Minimum number of special characters in a merged string

Jacob L. Fine

Jun 19th, 2024

Suppose we are merging pairs of k-mers of equal length l, where each pair of k-mers may overlap with each other by k characters. And each k-mer contains a subset of characters which we denote 'special characters', such that the total number of characters in each k-mer is at least p.

We are specifically interested in finding a lower bound on the relative proportion of special characters in the merged string, accounting for the fact that the special characters may fall entirely in the overlap region k. We will denote this as ℓ_{min}. This quantity may be of interest if we want to ensure that merging k-mers does not cause the proportion of special characters in each k-mer to drastically decrease, and want to restore our confidence in our methods by finding a lower bound.

To start, we consider the length of the merged k-mers L, which is

$$L = 2l - k$$

since each k-mer has length n, and the overlap region is of length k.

We will then consider the 'worst-case' scenario, where all special characters are found in the overlap region. This entails that $k = p$ gives rise to the worse-case scenario. Thus, the special characters in each k-mer are identical, which entails that each k-mer has no unique special characters.

The proportion of the merged k-mer of length L that contains the unique characters is therefore p/L, and since $k = p$, we can write that

$$\ell_{min} = \frac{p}{2l - p}$$

This is the desired lower bound.

We may 'naively' (and incorrectly) think that since each k-mer contains at least p/l proportion of special characters that the merged k-mer's lower bound of the proportion of special characters occurs when there is no overlap, so $k = 0$. This would imply that the merged

k-mer has a proportion of special characters $\ell = 2p/2l = p/l$. But to show that p/l is not the true minimum, we would like to prove the inequality

$$\begin{aligned} &\frac{p}{2l-p} \leq \frac{p}{l} \\ \iff &2l \geq l + p \\ \iff &l \geq p \end{aligned}$$

Which must be true since p is defined as a length of a substring of l.