Space-efficient Representations for a set of k-mers

Ren Kimura[1a] **Sankardeep Chakraborty**[1] **Roberto Grossi**[2]
Giulia Punzi[2] **Kunihiko Sadakane**[1] **Wiktor Zuba**[3]

The University of Tokyo, Tokyo, Japan[1] (`renkimura@g.ecc.u-tokyo.ac.jp`[1a])
University of Pisa, Pisa, Italy[2] University of Warsaw, Warsaw, Poland[3]

We propose a new algorithm which takes DNA reads as input and outputs its SPSS [1].

Given a finite set of symbols Σ, a string (of length n) $S = S[1]S[2]...S[n]$ ($n \in \mathbb{N}$) is a finite sequence of symbols $S[i] \in \Sigma$. A string S' is called a substring (of length $j - i + 1$) of S *iff* there exists $i, j \in \mathbb{N}$ such that both $1 \leq i \leq j \leq n$ and $S' = S[i]S[i+1]...S[j] =: S[i...j]$ holds.

Let a set of strings $\mathcal{S}$ be DNA reads ($|\Sigma| = 4$). A k-mer is a length-k substring of $S \in \mathcal{S}$. A k-spectrum of $\mathcal{S}$, denoted by $\mathrm{sp}^k(\mathcal{S})$, is the set of all distinct k-mers of all $S \in \mathcal{S}$. Spectrum Preserving String Sets (SPSS) of $\mathcal{S}$ is any set of strings $\mathcal{S}'$ such that $\mathrm{sp}^k(\mathcal{S}') = \mathrm{sp}^k(\mathcal{S})$, where $\mathcal{S}$ is the original DNA reads.

A dimension-k node- (resp. edge-) centric de Bruijn graph of $\mathcal{S}$ is a directed graph $G(V, E)$, where $V = \mathrm{sp}^k(\mathcal{S})$ and $E = \{(u, v) | u \neq v \in V, u[2...k] = v[1...k-1]\}$ (resp. $V = \mathrm{sp}^{k-1}(\mathcal{S})$ and $E = \{(u, v) | u \neq v \in V, u[1...k-1]v[k-1] \in \mathrm{sp}^k(\mathcal{S}), u[2...k-1] = v[1...k-2]\}$). In a node-centric de Bruijn graph $G(V, E)$, a finite sequence of nodes $v_1, ..., v_h \in V$ s.t. $h \geq 2$ and $\forall i \neq j \; v_i \neq v_j$ and $(v_1, v_2), ..., (v_{h-1}, v_h) \in E$ is a path (resp. cycle) of length h if $(v_h, v_1) \notin E$ (resp. $(v_h, v_1) \in E$). If $h = 1$, we handle it as a path of length 1.

The main focus of our study is to establish a new algorithm which finds a SPSS of DNA reads and encodes it in a space-efficient way. We developed a node-centric based approach, having a natural contrast with Eulertigs [2] by Schmidt and Alanko. Their edge-centric based approach reduces SPSS finding to Eulerian cycle problem. Eulertigs finds a set of paths.

A path of length h requires $h + k - 1$ symbols for storing, while a same length cycle needs only h symbols by using a circular string. Our algorithm finds cycles and paths, minimizing the number of paths at the same time.

We first construct the node-centric de Bruijn graph $G(V, E)$. Next we transform G into a bipartite graph $G'(V_{\mathrm{L}} \cup V_{\mathrm{R}}, E')$, where there exist bijections $f_{\mathrm{L}} : V \to V_{\mathrm{L}}$ and $f_{\mathrm{R}} : V \to V_{\mathrm{R}}$, and $E' = \{(f_{\mathrm{L}}(u), f_{\mathrm{R}}(v)) | (u, v) \in E\}$. Then, we find a maximum bipartite matching in G'. This matching provides a unique decomposition of G into paths and cycles. We observed that the number of paths in such a decomposition is guaranteed to be the minimum possible.

References

[1] Amatur Rahman and Paul Medevedev. Representation of k-mer sets using spectrum-preserving string sets. *Journal of Computational Biology*, 28(4):381–394, 2021.

[2] Sebastian Schmidt and Jarno N Alanko. Eulertigs: minimum plain text representation of k-mer sets without repetitions in linear time. *Algorithms for Molecular Biology*, 18(1):5, 2023.