensional graphical representation of DNA sequences based on binary images [41]. In this section, we introduce our method and demonstrate the notable effects of *weighting* for the construction of the graphical representations introduced first by the author and co-workers [40].

4.1. Vector assignment

We used the two-dimensional vector assignment to four bases shown in **Figure 9**, which is a modified version of Gates' assignment (**Figure 1**). We located both G and C on the same side so that the GC-contents of the target sequences can be represented on the graphs; the graphs for the sequences with high GC-contents tend to grow in the downward direction, although the tendency is not rigid due to the weighting mentioned below.

4.2. Introducing weighting

In order to extract potential information conveyed by individual bases in DNA sequences, we introduced *weighting* into the process of generating graphical representations; we calculated the weighting factors based on a Markov chain model and multiplied them to the vectors assigned to the bases. As the weighting factors, we used self-information, which is the amount of information that we will receive when a certain event occurs [42]. The self-information is defined by

$$I(E) = -\log P(E), \tag{5}$$

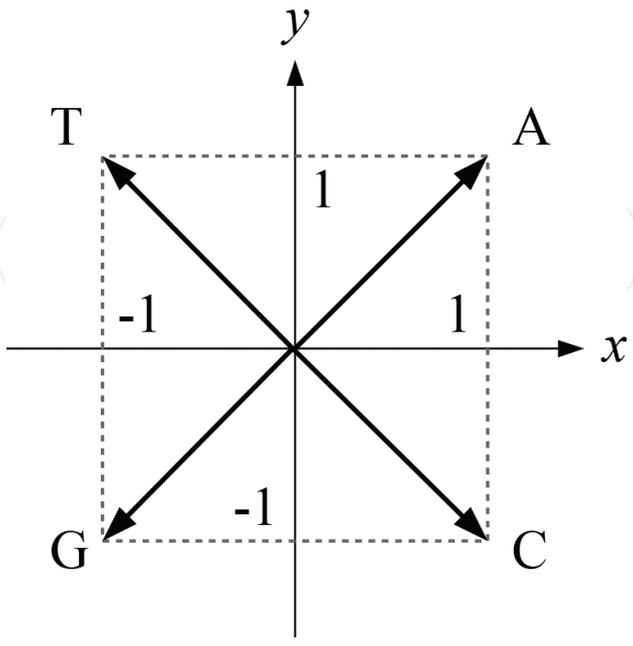

Figure 9. Two-dimensional vector assignment to four bases utilized by Kobori and Mizuta [41].

where $P(E)$ is the probability that event E occurs. We employed the conditional probability calculated based on the second-order Markov chain as $P(E)$ concerning about *codons*, which are triplets of bases in the coding regions of DNA sequences.

The conditional probability is calculated from the appearance frequencies of triplets of bases. For example, the probability that base A occurs after a pair of bases TC is given by

$$P(\mathrm{A}|\mathrm{TC}) = \frac{f(\mathrm{TCA})}{f(\mathrm{TCA}) + f(\mathrm{TCT}) + f(\mathrm{TCG}) + f(\mathrm{TCC})}, \tag{6}$$

where $f(S)$ is the number of appearances of triplet S. For the other combinations of bases, the conditional probabilities are calculated by a similar manner. The numbers of appearances of triplets were measured in all the DNA sequences analyzed.

Table 3 lists the weighting factors calculated with base 4 of the logarithm in Eq. (5) from 31 mammalian mitochondrial genomes. The weighting factor lower than 1.00 indicates that the pair of bases on the row tends to be followed by the base on the column, and on the other hand, the weighting factor higher than 1.00 indicates that, after the pair of bases on the row, the base on the column is hard to appear.

Let us illustrate the procedure of the graphical representation with weighting factors by a simple example. **Figure 10(a)** and **(b)** shows the graphical representations of sequence "ACATATG" by Kobori's vector assignment (**Figure 9**) without and with weighting, respectively. The weighting is not applied to the first two bases, because the weighting factors are not given for the first two bases by our weighting scheme. The weighting factors for the subsequent bases A, T, A, T, and G are 0.83, 0.90, 0.84, 0.92, and 1.42, respectively (see **Table 3**). The vectors for the bases are multiplied by the corresponding weighting factors. As a result, the graphical representation in **Figure 10(a)** is modified as shown in **Figure 10(b)**.

We demonstrate the notable effects of the weighting on the graphical representations by the real sequences. **Figure 11** depicts the graphical representations of three mammalian mitochondrial genomes without weighting and with weighting. Comparing the graphs with weighting (lower row) to the graphs without weighting (upper row), it can be recognized that the characteristics of the graphs are emphasized by the weighting and the individual species can be easily distinguished.

