Genomics Proteomics Bioinformatics 14 (2016) 103–112

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

Similarity Estimation Between DNA Sequences Based on Local Pattern Histograms of Binary Images

Yusei Kobori [a], **Satoshi Mizuta** [*,b]

Graduate School of Science and Technology, Hirosaki University, Hirosaki, Aomori 036-8561, Japan

Received 30 March 2015; revised 19 September 2015; accepted 23 September 2015
Available online 27 April 2016

Handled by Le Zhang

KEYWORDS

Genome sequence;
Mitochondria;
Bitmap image;
Occurrence frequency;
Distance measure

Abstract Graphical representation of DNA sequences is one of the most popular techniques for alignment-free sequence comparison. Here, we propose a new method for the feature extraction of DNA sequences represented by binary images, by estimating the similarity between DNA sequences using the frequency histograms of local bitmap patterns of images. Our method shows linear time complexity for the length of DNA sequences, which is practical even when long sequences, such as whole **genome sequences**, are compared. We tested five **distance measures** for the estimation of sequence similarities, and found that the histogram intersection and Manhattan distance are the most appropriate ones for phylogenetic analyses.

Introduction

Sequence alignment [1,2] is generally used to estimate similarities between relatively short sequences less than about several thousand characters, such as nucleotide sequences or amino acid sequences. However, the time complexity of the alignment is the square of the sequence length, thus the long sequence length may result in enormous amount of computation time [3]. Therefore, to reduce the time required for comparing long sequences such as whole genome sequences, developing so-called alignment-free methods becomes a necessity.

Graphical representation of biological sequences represents one of the most popular methods for the alignment-free sequence comparison [4]. Various methods based on graphical representation have been introduced, and almost all methods share the common basic procedure. Every base type in a DNA sequence is replaced by an individual vector in a two-dimensional (2D) [5,6], three-dimensional (3D) [7,8], or even higher-dimensional [9] expression space. These vectors are then connected successively, drawing a trajectory in the expression space and finally, the distances between the trajectories, or graphs, are calculated according to a predefined distance measure. While there exist many methods in the field of graphical representation of biological sequences as mentioned above,

* Corresponding author.
E-mail: slmizu@hirosaki-u.ac.jp (Mizuta S).
[a] ORCID: 0000-0003-0953-8723.
[b] ORCID: 0000-0002-9890-3448.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.09.007

further improvement in terms of the performance and the calculation time is still required.

In this study, we propose a new method for sequence comparison based on the graphical representation. We expressed a DNA sequence as a binary image—each pixel of a binary image was plotted in either black or white—in a two-dimensional space, and counted the occurrence frequencies of 3×3 bitmap patterns in the binary image. The distance between the binary images was measured based on the frequency histograms of the bitmap patterns. Five frequently-used distance measures were evaluated for their performance in determining the distance between histograms based on the phylogeny of 31 mitochondrial genome sequences. These include histogram intersection [10], Manhattan distance, Bhattacharyya distance [11], Jensen–Shannon divergence [12], and Kendall's rank correlation coefficient [13].

Methods

Generating a binary image from a DNA sequence

We describe here the step-by-step procedure used to generate a binary image from a DNA sequence.

Graphical representation of a DNA sequence

Firstly, we assigned 2D numerical vectors on the *xy*-plane, which are perpendicular or in opposite directions to each other, to A, T, G, and C. The number of independent variations of the assignments was 3!/2 = 3, including the assignments that can be transformed into each other by 90-degree rotations or the inversions with respect to vertical or horizontal axes (**Figure 1**). We chose the assignment presented in Figure 1A, where nucleotides A and T are placed in the upper quadrants, and G and C in the lower ones, so that the GC content of a DNA sequence can be grasped easily from the resultant graphical representation. Note that, the assignment presented in Figure 1B is also acceptable, but better results are obtained with the former assignment as shown in Figure 1A. Therefore, the assignment given in Figure 1A is adopted throughout this article. Then, a 2D graph can be drawn by consecutively connecting the vectors assigned to the nucleotides of a DNA sequence. A graphical representation of a sequence, "ACATATG", is represented in **Figure 2A**.

