DNA sequence representation without degeneracy

Stephen S. -T. Yau*, Jiasong Wang1, Amir Niknejad, Chaoxiao Lu, Ning Jin1 and Yee-Kin Ho2

Department of Mathematics, Statistics and Computer Science and 2Department of Biochemistry and Molecular Biology, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045, USA and 1Department of Mathematics, Nanjing University, Nanjing 210008, China

Received February 28, 2003; Revised April 9, 2003; Accepted April 22, 2003

ABSTRACT

Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures. A new two-dimensional graphical representation method using a two-quadrant Cartesian coordinates system has been derived for mathematical denotation of DNA sequence. The two-dimensional graphic representation resolves sequences’ degeneracy and is mathematically proven to eliminate circuit formation. Given x-projection and y-projection of any point on the graphical representation, the number of A, G, C and T from the beginning of the sequence to that point could be found. Compared with previous methods, this graphical representation is more inline with the conventional recognition of linear sequences by molecular biologists, and also provides a metaphor in two dimensions for local and global DNA sequence comparison.

Mathematical analysis of large volume genomic DNA sequence data is one of the challenges for bio-scientists. Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures. About twenty years ago, Hamori first used a three-dimensional H curve to represent a DNA sequence (1,2). Sophisticated computer graphic tools are needed to generate the H curve (3). Gates proposed a two-dimensional graphical representation that is simpler than the H curve (4,5). However, Gates’s graphical representation has high degeneracy. For example, the sequences AGTC, AGTCA, AGTCAG, etc. have many circuits.

To prove there is no circuit or degeneracy in our two-dimensional graphical representation, we assume that (1) the number of nucleotide forming a circuit is n; (2) the number of A, G, C and T in a circuit is a, g, c and t, respectively. So, a + g + c + t = n. Because aA, gG, cC and tT form a circuit, the following equation holds:

\[ a \left( \frac{1}{2}, -\frac{\sqrt{3}}{2} \right) + g \left( \frac{\sqrt{3}}{2}, -\frac{1}{2} \right) + c \left( \frac{\sqrt{3}}{2}, \frac{1}{2} \right) + t \left( \frac{1}{2}, \frac{\sqrt{3}}{2} \right) = 0 \]

i.e.,

\[ a + \sqrt{3}g + \sqrt{3}c + t = 0 \]  
\[ -\sqrt{3}a + g + \sqrt{3}t = 0 \]

Clearly 1 and 2 hold if, and only if, a = g = c = t = 0. Therefore, n = 0, which means no circuit exists in this graphical representation.

Furthermore, given x-projection and y-projection of any point \( p = (x, y) \) on the sequence, we have

\[ a \left( \frac{1}{2}, -\frac{\sqrt{3}}{2} \right) + g \left( \frac{\sqrt{3}}{2}, -\frac{1}{2} \right) + c \left( \frac{\sqrt{3}}{2}, \frac{1}{2} \right) + t \left( \frac{1}{2}, \frac{\sqrt{3}}{2} \right) = (x, y) \]

i.e.

*aTo whom correspondence should be addressed. Tel/Fax: +1 312 996 3065; Email: yau@uic.edu
where \( x \) is the \( x \)-projection and \( y \) is the \( y \)-projection of the point. 2\( x \) and 2\( y \) are irrational numbers of form \( m + n\sqrt{3} \), where \( m \) and \( n \) are integers. After uniquely determining \( m_x, n_x, m_y, n_y \) from 2\( x \) and 2\( y \), the number \( a_p, g_p, c_p \) and \( t_p \) of A, G, C and T from the beginning of the sequence to the point \( p \) can be found by solving linear system:

\[
\begin{align*}
    a + \sqrt{3}g + \sqrt{3}c + t &= 2x \\
    -\sqrt{3}a - g + c + \sqrt{3}t &= 2y
\end{align*}
\]

By successive \( x \)-projection and \( y \)-projection of points on the sequence, we can recover the original DNA sequence uniquely from the DNA graph.

The current scheme provides a direct plotting method to denote DNA sequences without degeneracy. Compared with previous methods, this graphical representation is more in-line with the conventional recognition of linear sequences from 5’ to 3’ end by molecular biologists (Fig. 2), and can easily be constructed without extensive computer graphic tools. The features of peaks and valleys generated from the DNA graph are distinct for specific DNA sequence. These long-range distinct patterns can be recognized visually. From the DNA graph, the A, T, G, C usage as well as the original DNA sequence can be recaptured mathematically without loss of
textual information. Hamori’s H curve can be established with extensive annotations designating the sites of known biological functions, genes, exons, introns and so on. Our two-dimensional curves can also be annotated in a similar fashion. High complexity and degeneracy are major problems in previous DNA graphical representations (6–8), which limit the application of DNA graphs. The current two-dimensional graphical representation of DNA sequences introduced in this paper overcomes these problems and therefore will provide a different approach for both computational scientists and molecular biologists to analyse DNA sequences efficiently.

ACKNOWLEDGEMENTS

We thank the referees for their suggestions of improving the presentation of this paper.

REFERENCES