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

Jacopo Notarstefano
jacopo.notarstefano [at] gmail.com

May 19, 2014

Phred quality score

Definition (Phred quality score)

Let P be the probability that a base call is incorrect. Then we define its *Phred quality score* Q as

$$Q = -10 \log_{10} P.$$

This quantity is encoded in the FASTQ format as a single byte, where the character `'!'` represents the lowest quality while `'~'` is the highest.

The FASTQ format

The DICT algorithm

Algorithm 1 DICT

Input: C, k, r

Output: D

```
1: D ← {}
2: A ← [0, ..., 0] ∈ ℕ^(4^k)
3: for x ∈ C_k do
4:     A[x]++
5: for x ∈ [4^k] do
6:     if A[x] ≥ r then
7:         D.append(x)
8: return D
```

The MarkKmer algorithm

Algorithm 2 MarkKmer

Input: x, D

Output: M

1: **if** $\Delta(x, D) > 1$ **then**
2: $\quad M \leftarrow [\text{false}, \ldots, \text{false}] \in \{\text{true}, \text{false}\}^k$
3: **else**
4: $\quad M \leftarrow [\text{true}, \ldots, \text{true}] \in \{\text{true}, \text{false}\}^k$
5: $\quad$ **for** $y \in D \mid \Delta(x, y) = 1$ **do**
6: $\quad\quad$ **for** $i \in [k]$ **do**
7: $\quad\quad\quad$ **if** $x_i \neq y_i$ **then**
8: $\quad\quad\quad\quad M_i \leftarrow \text{false}$
9: **return** M

The MarkRead algorithm

Algorithm 3 MarkRead

Input: γ, D

Output: $\mathcal{M}$

1: // Let x^a be the k-mer in γ starting at a.
2: // Cover γ by k-mers $\{x^{a_1}, \ldots, x^{a_n}\}$.
3: **for** $i \in [n]$ **do**
4: $\quad M^i \leftarrow \textsc{MarkKmer}(x^{a_i}, \mathsf{D})$
5: $\quad \overline{M}^i \leftarrow [\mathsf{false}, \ldots, \mathsf{false}] \in \{\mathsf{true}, \mathsf{false}\}^{\mathsf{length}(\gamma)}$
6: $\quad$ **for** $j \in [k]$ **do**
7: $\quad\quad \overline{M}^i_{j+a_i-1} \leftarrow M^i_j$
8: $\mathcal{M} \leftarrow \overline{M}^1 \vee \cdots \vee \overline{M}^n$
9: **return** $\mathcal{M}$

The SparsifyRQ algorithm

Algorithm 4 SparsifyRQ

Input: $\gamma, \mathsf{Q}, \mathsf{D}, \mathsf{Q}_{\mathsf{threshold}}$

Output: Q'

1: $\mathsf{Q}' \leftarrow \mathsf{Q}$
2: $\mathcal{M} \leftarrow \textsc{MarkRead}(\gamma, \mathsf{D})$
3: **for** $\mathsf{i} \in [\mathsf{length}(\gamma)]$ **do**
4: **if** $(\mathsf{Q}_\mathsf{i} > \mathsf{Q}_{\mathsf{threshold}})$ **or** $(\mathcal{M}_\mathsf{i} = \mathsf{true})$ **then**
5: $\mathsf{Q}'_\mathsf{i} \leftarrow \mathsf{Q}_{\mathsf{threshold}}$
6: **return** Q'