

Bioinformatics

by Biomedical Informatics Publishing Group

$x[n] = [A A T G C A T C A]$ then,
 $u_A[n] = [1 1 0 0 0 1 0 0 1]$,
 $u_G[n] = [0 0 0 1 0 0 0 0 0]$,
 $u_C[n] = [0 0 0 0 1 0 0 1 1]$ and
 $u_T[n] = [0 0 1 0 0 0 1 0 0]$.

Obviously, the sum of all binary indicators at any position n is 1 for all n .

$$\text{i.e. } u_A[n] + u_G[n] + u_C[n] + u_T[n] = 1 \text{ for } n=0, 1, 2, \dots, N-1. \quad (1)$$

Let $U_A[k]$, $U_G[k]$, $U_C[k]$ and $U_T[k]$ be the Discrete Fourier Transforms (DFT) of the binary sequences $u_A[n]$, $u_G[n]$, $u_C[n]$ & $u_T[n]$ respectively which are given by,

$$U_X[k] = \sum_{n=0}^{N-1} u_X[n] e^{(-j2\pi kn/N)} \quad , \quad X=A, G, C, \text{ or } T \text{ and } k = 0, 1, 2, \dots, (N-1) \quad (2)$$

$$S[k] = \sum |U_X[k]|^2 \text{ for } X=A, G, C \text{ or } T. \quad \& \quad k = 0, 1, 2, \dots, (N-1) \quad (3)$$

$S[k]$ may be used as a preliminary indicator of a coding region as a plot of $S[k]$ against k reveals a peak at $k=N/3$ for coding region and shows no such peak for noncoding region. [2] It has been proved that the pronounced peak actually springs from the nonuniform distribution of the nucleotides in the three coding positions of codons in a coding area. [3] And $S[k]$ as a coding measure is model independent as it is not specific to any particular genome.