

Author's Accepted Manuscript

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

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



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>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and a review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain

could deduce the following recurrence formula:

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

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 $S = ACAGCTCTTGATACG$, $d^A = (2, 8, 2, 2)$, $d^G = (6, 5, 0)$, $d^C = (3, 2, 7, 1)$, $d^T = (2, 1, 3, 3)$. According to the proposition we could take the following procedure to reproduce S .

$d^A(4) = 2$	-----	A	--
$d^A(3) = 2$	-----	A	-A--
$d^A(2) = 8$	--	A	-----A-A--
$d^A(1) = 2$	A	A	-----A-A--
$d^G(3) = 0$	A	A	-----A-A
$d^G(2) = 5$	A	A	-----GA-A-G
$d^G(1) = 6$	A	A	GA-----GA-A-G
$d^C(4) = 1$	A	AG	-----GA-ACG
$d^C(3) = 7$	A	AG	--C--GA-ACG
$d^C(2) = 2$	A	AG	C-C--GA-ACG
$d^C(1) = 3$	A	CAG	C-C--GA-ACG
S	ACAGCTCTTGATACG		

2.2 20 dimensional feature vector

For a DNA sequence S with length n , suppose that the base x ($x \in \{A, G, C, T\}$) occurs m times with occurrence positions i_1, i_2, \dots, i_m . From equation (2) we have $d^x(m) = n - i_m$, the distance between the i_m th base x and the n th 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* d^x . Arrange all elements in d^x in ascending order and then obtain an *ordered precise inter-nucleotide distance sequence* \vec{d}^x . Based on d^x and \vec{d}^x we extract five basic statistical quantities.

Download English Version:

<https://daneshyari.com/en/article/5760303>

Download Persian Version:

<https://daneshyari.com/article/5760303>

[Daneshyari.com](https://daneshyari.com)