Overview

Peter A Noble PhD

Below was extracted from BMC Genomics 19:675 (2018)

Extracting 2- to 9-mers from transcript sequences

An alignment-free sequence comparison method called 'Chaos Genome Representation' (CGR) [1, 2, 3] was used to extract mers from the transcript sequences because it was more practical (computationally efficient) than string-based search algorithms (see Proof) [4]. CGR is an iterative mapping technique that processes nucleotides in a sequence to find the *x-, y-* coordinates for their position in a continuous space. The *x-* and *y-* coordinates can then be used to recover sequence, which in this study were oligomers. Once the coordinates of a sequence are known, the presence/absence/ frequency of a mer of any size in a transcript sequence can easily be determined, as demonstrated below.

Reading a sequence into CGR space.

The processing of a transcript sequence involves converting each nucleotide into *x-* and *y-* coordinates and assembling the coordinates into a CGR database. For example, the sequence 'AAACC' is represented by the *x-* and *y-* coordinates of +0.53125 and -0.53125, respectively. The coordinates are determined by reading the sequence into CGR space. The space is confined by the four possible nucleotides as vertices of a binary square with *x, y* position (-1, +1) being the vertex A, (+1, +1) being the vertex T, (-1, +1) being the vertex G and (-1, -1) being the vertex C. The position of a nucleotide in the fragment is calculated by moving a pointer to half the distance between the previous position and the current binary representation.

An example. Starting at point *x, y* (0, 0), the first nucleotide 'A' is plotted at half way to the vertex of A (-1, +1), which is coordinate (-0.5, +0.5). The next nucleotide is also 'A', therefore half way from the coordinate (-0.5, +0.5) to vertex of A (-1, +1) is (-0.75, +0.75). The next nucleotide is also 'A' so half way from the coordinate (-0.75, +0.75) to the vertex of A (-1, +1) is the coordinate (-0.875, +0.875). The next nucleotide 'C', so half-way from the coordinate (-0.875, +0.875) to the vertex of C (-1, -1) is the coordinate (+0.0625, -0.0625) and so on up to the last nucleotide of the sequence with the last coordinates of *x*=+0.53125 and *y*=-0.53125. A depiction of reading a sequence into CGR space is shown in Figure 1a of the Almeida et al. [3] study.

Once all the sequences have been read into CGR space and their coordinates stored in a database, it is possible to determine the presence/absence/frequency of mers by their coordinates and mer length (i.e., 1/resolution), which is outlined in the Mer analysis section below.

Mer analysis

Mer analysis determines the presence/absence/frequency of a mer of length *z* (where *z* is 2 to 9) in a gene transcript.

Finding a specific mer in a transcript. Let us assume that a database of the *x-, y-* coordinates of the target sequence has been assembled and we want to determine the presence/absence of the mer 'AAACAA' in a target sequence. There are three steps.

First, we process the mer AAACAA into CGR space to find it *x-, y-* coordinates, which are -0.734375 and 0.734375, respectively.

Second, we determine the resolution of the search, which depends on mer length (i.e., resolution = $2^{(\text{mer length})}$). A 6-mer requires a resolution of 64. The inverse of the resolution (1/resolution) is the CGR space around the coordinates that contain the specific mer. The CGR space around the coordinates is expressed by the following equation:

$$\boxed{x^{'} = x \pm \frac{1}{r}, y^{'} = y \pm \frac{1}{r}\text{, where } r \text{ is } 2^{\text{mer\_length}}}$$

For the 6-mer AAACAA

$$\boxed{x^{'} = -0.734375 \pm \frac{1}{64}, y^{'} = 0.734375 \pm \frac{1}{64}}$$

Third, the coordinates and CGR space of the mer is then used to search the CGR space of the target transcript sequence in the database. Any transcript that have coordinates within the box of x' and y' of the mer represents the sequence 'AAACAA'. Furthermore, one can tally the number of hits within the box, which represents the frequency of the mer in a target sequence. We verified the presence of the mers in the identified target sequences by textual comparisons.

REFERENCES

1. **Jeffrey HJ.** Chaos game representation of gene structure. Nucleic Acids Res 1990; 18:2163-70. PMID: 2336393
2. **Noble PA, Citek RW, Ogunseitan OA.** Tetranucleotide frequencies in microbial genomes. Electrophoresis 1998;19:528-35. PMID: 9588798
3. **Almeida JS, Carriço JA, Maretzek A, Noble PA, Fletcher M.** Analysis of genomic sequences by Chaos Game Representation. Bioinformatics 2001;17:429-37. PMID: 11331237
4. **Noble P.A., Pozhitkov A.** (2018) Cryptic sequencing features in the active postmortem transcriptome, BMC Genomics 19:675, PMID: 30217147