4.3. Generating binary images

A binary image is defined as a digitized image composed of the pixels with two possible values (e.g., 0 and 1), which are typically assigned by *white* and *black*, respectively, on the image. From the graphical representation, a binary image is generated in the following ways: if the pixels include at least a portion of a curve of the graphical representation, they are assigned 1; otherwise, they are assigned 0.

Preceding pair of bases	Third base			
	A	T	G	C
AA	0.82	0.92	1.47	0.95
AT	0.84	0.90	1.42	0.97
AG	1.04	1.11	1.11	0.79
AC	0.83	0.88	1.64	0.90
TA	0.86	0.93	1.28	1.01
TT	0.77	0.97	1.51	0.94
TG	0.73	1.14	1.16	1.06
TC	0.77	0.93	1.69	0.91
GA	0.79	1.08	1.15	1.03
GT	0.67	1.04	1.36	1.11
GG	0.79	1.14	1.22	0.93
GC	0.85	0.93	1.97	0.75
CA	0.84	0.90	1.44	0.96
CT	0.68	1.02	1.51	1.02
CG	0.91	1.00	1.22	0.91
CC	0.90	0.82	1.79	0.85

The 31 mammalian species are (with the accession numbers in the parentheses), human (V00662), pygmy chimpanzee (D38116), common chimpanzee (D38113), gorilla (D38114), gibbon (X99256), baboon (Y18001), Bornean orangutan (D38115), African green monkey (AY863426), cat (U20753), dog (U96639), wolf (EU442884), pig (AJ002189), sheep (AF010406), cow (V00654), buffalo (AY488491), tiger (EF551003), leopard (EF551002), Indian rhinoceros (X97336), white rhinoceros (Y07726), harbor seal (X63726), gray seal (X72004), African elephant (AJ224821), Asiatic elephant (DQ316068), black bear (DQ402478), brown bear (AF303110), polar bear (AF303111), rabbit (AJ001588), hedgehog (X88898), Norway rat (X14848), vole (AF348082), and squirrel (AJ238588).

Table 3. Weighting factors calculated from 31 mammalian mitochondrial genomes.

4.4. Numerical characterization by local pattern histograms

In this work, each graph is characterized by the frequency distributions of *local patterns* that appear on the graph. A local pattern is defined here as a small bitmap image of a certain size. Because each pixel of a binary image takes two variations (0 and 1), the number of the local patterns is 2^n, where n is the number of the pixels of the local pattern. Local patterns of large size are dominated by white pixels, while, on the other hand, those of small size do not have enough variations to express the characteristics of the local area of a graph. For this study, therefore, we chose 3×3 as the size of the local patterns (the number of the local patterns is 2^9=512). **Figure 12** shows the examples of the local patterns of window size 3×3. Excluding the pattern of which the pixels are all white, we construct a feature vector, or a *local pattern histogram*, of dimension 511 for each graph from the appearance frequencies of the local patterns on the graph.

