Genomic signature distance

Campbell et al. (1999) defined the **Genomic Signature Distance** *d* between array $(X_i)_{(1\le i\le n)}$ and array $(Y_i)_{(1\le i\le n)}$ for dinucleotides with the following formula:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left| f'_{x\,i} - f'_{y\,i} \right|$$

where f'_{xi} and f'_{yi} are the standardized frequencies of the *i*-th oligonucleotide in arrays X and Y and n is equal to 4^k (where k is the number of bases of the oligonucleotides). Campbell et al. (1999) applied the formula specifically to compare dinucleotide relative abundance profiles and named it average absolute dinucleotide relative abundance difference (δ*).

Wang et al. (2005) applied the same formula for longer oligonucleotides, calling it the Hamming distance, but as the Hamming distance is usually defined for strings or vectors and only accounts for the number of positions at which they differ, the present study will call this the Genomic Signature Distance.

The oligonucleotide occurrences obtained from DNA sequences must be standardized using the following formula:

$$f'_{xi} = \frac{\mathrm{n}\ f_{x\,i}}{\sum_{i=1}^{n} f_{x\,i}} \qquad f'_{yi} = \frac{\mathrm{n}\ f_{y\,i}}{\sum_{i=1}^{n} f_{y\,i}}$$

where f_{xi} and f_{yi} are the number of occurrences of the *i*-th and *j*-th oligonucleotides within the arrays X_i and Y_i. For standardized data, the sum of all values equals the number of elements.

We may merge and simplify the above formulas:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left| f'_{x\,i} - f'_{y\,i} \right| = \frac{1}{n}\sum_{i=1}^{n}\left| \frac{\mathrm{n}.fx_i}{\sum_{i=1}^{n} fx_i} - \frac{\mathrm{n}.fx_i}{\sum_{i=1}^{n} fy_i} \right|$$

$$d = \sum_{i=1}^{n}\left| \frac{fx_i}{\sum_{i=1}^{n} fx_i} - \frac{fy_i}{\sum_{i=1}^{n} fy_i} \right|$$

so that the new formula allows for the computation of the Genomic Signature Distance without needing to standardize the oligonucleotide frequencies.