Information dimension of gene CGR representation

Information dimension is an index for characterizing fractal patterns or sets by quantifying their complexity as a ratio of the change in detail to the change in scale [21]. We divide the CGR pattern into ε^2 square grids of equal size, and then the side length of each grid is $1/\varepsilon$. The numbers of non-empty grids are labeled by $Z(\varepsilon)$. Dividing the number of points falling in the *i*-th grid by the total point number in the CGR square yields a probability p_i for the *i*-th grid. Information function and information dimension for the CGR are respectively defined as

$$I(\varepsilon) = -\sum_{i=1}^{Z(\varepsilon)} p_i \log p_i \tag{5}$$

$$D_I = \lim_{1/\varepsilon \to 0} \frac{I(\varepsilon)}{\log(1/\varepsilon)} \tag{6}$$

The information function $I(\varepsilon)$ during a range of $\log(1/\varepsilon)$has a scaling region. The information dimension D_I can be obtained from the slope in the scaling region.

Hurst exponent of gene time series

The Hurst exponent is the measure of the smoothness of fractal time series based on the asymptotic behavior of the rescaled range of the process[22]. In this study, rescaled range (R/S) analysis, a statistical method is developed to estimate the Hurst exponent of the times series of gene sequence. It involved the following basic steps. For a given gene sequence $x(s)$, F is a transformed times series over a total duration *N*. for a deterministic integer τ ,the cumulative total at each point in times is defined as

$$\Gamma_{\tau,k} = \sum_{i=1}^{k} (F_i - \mu_\tau) \quad 0 < k \le \tau \tag{7}$$

Where, F_i is the value of the time series at time *i*, μ_τ is the mean over the whole data set given by

$$\mu_\tau = \left(1/\tau\right)\sum_{i=1}^{\tau} F_i \tag{8}$$

The range R of given by

$$R_\tau = Max\left(\Gamma_{\tau,k}\right) - Min\left(\Gamma_{\tau,k}\right) \tag{9}$$

The standard deviation of the values over the whole data set is given

$$S_\tau = \sqrt{\left(1/\tau\right)\sum_{i=1}^{\tau}\left(F_i - \mu_\tau\right)^2} \tag{10}$$

The rescale range is given by R/S. The Hurst exponent is estimated by plotting the values of log(R/S) versus $\log\tau$. The slope of the best fitting line gives the estimate of the Hurst exponent.

Topological entropy of gene sequences

Topological entropy is a measure of complex regulation of a gene sequence. For a given gene sequence $x(s)$, N corresponds to the length of $x(s)$ and n is defined as a unique integer by the following equation,

$$4^n + n - 1 \le N < 4^{n+1} + (n+1) - 1 \tag{1}$$

Based on deferent n, the complexity function C_x is defined as:

$$C_x(n) = \left|\left\{m : |m| = n \text{ and } m \text{ appears as a subword of } x\right\}\right| \tag{2}$$

Where, C_x represents the number of different n-length sub-words (overlaps allowed) that appear in $x(s)$.

Then for x^{4^n+n-1} the first $4^n + n - 1$ letters of x, the definition of topological entropy of the finite sequence is

$$H_{top}(x) = \frac{\log_4\left(C_{x^{4^n+n-1}}(n)\right)}{n} \tag{3}$$

Where, $H_{top}(x)$ is the topological entropy of $x(s)$.