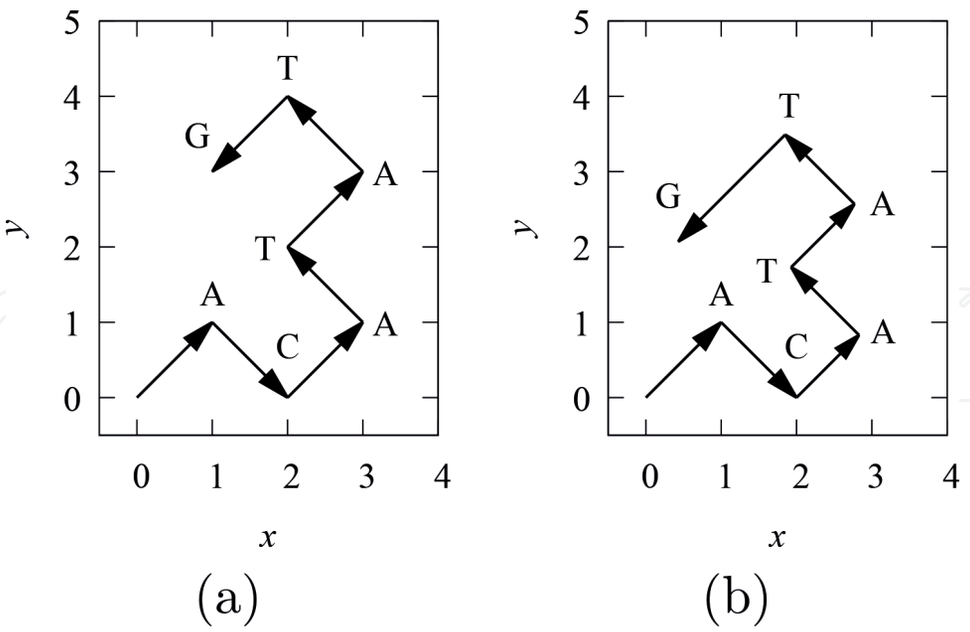

Figure 10. Graphical representation of sequence "ACATATG" by Kobori's vector assignment (**Figure 9**) without weighting (a) and with weighting (b).

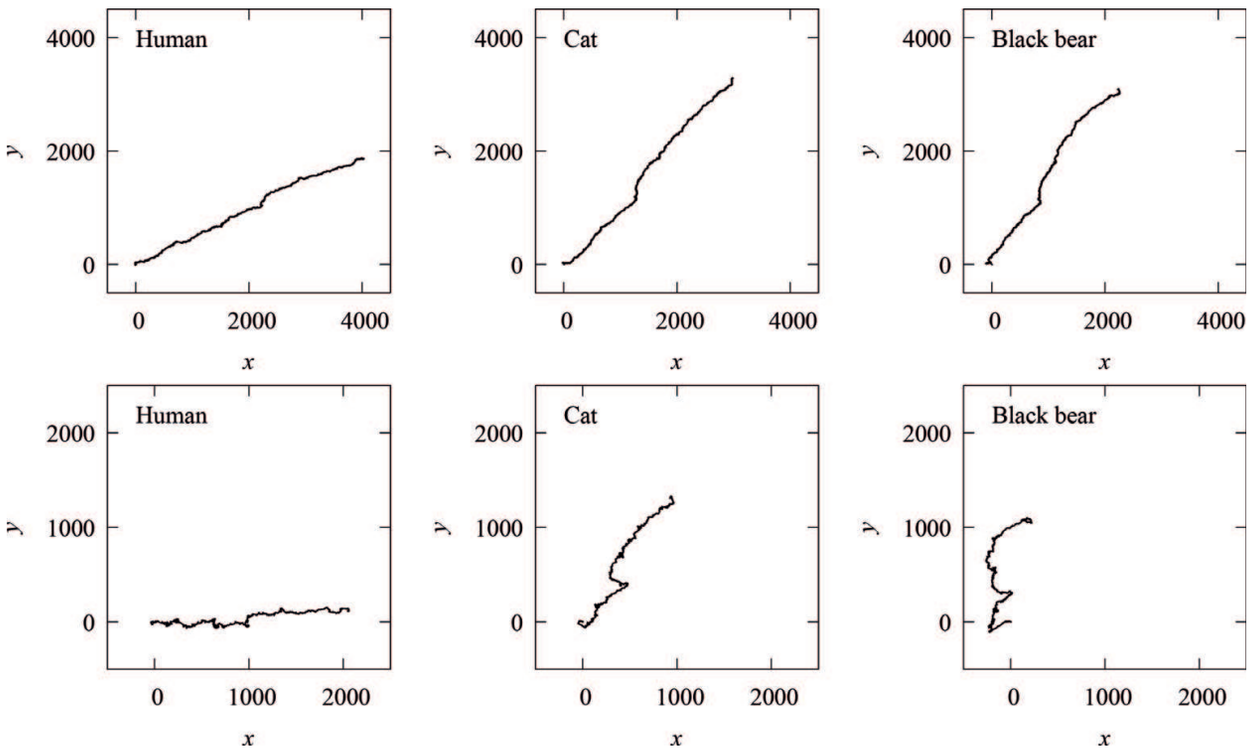

Figure 11. Graphical representations of mitochondrial genomes of three mammalian species without weighting (upper row) and with weighting (lower row). The arrow heads of the vectors are eliminated.

