Bioinformatics, 38(20), 2022, 4848–4849
https://doi.org/10.1093/bioinformatics/btac604
Advance Access Publication Date: 5 September 2022
Letter to the Editor

Sequence analysis

A closed formula relevant to 'Theory of local k-mer selection with applications to long-read alignment' by Jim Shaw and Yun William Yu

John L. Spouge

Intramural Research Program, National Library of Medicine, Bethesda, MD 20894, USA

Contact: spouge@nih.gov

Associate Editor: Janet Kelso

Received on October 8, 2021; revised on July 11, 2022; editorial decision on August 30, 2022; accepted on September 1, 2022

1 Introduction

To handle the volume from next-generation sequencing data, modern sequence comparison often relies on summary sketches such as minimizers (Roberts *et al.*, 2004; Schleimer *et al.*, 2003), syncmers (Edgar, 2021) and minimally overlapping words (Frith *et al.*, 2021). Let us call a substring of length k within a sequence a k-mer. Sequence sketches are often the consequence of a rule f for selecting k-mers from a sequence. If the rule depends only on the k-mer under scrutiny and not on the sequence context (Shaw and Yu, 2021), call the rule 1-local. In this context, consider a long sequence where bases are mutated independently with probability θ. Eyeing applications where the mutated sequence is mapped onto the original sequence by k-mer matches, Theorem 2 of Shaw and Yu (2021) quantifies how frequently k-mers in a sketch are conserved under mutation of the original sequence.

Theorem 2 concerns itself with two vectors each of k probabilities, denoted $\Pr(\alpha(\theta,k))$ and $\Pr(f)$. To explain $\Pr(\alpha(\theta,k))$, call a run of α consecutive unmutated k-mers, i.e. a run of $k+\alpha-1$ unmutated letters, an α-run. On the one hand, $\Pr(\alpha(\theta,k))$ focuses on a letter chosen randomly from the middle of the long unmutated sequence. The k-mers containing the chosen letter include a total of $2k-1$ letters. Let $\Pr(\alpha(\theta,k)=\alpha)$ be the probability that the longest unmutated run within the $2k-1$ letters is an α-run. A classical formula (Shaw and Yu, 2021) determines $\Pr(\alpha(\theta,k)) = (\Pr(\alpha(\theta,k)=\alpha) : \alpha=1,2,\ldots,k)$ explicitly. To explain $\Pr(f)$, it relates α-runs directly to the sketch determined by the rule f. Consider an α-run ($\alpha=1,2,\ldots,k$) chosen randomly from the middle of a long random sequence. Let the α-run probability $\Pr(f,\alpha)$ be the probability that f selects at least one k-mer from the α-run. For any rule f, then, we can define the vector $\Pr(f) = (\Pr(f,\alpha) : \alpha=1,2,\ldots,k)$ of α-run probabilities. Loosely, $\Pr(f)$ quantifies the spread of the sketch with rule f: if f bunches the k-mers it selects too closely, the sketch is less likely to include a k-mer from a random α-run in the middle of a long sequence. Further details may be found in Shaw and Yu (2021).

Among other results in Shaw and Yu (2021), Theorem 2 gave a dot-product anticipating the practical performance of a sketch using a 1-local rule in mapping applications. In particular, the probability that a randomly chosen letter is within an unmutated k-mer selected by a rule f is

$$Cons(f,\theta,k) = \Pr(\alpha(\theta,k)) \cdot \Pr(f), \quad (1)$$

where the right side is the probability that the longest unmutated run containing the letter is an α-run times the probability that the rule f includes a k-mer from the α-run in the sketch, summed over $\alpha=1,2,\ldots,k$ by a dot-product. Details may be found in the original article (Shaw and Yu, 2021).

Shaw and Yu (2021) examine the consequences of Equation (1) for minimizers (Roberts *et al.*, 2004; Schleimer *et al.*, 2003) and for both closed and open syncmers (Edgar, 2021). Note that the rule for syncmers is 1-local, unlike the rule for minimizers. Section 4 in Shaw and Yu (2021) analyzes rules for selecting minimizers and syncmers under the assumption of a randomized hash function, neglecting equal k-mers as rare and thereby imposing a uniform distribution on the permutation ordering the relevant k-mer hashes. Recursions on four variables calculated $\Pr(f,\alpha)$, with variants tailored for the different rules under scrutiny. For closed syncmers, the recursion was equivalent to a closed formula for $\Pr(f,\alpha)$, but for minimizers and open syncmers, closed formulas appeared unavailable. From a practical point of view, the original four-variable recursions pose programming difficulties and they are computationally expensive for large parameter values. The purpose of this letter is to replace the recursion for minimizers with a simple explicit formula that alleviates these problems and to justify it directly with a combinatorial heuristic. The Section 3 points out that the formula is likely to generalize to other sketches.

