Traversing the k-mer Landscape of NGS Read Datasets for Quality Score Sparsification

Jacopo Notarstefano
jacopo.notarstefano [at] gmail.com

May 19, 2014

The DICT algorithm

Algorithm 1 DICT

Input: C, k, r

Output: D

    $D \leftarrow \{\}$

    $A \leftarrow [0, \ldots, 0] \in \mathbb{N}^{4^k}$

    for $\mathrm{x} \in C_k$ **do**

      $A[\mathrm{x}] + +$

    for $\mathrm{x} \in [4^k]$ **do**

      if $A[\mathrm{x}] \geq r$ **then**

        $D.\mathrm{append}(\mathrm{x})$

    return D

The MarkKmer algorithm

Algorithm 2 MarkKmer

Input: x, D

Output: M

The MARKREAD algorithm

Algorithm 3 MARKREAD

Input: γ, D
Output: $\mathcal{M}$

Algorithm 4 SPARSIFYRQ

Input: $\gamma, Q, D, Q_{\text{threshold}}$
Output: Q'
1: $Q' \leftarrow Q$
2: $\mathcal{M} \leftarrow \text{MARKREAD}(\gamma, D)$
3: **for** i $\in [\text{length}(\gamma)]$ **do**
4:    **if** $(Q_i > Q_{\text{threshold}}) \vee (\mathcal{M}_i = \text{true})$ **then**
5:       $Q'_i \leftarrow Q_{\text{threshold}}$
6: **return** Q'