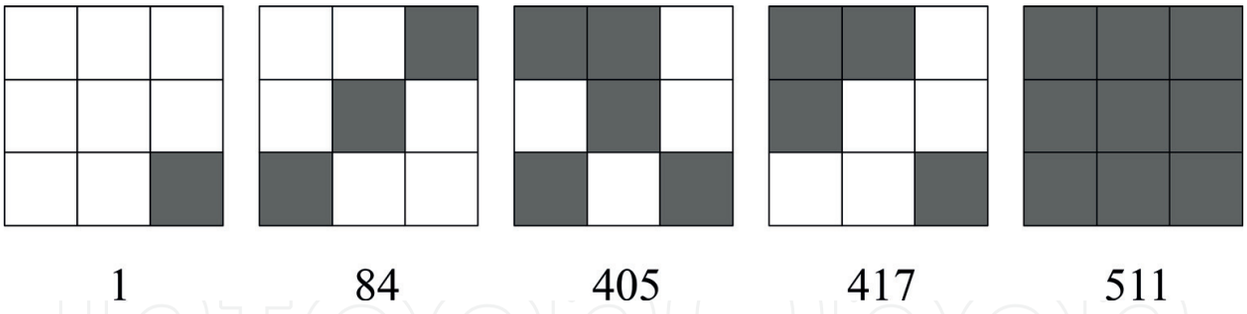

Figure 12. Examples of local patterns. The numbers below each local pattern are the serial numbers assigned to the local patterns.

4.5. Distance measures between local pattern histograms

There are several measures to estimate similarity/dissimilarity between two histograms. Here, we briefly mention five frequently used methods. In the following formulas, K is the number of the local patterns ($K = 511$), and p_i and q_i are the normalized appearance frequencies of the local pattern of serial number i in histograms P and Q, respectively $\left(\sum_{i=1}^{K} p_i = \sum_{i=1}^{K} q_i = 1\right)$.

4.5.1. Histogram intersection

Histogram intersection was proposed by Swain et al. [57] for color indexing of images, which is defined as

$$HI(P,Q) = \sum_{i=1}^{K} \min\ (p_i, q_i). \tag{7}$$

It ranges from 0 to 1, with 1 for P and Q being identical. It is converted to a distance by $D_{HI}(P,Q) = 1 - HI(P,Q)$.

4.5.2. Manhattan distance

Manhattan distance, also known as *city block distance* or L_1-norm, is defined as

$$D_{MD}(P,Q) = \sum_{i=1}^{K} |p_i - q_i|, \tag{8}$$

which ranges from 0 to 2, with 0 for P and Q being identical.

4.5.3. Bhattacharyya distance

Bhattacharyya distance [58] is defined between two probability distributions from a divergence

$$BD(P,Q) = \sum_{i=1}^{K} \sqrt{p_i q_i}, \tag{9}$$

which ranges from 0 to 1, with 1 for P and Q being identical. The Bhattacharyya distance is defined from the divergence by $D_{BD}(P,Q) = -\ln BD(P,Q)$.

4.5.4. Jensen-Shannon divergence

Jensen-Shannon divergence [59] is a symmetrized and smoothed version of Kullback–Leibler divergence [60], which is defined as

$$D_{JS}(P,Q) = \frac{1}{2}KL(P,M) + \frac{1}{2}KL(Q,M), \tag{10}$$

where $M = (P+Q)/2$ and $KL(\cdot, M)$ is the Kullback-Leibler divergence calculated by

$$KL(P,M) = \sum_{i=1}^{K} p_i \log_2 \frac{p_i}{m_i}, \tag{11}$$

$$KL(Q,M) = \sum_{i=1}^{K} q_i \log_2 \frac{q_i}{m_i}. \tag{12}$$

Here, $m_i = (p_i + q_i)/2$. Note that the local patterns having $p_i = p_i = 0$ are excluded from the calculation. The Jensen-Shannon divergence ranges from 0 to 1, with 0 for P and Q being identical.

4.5.5. Kendall's rank correlation coefficient

Kendall's rank correlation coefficient [61], also known as Kendall's τ, is defined as

$$\tau = \frac{X - Y}{\sqrt{X + Y + r}\sqrt{X + Y + s}}, \tag{13}$$

where X is the number of *concordant* $i, j (i > j)$ pairs, which are the i, j pairs that satisfy $\left(p_i - p_j\right)\left(q_i - q_j\right) > 0$; Y is the number of *discordant* pairs, which are the i,j pairs that satisfy $\left(p_i - p_j\right)\left(q_i - q_j\right) < 0$; r is the number of one kind of *tie* pairs, which are the i,j pairs that satisfy $p_i = p_j$ and $q_i \neq q_j$; and s is the number of the other kind of *tie* pairs, which are the i, j pairs that satisfy $p_i \neq p_j$ and $q_i = q_j$. The i, j pairs that satisfy both $p_i = p_j$ and $q_i = q_j$ are excluded from the calculation. Kendall's τ lies between -1 and 1, with 1 for the rank orders of $p_i s$ and $q_i s$ being completely in agreement with each other and with -1 for them being completely reversal with each other. The Kendall's τ is rescaled by

$$D_\tau(P,Q) = 1 - \frac{\tau + 1}{2}, \tag{14}$$

so that $D_\tau(P,Q)$ ranges from 0 to 1, with 0 for the rank orders of P and Q being identical.

