Supplementary Material 1
Frequency Chaos Game Representation

Fatemeh Alipour[1] and Kathleen A. Hill[2], Lila Kari[1]

[1] School of Computer Science, University of Waterloo, Waterloo, ON, Canada
[2] Department of Biology, University of Western Ontario, London, ON, Canada
falipour@uwaterloo.ca

1 FCGR of resolution k

Definition S1.1. *A Frequency Chaos Game Representation ($FCGR_k$) of a sequence $s \in \Delta^n$ with resolution k with $n \geq k$, is a matrix in $R^{2^k \times 2^k}$ derived from X_s, the CGR image of s. Its (i, j)th entry f_{ij} satisfies:*

$$f_{ij} = \frac{Number\ of\ points\ of\ X_s\ in\ cell\ (i, j)}{n} \tag{1}$$

where $cell(i, j)$ is the (i, j)th subsquare, starting from the bottom left, of the square $\mathcal{G}$ if we subdivide $\mathcal{G}$ into $2^k \times 2^k$ equal size subsquares.

It is worth remarking that each of the $2^k \times 2^k$ $cell(i, j)$ corresponds to one of the 4^k k-mer, that is, the frequency f_{ij} that pixels of CGR image X_s of S falling into $cell(i, j)$ is the frequency that the corresponding k-mer occurs in the sequence s. Note that $cell(i, j)$ is uniquely determined by its upper left corner $(x_i, y_j) = (\frac{2(i-1)}{2^k} - 1, \frac{2j}{2^k} - 1)$, or equivalently, $i = 1 + 2^{k-1}(x_i + 1)$, $j = 2^{k-1}(y_i + 1)$. And the upper left (x_i, y_j) is determined by the k-mer s_k recursively as follows: $C(s_k) = (x_i, y_j),\ C(s_k(1 : k - 1)) = (x'_i, y'_j),\ C(empty) = (-1, 1)$,

$$C(s_k) = \begin{cases} ((x'_i - 1)/2, (y'_j - 1)/2), & if\ s(k) = A, \\ ((x'_i + 1)/2, (y'_j - 1)/2), & if\ s(k) = T, \\ ((x'_i + 1)/2, (y'_j + 1)/2), & if\ s(k) = G, \\ ((x'_i - 1)/2, (y'_j + 1)/2), & if\ s(k) = C. \end{cases}$$

The FCGR matrix provides a compact and informative representation of the sequence s.