PROJET ALGORITHMES DE RECHERCHE - 2015 -

Codes circulaires de tétranucléotides autocomplémentaires

Professeur Christian MICHEL

Département d'Informatique, Université de Strasbourg

1. The graph theory of n-nucleotide circular codes

Throughout this section let $B = \{A, C, G, T\}$ be the set of nucleotide bases, where A stands for *Adenine*, C stands for *Cytosine*, G stands for *Guanine*, and T stands for *Thymine*. For $n \in \mathbb{N}$ with $n \geq 2$ an *n-nucleotide code* is a subset $X \subseteq B^n$. The following definition relates a directed graph to any n-nucleotide code. Recall from graph theory (Clark and Holton, 1991) that a *graph* $\mathcal{G}$ consists of a finite set of *vertices (nodes)* V and a finite set of *edges* E. Here, an edge is a set $\{v, w\}$ of vertices from V. The graph is called *oriented* if the edges have an orientation, i.e. edges are considered to be ordered pairs $[v, w]$ in this case.

Definition 1. *Let $X \subseteq B^n$ be an n-nucleotide code ($n \in \mathbb{N}$). We define a directed graph $\mathcal{G}(X) = (V(X), E(X))$ with set of vertices $V(X)$ and set of edges $E(X)$ as follows:*

- $V(X) = \{N_1...N_i, N_{i+1}...N_n : N_1N_2N_3...N_n \in X, 1 \leq i \leq n-1\}$
- $E(X) = \{[N_1...N_i, N_{i+1}...N_n] : N_1N_2N_3...N_n \in X, 1 \leq i \leq n-1\}$

The graph $\mathcal{G}(X)$ is called the representing graph of X or the graph associated to X.

Basically, the graph $\mathcal{G}(X)$ associated to a code X interprets n-nucleotide words from X in $(n-1)$ ways by pairs of i-nucleotides and $(n-i)$-nucleotides for $1 \leq i \leq n-1$. The following pictures give examples of codes and their representing graphs in the case of $n = 2$ (dinucleotide code), $n = 3$ (trinucleotide code) and $n = 4$ (tetranucleotide code).

Example 1. *In Figures 1-3, we show three examples of dinucleotide, trinucleotide and tetranucleotide codes and their representing graphs.*

As we can see the graph of the tetranucleotide code has four disjoint parts. However, note that two parts are built by vertices labeled with dinucleotides and two parts are built by vertices labeled with nucleotides and trinucleotides. These parts are called *components* of $\mathcal{G}$. Recall that a subset V' of the set of vertices V is called *connected* if for any two nodes