4.6. Reconstruction of phylogenetic tree

Among the five distance measures mentioned above, histogram intersection and Manhattan distance showed the best performance. **Figure 13** shows the phylogenetic tree of 31 mammalian species reconstructed by our method using Unweighted Pair Group Method with

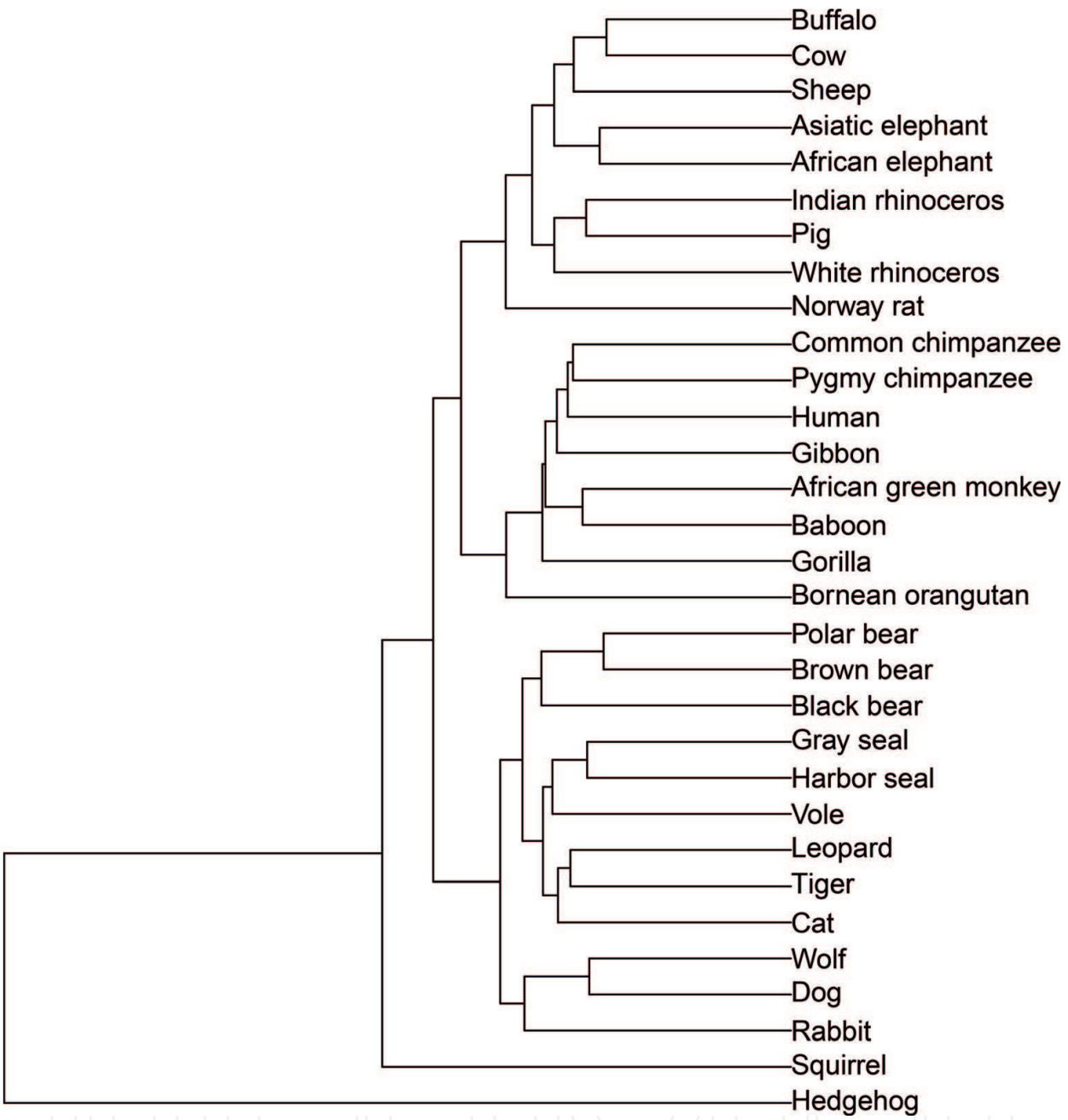

Figure 13. Phylogenetic tree of 31 mammalian species reconstructed by Kobori's method [41] using UPGMA based on the histogram intersection distance measure. The tree is generated by statistical analysis software R with package "ape".

