Whole DNA Sequences of *Cebus capucinus* on Variant Maps

Yuyuan Mao, Jeffrey Zheng and Wenjia Liu

Abstract DNA sequences as a big data stream have been researched for years. However, researches on whole DNA sequences have various limitations to use existing research methods. A new scheme is proposed to map whole DNA sequences as 2D maps in this chapter, the whole DNA sequence of Capuchin monkey (*Cebus capucinus*) in apes was used as an example to demonstrate the mapping results.

Keywords Gene sequence · *Cebus capucinus* · Mapping method
Sequential model · Variant map

1 Introduction

In modern biologics, DNA sequences are being sequenced from wider species from human to simple cells in DNA data banks as big data streams. It is difficult to process various DNA streams for classification and identification on various species from whole sequences. The main task of present genomic research [1, 2] is to obtain

This work was supported by the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002), NSF of China (61362014) and Yunnan Advanced Overseas Scholar Project.

Y. Mao
School of Software, Yunnan University, Kunming, China
e-mail: m805792943@foxmail.com

J. Zheng (✉)
Key Laboratory of Quantum Information of Yunnan, Yunnan University, Kunming, China
e-mail: conjugatelogic@yahoo.com

J. Zheng
Key Laboratory of Software Engineering of Yunnan, Yunnan University, Kunming, China

W. Liu
Yunnan University, Kunming, China
e-mail: 8avalon8@gmail.com

© The Author(s) 2019
J. Zheng (ed.), *Variant Construction from Theoretical Foundation to Applications*,
https://doi.org/10.1007/978-981-13-2282-2_24

more biological information by processing and analyzing of the DNA sequence from multi-angles and multilevels [4–7]. In recent years, the processing and utilization of biological gene data are being carried out in a variety of ways, such as gene feature extraction, gene sequence location [7–9], and so on.

Variant map is an emerging technology to handle four symbols as meta-structure to process random sequences from cryptographic sequences, DNA sequences [3, 10] to ECG signals. Multiple statistical probability distributions are generated from selected sequences to form 2D–3D visual maps in representation. This scheme makes whole data sequences more compact and effectively visualized, and mapping results may be useful to explore nonlinear complex behaviors of whole genomics. A whole DNA sequence of a night monkey has mapped [11] on variant maps.

In this chapter, a special scheme is proposed to show a series of mapping results from a selected gene sequence of a capuchin monkey.

2 Process Model

A. Architecture

The architecture of the process model is shown in Fig. 1a. The process model consists of five parts: input, processing, measurement, projection, and output. There are three modules: Processing, Measurement, and Projection.

Input: A DNA sequence

Output: A 2D map

Modules: Processing, Measurement, and Projection

Process: From a selected DNA sequence, multiple segments are divided by a fixed length m on the whole sequence sequentially in the Processing module. Each segment needs to count four symbols: $\{A, C, G, T\}$ in the segment to transfer all segments into a measuring sequence of four measures in Measurement module. A special combination on $X: \{AT\}$ and $Y: \{AG\}$ is selected to determine four measures in a projection position and the whole measuring sequence projected to be a 2D map in Projection module.

B. Processing Module

From an input DNA sequence, multiple segments can be separated by a fixed length m to generate a sequence of segments.

Input: a DNA sequence

Output: a sequence of segments

C. Measurement Module

In this module, shown in Fig. 1b, each segment counts four numbers of $\{A, G, C, T\}$ in each proportions, respectively. As the result, each count is an integer number between 0 and m to transfer a segment sequence into a measuring sequence of four measures.

Input: a sequence of segments
Output: a sequence of four measures

D. Projection Module

The projection module is shown in Fig. 1c as two units: Position and Projecting. For each four measures, two axis positions are determined by $X(AT)$ and $Y(AG)$, respectively. When all measures are processed, a 2D histogram is established as a statistical distribution as a 2D map.

Input: a sequence of four measures
Output: a 2D map

Fig. 1 Architecture of mapping scheme (a)–(c). **a** Architecture; **b** Measurement module; **c** Projection module

3 Details

A. Relevant Parameters

m : segment length

V : Two bases of combination: $\{AT, AG\}$

$$\text{num}(AT) = \text{num}(A) + \text{num}(T);$$

$$\text{num}(AG) = \text{num}(A) + \text{num}(G);$$

$$P_v = \text{num}(V)$$

P_v : The proportion of a base or combinatorial base

$(X_{P_{AT}}, Y_{P_{AG}})$: a pair of XY mapping positions.

B. Parameter in Module

Since the output quality of generating maps is dependent on the number of projection points, it is necessary for a refined map to include a larger number of coordinate points. The mapping projection forms the superposition to add up a larger number of coordinate points in 2D histogram representing a color map.

C. Measurement module.

- m : subsection length of a DNA sequence
- $\text{num}(AT) = \text{num}(A) + \text{num}(T)$
- V : AT or AG , $\{AT, AG\} \in D$.
- P_v : The proportion of AT or AG on the length of the sequence M .
- $P_v = \text{num}(V)/m$
- P : The proportion of AT
- P_{AG} : The proportion of AG
- $(X_{P_{AT}}^i, Y_{P_{AG}}^j)$: a pair of XY mapping coordinates. i, j are different subsections.