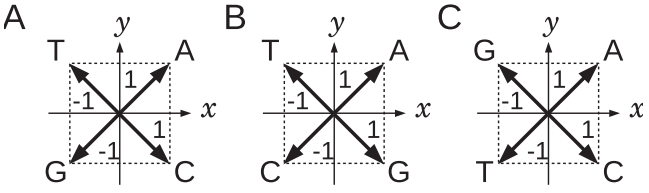

Figure 1 Three independent assignments of vectors on the *xy*-plane to individual nucleotides

There are three independent assignments under the symmetries of 90-degree rotations and the inversion with respect to the vertical or horizontal axis. Four nucleotides A, T, G, and C are arranged counterclockwise on the *xy*-plane "ATGC" (**A**), "ATCG" (**B**), and "AGTC" (**C**). Assignment **A** is used throughout the study.

Multiplying weighting factors

In order to extract potential information conveyed by individual nucleotides, we introduced weighting factors, based on a Markov chain model, into the process of binary images generation [14]. To emphasize rare patterns that appear in genome sequences, we used self-information $I(E)$, the amount of information that is received when a certain event E occurs, as the weighting factor. Let $P(E)$ be the probability that event E occurs, $I(E)$ is defined as $I(E) = -\log_2 P(E)$ in bit units. A trajectory for each genome sequence in a 2D plane is drawn as follows:

$$\boldsymbol{R}_i = \sum_{k=1}^{i} w_k \boldsymbol{V}_k, \tag{1}$$

where $\boldsymbol{R}_i$ is the coordinate of the ith point on the trajectory, $\boldsymbol{V}_k$ is the vector assigned to the kth nucleotide of the genome sequence, and w_k is the corresponding weighting factor $I(E)$. Here, we defined $P(E)$ according to the second order Markov chain. The probability that nucleotide z occurs after a pair of nucleotides xy (x, y, $z \in \{A, T, G, C\}$) is calculated using

$$P(z|xy) = \frac{N_{xyz}}{\sum_{s\in\{A,T,G,C\}} N_{xys}}, \tag{2}$$

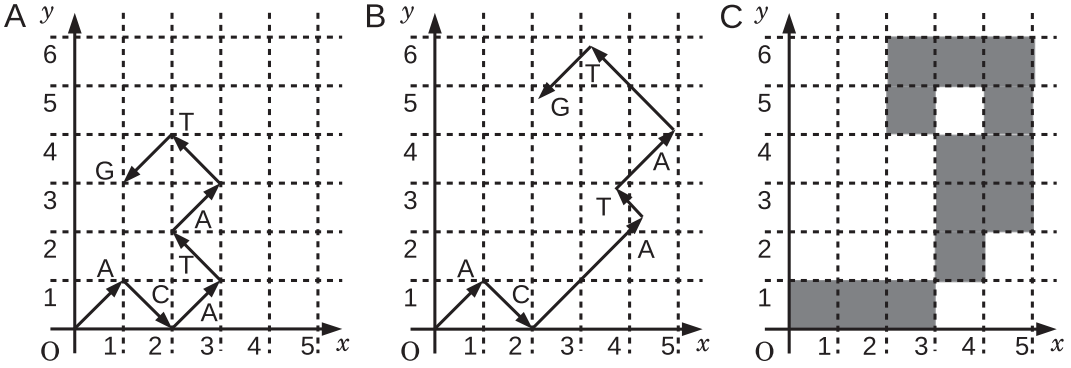

Figure 2 Generating a binary image of sequence "ACATATG"

A. The primary graphical representation. **B.** The graphical representation modified with weighting factors. **C.** The generated binary image. Each grid represents an individual pixel of a binary image.