Arithmetic mean (UPGMA) with the histogram intersection distance measure. The same tree is given by Manhattan distance.

5. Conclusion

With the rapid growth of the data size in the archives of biological sequences, the demand for the alignment-free sequence comparison methods is increasing. Graphical representation is

one of the major alignment-free sequence comparison methods. In addition to the visual discrimination abilities of the sequences, the graphical representation has an advantage of requiring only small computational time. The similarity/dissimilarity between a pair of sequences is calculated from the feature vectors constructed based on the graphical representation. The time complexity of the calculation is estimated to be $O(K)$, where K is the dimension of the feature vector and K is usually independent of the sequence length (except for a few methods). Even though the computational time to make a graph, and to construct a feature vector from the graph, may depend on the sequence length N, typically $O(N)$, the construction of the graph and the feature vector is needed to be done for each sequence only once. Thus, the time complexity of the sequence comparison based on the graphical representation is regarded as $O(K)$, which is much less than that of the alignment, $O(N^2)$. From the above considerations, the graphical representation is expected to stay in the main stream of the alignment-free sequence comparison methods from now on, too.

Author details

Satoshi Mizuta

Address all correspondence to: slmizu@hirosaki-u.ac.jp

Graduate School of Science and Technology, Hirosaki University, Hirosaki, Aomori, Japan

References

[1] Roy A, Raychaudhury C, Nandy A. Novel techniques of graphical representation and analysis of DNA sequences—A review. Journal of Biosciences. 1998;**23**(1):55-71. DOI: 10.1007/BF02728525

[2] Nandy A, Harle M, Basak SC. Mathematical descriptors of DNA sequences: development and applications. ARKIVOC. 2006;**2006**(9):211-238. DOI: 10.3998/ark.5550190.0007.907

[3] Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. Journal of Biological Chemistry. 1983;**258**(2):1318-1327

[4] Gates MA. Simpler DNA sequence representations. Nature. 1985;**316**(6025):219. DOI: 10.1038/316219a0

[5] Nandy A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Current Science. 1994;**66**:309-314

[6] Leong PM, Morgenthaler S. Random walk and gap plots of DNA sequences. Bioinformatics. 1995;**11**(5):503-507. DOI: 10.1093/bioinformatics/11.5.503

[7] Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2003;**368**(1–2): 1-6. DOI: 10.1016/S0009-2614(02)01784-0

[8] Randić M, Vračko M, Lerš N, Plavšić D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chemical Physics Letters. 2003; **371**(1–2):202-207. DOI: 10.1016/S0009-2614(03)00244-6

[9] Yau SST, Wang J, Niknejad A, Lu C, Jin N, Ho YK. DNA sequence representation without degeneracy. Nucleic Acids Research. 2003;**31**(12):3078-3080. DOI: 10.1093/nar/gkg432

[10] Liu Y, Guo X, Xu J, Pan L, Wang S. Some notes on 2-D graphical representation of DNA sequence. Journal of Chemical Information and Modeling. 2002;**42**(3):529-533. DOI: 10.1021/ci010017g

[11] Liu XQ, Dai Q, Xiu Z, Wang T. PNN-curve: A new 2D graphical representation of DNA sequences and its application. Journal of Theoretical Biology. 2006;**243**(4):555-561. DOI: 10.1016/j.jtbi.2006.07.018

[12] Wu Y, Liew AWC, Yan H, Yang M. DB-curve: A novel 2D method of DNA sequence visualization and representation. Chemical Physics Letters. 2003;**367**(1–2):170-176. DOI: 10.1016/S0009-2614(02)01684-6

[13] Liao B, Wang TM. New 2D graphical representation of DNA sequences. Journal of Computational Chemistry. 2004;**25**(11):1364-1368. DOI: 10.1002/jcc.20060

[14] Song J, Tang H. A new 2-D graphical representation of DNA sequences and their numerical characterization. Journal of Biochemical and Biophysical Methods. 2005;**63**(3):228-239. DOI: 10.1016/j.jbbm.2005.04.004

[15] Zhang Y, Chen W. Invariants of DNA sequences based on 2DD-curves. Journal of Theoretical Biology. 2006;**242**(2):382-388. DOI: 10.1016/j.jtbi.2006.03.012

[16] Bielińska-Wąż D, Clark T, Wąż P, Nowak W, Nandy A. 2D-dynamic representation of DNA sequences. Chemical Physics Letters. 2007;**442**(1–3):140-144. DOI: 10.1016/j.cplett.2007.05.050