D. Parameter in Module

Calculating the proportion of AT and AG in the subsection according to the basic rules of mathematics. Two proportions can form a coordinate $(X_{P_{AT}}^i, Y_{P_{AG}}^j)$, which map a point on the two-dimensional graph.

The mapping relation between x and y :

$$X : P_{AT}$$

$$Y : P_{AG}$$

It is necessary for a distinct graph that includes a large number of coordinate points. Only a large number of DNA sequences can get a large number of coordinates points and pretty projection results. The graphics projection module completes the superposition of a large number of coordinate points.

4 Results Display

4.1 Maps on Various Segmented Length

Different parameters are shown in Fig. 2a–l for $m = \{20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 200\}$, Fig. 3a–f for $m = \{54, 56, 58, 60, 62, 64\}$, Fig. 4a–d for $m = \{59, 60, 61, 62\}$ and Fig. 5 for $m = 60$, respectively.

In the map, similar color of pixels indicates the similar number of segments in the cluster.

4.2 Brief Analysis

From Fig. 2, it is interesting to notice that when $m < 50$, maps have more symmetric properties than larger numbers. Changing segmented lengths, significant patterns appear in $m = 54$ – 64 region shown in Fig. 3 and refined lengths are shown in Fig. 4.

From a visual observation, when $m = 60$, the map has shown the better effects.

5 Conclusion

Using the proposed mapping scheme, it is feasible to transfer a whole DNA sequence as a color map with significant visual features. In addition to mapping method and selected functions, a set of sample sequences in various segmented lengths illustrate colorful distributions as variant maps.

Checking symmetric information among different maps, it is possible to identify specific spatial features under different configurations.

Fig. 2 Variant maps of *Cebus capucinus* on various segmented lengths (a)–(l) $m = \{20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 200\}$. **a** $m = 20$; **b** $m = 30$; **c** $m = 40$; **d** $m = 50$; **e** $m = 60$; **f** $m = 70$; **g** $m = 80$; **h** $m = 90$; **i** $m = 100$; **j** $m = 120$; **k** $m = 150$; **l** $m = 200$

Fig. 3 Variant maps of *Cebus capucinus* on various segmented lengths (a)–(f) $m = \{54, 56, 58, 60, 62, 64\}$; **a** $m = 54$; **b** $m = 56$; **c** $m = 58$; **d** $m = 60$; **e** $m = 62$; **f** $m = 64$

Fig. 4 Variant maps of *Cebus capucinus* on various segmented lengths (a)–(d) $m = \{59, 60, 61, 62\}$. **a** $m = 59$; **b** $m = 60$; **c** $m = 61$; **d** $m = 62$

Fig. 5 Variant maps of *Cebus capucinus* on segmented lengths $m = 60$

Since this is an initial step to make a whole DNA sequence in mapping operation, further researches and explorations are required.

References

1. J.A. Berger, S.K. Mitra, M. Carli, A. Neri, Visualization and analysis of DNA sequences using DNA walks. *J. Franklin Inst.* **341**(1/2) (2004)
2. J.N. Pitt, I. Rajapakse, A.R. Ferré-D'Amaré, SEWAL: an open-source platform for next-generation sequence analysis and visualization. *PMC* **38**(22), 7908–7915 (2010)
3. L. Yuqian, Z. Zhijie, The Visual Analysis of Coding and Non-Coding DNA Sequences. *Hans J. Comput. Biol.* **4**, 20–31 (2014)
4. J. Hellman, S. Drucker, N.R. Riche, B. Lee, A deeper understanding of sequence in narrative visualization. *IEEE Trans. Vis. Comput. Graph.* **19**(12) (2013)
5. G.-D. Sun, Y.-C. Wu, R.-H. Liang, S.-X. Liu, A survey of visual analytics techniques and applications: state-of-the-art research and future challenges (2013). <https://doi.org/10.1007/s11390-013-1383-8>
6. J. Batley, D. Edwards, Genome sequence data: management, storage, and visualization. *BioTechniques* **46**(5) (2009)
7. Y. Nakamura, T. Gojobori, T. Ikemura, Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucl. Acids. Res.* **28**, 292 (2000)
8. N. Rusk, Focus on next-generation sequencing data analysis. *Nat. Methods* **6**, S1 (2009)
9. R. Durrett, *Probability Models for DNA Sequence Evolution* (Springer, 2008)

10. J. Zheng, W. Zhang, J. Luo, W. Zhou, R. Shen, Variant map system to simulate complex properties of DNA interactions using binary sequences. *Adv. Pure Math.* **3**(7A), 5–24 (2013)
11. Y. Mao, J. Zheng, W. Liu, Mapping Whole DNA Sequence on Variant Maps, *ASONAM '17 Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 1037–1040

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.