Sequence comparison and essential gene identification with new inter-nucleotide distance sequences

Yushuang Li, Yanfen Lv, Xiaonan Li, Wenli Xiao, Chun Li

www.elsevier.com/locate/yjtbi

PII: S0022-5193(17)30031-0
DOI: http://dx.doi.org/10.1016/j.jtbi.2017.01.031
Reference: YJTBI8941

To appear in: *Journal of Theoretical Biology*

Received date: 19 October 2016
Revised date: 17 January 2017
Accepted date: 19 January 2017

Cite this article as: Yushuang Li, Yanfen Lv, Xiaonan Li, Wenli Xiao and Chun Li, Sequence comparison and essential gene identification with new inter nucleotide distance sequences, *Journal of Theoretical Biology* http://dx.doi.org/10.1016/j.jtbi.2017.01.031

could deduce the following recurrence formula:

$$\begin{cases} i_{L_1}^{\mathrm{A}} = n - a_{L_1}, \\ i_j^{\mathrm{A}} = i_{j+1}^{\mathrm{A}} - d^{\mathrm{A}}(j),\ j = L_1 - 1, L_1 - 2, \ldots, 2, 1. \end{cases} \tag{3}$$

By the similar procedure we are able to obtain the occurrence positions of the bases G and C in S respectively, and the base T must occur in the remaining positions. Complete the proof.

For example: To the above DNA fragment $S = \mathrm{ACAGCTCTTGATACG}$, $d^{\mathrm{A}} = (2,8,2,2)$, $d^{\mathrm{G}} = (6,5,0)$, $d^{\mathrm{C}} = (3,2,7,1)$, $d^{\mathrm{T}} = (2,1,3,3)$. According to the proposition we could take the following procedure to reproduce S.

$d^{\mathrm{A}}(4) = 2$ - - - - - - - - - - - - A - -

$d^{\mathrm{A}}(3) = 2$ - - - - - - - - - - A - A - -

$d^{\mathrm{A}}(2) = 8$ - - A - - - - - - - A - A - -

$d^{\mathrm{A}}(1) = 2$ A - A - - - - - - - A - A - -

$d^{\mathrm{G}}(3) = 0$ A - A - - - - - - - A - A - G

$d^{\mathrm{G}}(2) = 5$ A - A - - - - - - G A - A - G

$d^{\mathrm{G}}(1) = 6$ A - A G - - - - - G A - A - G

$d^{\mathrm{C}}(4) = 1$ A - A G - - - - - G A - A C G

$d^{\mathrm{C}}(3) = 7$ A - A G - - C - - G A - A C G

$d^{\mathrm{C}}(2) = 2$ A - A G C - C - - G A - A C G

$d^{\mathrm{C}}(1) = 3$ A C A G C - C - - G A - A C G

S A C A G C T C T T G A T A C G

2.2 20 dimensional feature vector

For a DNA sequence S with length n, suppose that the base x ($x \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$) occurs m times with occurrence positions $i_1, i_2, \ldots, i_m$. From equation (2) we have $d^x(m) = n - i_m$, the distance between the i_mth base x and the nth base y in S, it is not in fact the inter-nucleotide distance between two same bases, so we remove it from d^x and call the remaining integer sequence *precise inter-nucleotide distance sequence* $\boldsymbol{d}^x$. Arrange all elements in $\boldsymbol{d}^x$ in ascending order and then obtain an *ordered precise inter-nucleotide distance sequence* $\vec{\boldsymbol{d}}^x$. Based on $\boldsymbol{d}^x$ and $\vec{\boldsymbol{d}}^x$ we extract five basic statistical quantities.