[17] Qi ZH, Qi XQ. Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chemical Physics Letters. 2007;**440**(1–3):139-144. DOI: 10.1016/j.cplett.2007.03.107

[18] Qi ZH, Li L, Qi XQ. Using Huffman coding method to visualize and analyze DNA sequences. Journal of Computational Chemistry. 2011;**32**(15):3233-3240. DOI: 10.1002/jcc.21906

[19] Zhang ZJ. DV-curve: A novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics. 2009;**25**(9):1112-1117. DOI: 10.1093/bioinformatics/btp130

[20] Jafarzadeh N, Iranmanesh A. A novel graphical and numerical representation for analyzing DNA sequences based on codons. MATCH Communications in Mathematical and in Computer Chemistry. 2012;**68**:611-620

[21] Wąż P, Bielińska-Wąż D, Nandy A. Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences. Journal of Mathematical Chemistry. 2013;**52**(1):132-140. DOI: 10.1007/s10910-013-0249-1

[22] Zou S, Wang L, Wang J. A 2D graphical representation of the sequences of DNA based on triplets and its application. EURASIP Journal on Bioinformatics and Systems Biology. 2014;**2014**(1):1. DOI: 10.1186/1687-4153-2014-1

[23] Hamori E. Novel DNA sequence representations. Nature. 1985;**314**:585. DOI: 10.1038/314585a0

[24] Randić M, Vračko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. Journal of Chemical Information and Computer Sciences. 2000;**40**(5):1235-1244. DOI: 10.1021/ci000034q

[25] Randić M, Balaban AT. On a four-dimensional representation of DNA primary sequences. Journal of Chemical Information and Computer Sciences. 2003;**43**(2):532-539. DOI: 10.1021/ci020051a

[26] Liao B, Wang TM. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. Chemical Physics Letters. 2004;**388**(1–3):195-200. DOI: 10.1016/j.cplett.2004.02.089

[27] Liao B, Ding K. A 3D graphical representation of DNA sequences and its application. Theoretical Computer Science. 2006;**358**(1):56-64. DOI: 10.1016/j.tcs.2005.12.012

[28] Liao B, Li R, Zhu W, Xiang X. On the similarity of DNA primary sequences based on 5-D representation. Journal of Mathematical Chemistry. 2007;**42**(1):47-57. DOI: 10.1007/s10910-006-9091-z

[29] Yao YH, Nan XY, Wang TM. Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation. Chemical Physics Letters. 2005;**411**(1–3):248-255. DOI: 10.1016/j.cplett.2005.06.040

[30] Chi R, Ding K. Novel 4D numerical representation of DNA sequences. Chemical Physics Letters. 2005;**407**(1–3):63-67. DOI: 10.1016/j.cplett.2005.03.056

[31] Qi ZH, Fan TR. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2007;**442**(4–6):434-440. DOI: 10.1016/j.cplett.2007.06.029

[32] Yu JF, Sun X, Wang JH. TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. Journal of Theoretical Biology. 2009;**261**(3):459-468. DOI: 10.1016/j.jtbi.2009.08.005

[33] Xie G, Mo Z. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. Journal of Theoretical Biology. 2011;**269**(1):123-130. DOI: 10.1016/j.jtbi.2010.10.018

[34] Jafarzadeh N, Iranmanesh A. C-curve: A novel 3D graphical representation of DNA sequence based on codons. Mathematical Biosciences. 2013;**241**(2):217-224. DOI: 10.1016/j.mbs.2012.11.009

[35] Wąż P, Bielińska-Wąż D. 3D–dynamic representation of DNA sequences. Journal of molecular modeling. 2014;**20**(3):2141. DOI: 10.1007/s00894-014-2141-8.

[36] Wąż P, Bielińska-Wąż D. Non-standard similarity/dissimilarity analysis of DNA sequences. Genomics. 2014;**104**:464-471. DOI: 10.1016/j.ygeno.2014.08.010

[37] Liao B, Tan M, Ding K. Application of 2-D graphical representation of DNA sequence. Chemical Physics Letters. 2005;**414**(4–6):296-300. DOI: 10.1016/j.cplett.2005.08.079

[38] Yu C, Liang Q, Yin C, He RL, Yau SST. A novel construction of genome space with biological geometry. DNA Research. 2010;**17**(3):155-168. DOI: 10.1093/dnares/dsq008