2 Methods and results

Our set-up follows Section 2.2.1 in Shaw and Yu (2021). In windows consisting of w k-mers, therefore, the minimizers are the smallest k-mers, where a fixed random hash function determines the ordering O on the k-mers. Minimizers are the earliest sketch (Roberts *et al.*, 2004; Schleimer *et al.*, 2003) and they come with two very attractive properties. First, they have a window guarantee that every substring of length $w+k-1$ contains at least one minimizer. Second, the distance between consecutive minimizers follows a uniform first-occurrence distribution: their spacing is uniform on the set $\{1, 2, \ldots, w\}$ (Edgar, 2021).

For brevity, this letter identifies the k-mers with their random hashes, so for our purposes below a k-mer or a minimizer has length 1; a k-mer is positioned at the sequence index of its start; an α-run has length α; every w consecutive k-mers contains at least one minimizer; and if a minimizer is at index 0, the next minimizer has a random index chosen uniformly from the set $\{1, 2, \ldots, w\}$.

Let $\overline{F}_{w,\alpha}$ be the event where the random α-run of the Section 1 contains no minimizer. Every window of length w or more contains a minimizer, so on the one hand for $\alpha \geq w$, $\Pr(\overline{F}_{w,\alpha}) = 0$. For $1 \leq \alpha < w$, on the other hand, there is a rightmost minimizer M_- strictly to the left of the α-run. For convenience, set up a sequence coordinate system assigning index 0 to M_-. Let M_+ be the next minimizer to the right of M_-. The minimizer M_+ is at some uniformly distributed index $d \in \{1, 2, \ldots, w\}$ (Edgar, 2021). The α-run starts (by stationarity) at some uniformly distributed index $b \in \{1, 2, \ldots, d\}$ between M_- and M_+. The total number of configurations for the minimizer M_+ and the α-test window is therefore $\sum_{d=1}^{w} \sum_{b=1}^{d} 1 = \frac{1}{2} w(w+1)$.

On the event $\overline{F}_{w,\alpha}$, the α-run contains no minimizer, so M_+ must be strictly to the right of the α-run, i.e. $1 + \alpha \leq b + \alpha \leq d \leq w$. The total number of configurations allowed under $\overline{F}_{w,\alpha}$ for the minimizer M_+ and the α-run is therefore $\sum_{d=\alpha+1}^{w} \sum_{b=1}^{d-\alpha} 1 = \frac{1}{2}(w-\alpha)(w-\alpha+1)$. For minimizers, all distributions involved are uniform (in particular, the first-occurrence distribution of distance between consecutive minimizers), so the probabilities are proportional to the configuration counts. Thus,

$$\Pr(\overline{F}_{w,\alpha}) = \frac{(w-\alpha)(w+1-\alpha)}{w(w+1)}. \quad (2)$$

The present author and others (J.Shaw and Y.W.Yu, personal communication) performed extensive numerical computations looping over both α and k to compare Equation (2) with the recursion in Theorem 7 of Shaw and Yu (2021), confirming empirically that $\Pr(\overline{F}_{w,\alpha}) = 1 - \Pr(f, \alpha)$ for minimizers. Notably for $\alpha = 1$, Equation (2) yields $\Pr(\overline{F}_{w,1}) = (w-1)/(w+1)$, yielding the density of minimizers $1 - \Pr(\overline{F}_{w,1}) = 2/(w+1)$, a classical result (Roberts *et al.*, 2004; Schleimer *et al.*, 2003).

3 Discussion

Although the uniform first-occurrence distribution between consecutive minimizers simplifies formulas in Section 2, it is inessential to the heuristic there (J.Shaw and Y.W.Yu, personal communication). Our results therefore suggest the existence of a simple general formula for interconversion of first-occurrence distributions and α-run probabilities. Presently, the interconversion requires complicated recursive methods (Dutta *et al.*, 2022). The results presented may therefore be useful in accelerating the current interest and progress in understanding k-mer sketches (Belbasi *et al.*, 2022).

Acknowledgements

The author gratefully acknowledges useful conversations with Dr Martin C. Frith, Dr Jim Shaw and Dr Yun William Yu.

Funding

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Conflict of Interest: none declared.

Data availability

The article introduces no new data, so vacuously all links and identifiers for relevant data are present in the manuscript.

References

Belbasi,M. *et al.* (2022) The minimizer Jaccard estimator is biased and inconsistent. *bioRxiv*. https://doi.org/10.1101/2022.01.14.476226.

Dutta,A. *et al.* (2022) Parameterized syncmer schemes improve long-read mapping. *bioRxiv*. https://doi.org/10.1101/2022.01.10.475696.

Edgar,R. (2021) Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences. *PeerJ*, **9**, e10805.

Frith,M.C. *et al.* (2021) Minimally-overlapping words for sequence similarity search. *Bioinformatics*, **36**, 5344–5350.

Roberts,M. *et al.* (2004) Reducing storage requirements for biological sequence comparison. *Bioinformatics*, **20**, 3363–3369.

Schleimer,S. *et al.* (2003) Winnowing: local algorithms for document fingerprinting. In: *SIGMOD 2003*. ACM, San Diego, CA, pp. 76–85.

Shaw,J. and Yu,Y.W. (2021) Theory of local k-mer selection with applications to long-read alignment. *bioRxiv*. https://doi.org/10.1011/2021.05.22.445262.