[39] Huang G, Zhou H, Li Y, Xu L. Alignment-free comparison of genome sequences by a new numerical characterization. Journal of Theoretical Biology. 2011;**281**(1):107-112. DOI: 10.1016/j.jtbi.2011.04.003

[40] Mizuta S, Yamaguchi K. A novel 2-dimensional graphical representation of DNA sequences using weighted vector assignments. In: The Proceedings of the 6th International Conference on Bioinformatics Computational Biology (BICoB2014); Las Vegas; 2014. pp. 33-38

[41] Kobori Y, Mizuta S. Similarity estimation between DNA sequences based on local pattern histograms of binary images. Genomics, Proteomics & Bioinformatics. 2016;**14**(2):103-112. DOI: 10.1016/j.gpb.2015.09.007

[42] Yamaguchi K, Mizuta S. A new graphical representation of DNA sequences using symmetrical vector assignment. Review of Bioinformatics and Biometrics. 2014;**3**:14-21

[43] Zhang R, Zhang CT. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. Journal of Biomolecular Structure & Dynamics. 1994;**11**(4):767-782. DOI: 10.1080/07391102.1994.10508031

[44] Randić M, Butina D, Zupan J. Novel 2-D graphical representation of proteins. Chemical Physics Letters. 2006;**419**(4–6):528-532. DOI: 10.1016/j.cplett.2005.11.091

[45] Wen J, Zhang Y. A 2D graphical representation of protein sequence and its numerical characterization. Chemical Physics Letters. 2009;**476**(4–6):281-286. DOI: 10.1016/j.cplett.2009.06.017

[46] Wu ZC, Xiao X, Chou KC. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. Journal of Theoretical Biology. 2010;**267**(1):29-34. DOI: 10.1016/j.jtbi.2010.08.007

[47] He PA, Zhang YP, Yao YH, Tang YF, Nan XY. The graphical representation of protein sequences based on the physicochemical properties and its applications. Journal of Computational Chemistry. 2010;**31**(11):2136-2142. DOI: 10.1002/jcc.21501

[48] Yu C, Cheng SY, He RL, Yau SST. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. Gene. 2011;**486**(1–2):110-118. DOI: 10.1016/j.gene.2011.07.002

[49] Yu HJ, Huang DS. Novel 20-D descriptors of protein sequences and it's applications in similarity analysis. Chemical Physics Letters. 2012;**531**:261-266. DOI: 10.1016/j.cplett.2012.02.030

[50] Abo el Maaty MI, Abo-Elkhier MM, Abd Elwahaab MA. 3D graphical representation of protein sequences and their statistical characterization. Physica A: Statistical Mechanics and its Applications. 2010;**389**(21):4668-4676. DOI: 10.1016/j.physa.2010.06.031

[51] He P, Wei J, Yao Y, Tie Z. A novel graphical representation of proteins and its application. Physica A: Statistical Mechanics and its Applications. 2012;**391**(1–2):93-99. DOI: 10.1016/j.physa.2011.08.015

[52] He P, Li D, Zhang Y, Wang X, Yao Y. A 3D graphical representation of protein sequences based on the gray code. Journal of Theoretical Biology. 2012;**304**(0):81-87. DOI: 10.1016/j.jtbi.2012.03.023

[53] Czerniecka A, Bielińska-Wąż D, Wąż P, Clark T. 20D-dynamic representation of protein sequences. Genomics. 2016;**107**(1):16-23. DOI: 10.1016/j.ygeno.2015.12.003

[54] Randić M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D. Graphical representation of proteins. Chemical Reviews. 2011;**111**(2):790-862. DOI: 10.1021/cr800198j

[55] Raychaudhury C, Nandy A. Indexing scheme and similarity measures for macromolecular sequences. Journal of Chemical Information and Computer Sciences. 1999;**39**(2):243-247

[56] Randić M, Kleiner AF, De Alba LM. Distance/distance matrixes. Journal of Chemical Information and Modeling. 1994;**34**(2):277-286. DOI: 10.1021/ci00018a008

[57] Swain MJ, Ballard DH. Color indexing. International Journal of Computer Vision. 1991;**7**(1):11-32. DOI: 10.1007/BF00130487

[58] Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of Calcutta Mathematical Society. 1943;**35**(1):99-109

[59] Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991;**37**(1):145-151. DOI: 10.1109/18.61115

[60] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics. 1951;**22**(1):79-86. DOI: 10.1214/aoms/1177729694

[61] Kendall MG. A new measure of rank correlation. Biometrika. 1938;**30**(1–2):81-93. DOI: 10.1093/biomet/30.1-2.81

IntechOpen