Xin *et al. Algorithms Mol Biol* (2020) 15:10
https://doi.org/10.1186/s13015-020-00172-3

Algorithms for Molecular Biology

RESEARCH Open Access

Context-aware seeds for read mapping

Hongyi Xin[1,4], Mingfu Shao[2] and Carl Kingsford[3*]

Abstract

Motivation: Most modern seed-and-extend NGS read mappers employ a seeding scheme that requires extracting t non-overlapping seeds in each read in order to find all valid mappings under an edit distance threshold of t. As t grows, this seeding scheme forces mappers to use more and shorter seeds, which increases the seed hits (seed frequencies) and therefore reduces the efficiency of mappers.

Results: We propose a novel seeding framework, context-aware seeds (CAS). CAS guarantees finding all valid mappings but uses fewer (and longer) seeds, which reduces seed frequencies and increases efficiency of mappers. CAS achieves this improvement by attaching a confidence radius to each seed in the reference. We prove that all valid mappings can be found if the sum of confidence radii of seeds are greater than t. CAS generalizes the existing pigeonhole-principle-based seeding scheme in which this confidence radius is implicitly always 1. Moreover, we design an efficient algorithm that constructs the confidence radius database in linear time. We experiment CAS with *E. coli* genome and show that CAS significantly reduces seed frequencies when compared with the state-of-the-art pigeonhole-principle-based seeding algorithm, the Optimal Seed Solver.

Availability: https://github.com/Kingsford-Group/CAS_code

Keywords: Read mapping, Seeds, Error tolerance, Seed and extend

Introduction

Read mapping is used ubiquitously in bioinformatics. With $|\cdot|$ denoting the length of a string; $T[\cdot,\cdot]$ denoting the substring of a string T at a fixed interval; and $D(\cdot,\cdot)$ as the edit distance measurement between a string pair, commonly, read mapping is defined as follows:

Problem 1 (Read Mapping) *Given read R and reference T (usually with $|T| \gg |R|$), and an error tolerance threshold t, we say a substring of T at location $[l_1, l_2]$, i.e., $T[l_1, l_2]$, is a valid mapping of R if we have $D(R, T[l_1, l_2]) < t$.*

To efficiently map reads, modern mappers usually employ the *seed-and-extend* mapping strategy [1–4]: a mapper extracts a substring of R as a *seed*, s; iterates through all seed locations of s in T; at each seed location, performs sequence alignment of R against the surrounding text in T; reports alignments that have edit distances below t as valid mappings.

For mappers that use non-overlapping seeds, the number of seeds to extract from a read R is governed by the pigeonhole principle: to find all valid mappings of R, the mapper must divide R into at least t non-overlapping seeds. Otherwise, the mapper will not be able to consistently find all valid mappings of R in T. As t grows, the length of seeds is reduced. Using short seeds significantly increases the workload of a mapper [5, 6]. Shorter seeds appear more frequently in T, hence increasing the number of alignments while mapping a read. To improve the performance of mappers, it is desirable to use fewer non-overlapping seeds under a fixed t, which lets a mapper not only use fewer seeds, but also use longer seeds.

In this paper, we focus on improving seed-and-extend mappers that use non-overlapping seeds. We propose a

*Correspondence: carlk@cs.cmu.edu
[3] Computational Biology Department, Carnegie Mellon University, Pittsburgh 15213, USA
Full list of author information is available at the end of the article

novel seeding scheme, called *context-aware* seeds (CAS). CAS enables a mapper to use fewer than t seeds without missing any valid mappings. CAS attaches each seed s with a *confidence radius* score, c_s, with $c_s \geq 1$. Let S be a set of non-overlapping seeds from R. CAS ensures that as long as $\sum_{s \in S} c_s \geq t$, then S is sufficient to find all valid mappings of R under an error tolerance threshold of t. When S includes any seed s with $c_s > 1$, then $|S| < t$ and all valid mappings are secured with fewer-than-t seeds ($|S|$ denotes the number of seeds in S). In the worst case where $c_s = 1$ for all $s \in S$, CAS degenerates into the case governed by pigeonhole principle with $|S| = t$.

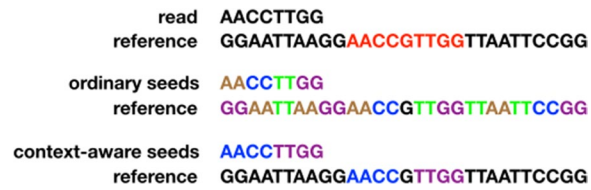

Fig. 1 Illustration of CAS. The upper part shows a read and a reference. Suppose that $t = 4$, i.e., we want to find all alignments of the read in the reference with fewer than 4 edits. There is only one such locally optimal alignment (marked as red). The middle part shows the seed extraction result with the pigeonhole principle, which splits the read into $t = 4$ seeds. This gives many seed locations and thus many alignments. With CAS (in the lower part), we can split the read into 2 long seeds while still guarantee to find all valid mappings. The two long seeds together have a total seed frequency of 2, drastically reducing the number of alignments

Algorithm 1: Linear Time Algorithm for Problem 2

Input: Suffix trie $Trie = (V, E)$ and the edit-distance threshold t
Output: X_v and Y_v for each $v \in V$

1. Assign ranks for all nodes using breadth-first search on T
2. Initialize X_r and Y_r for root $r \in V$.
3. **FOR** each node $v \in V$ in ascending order:
4. Initialize pointer $k_v = 0$ for arrays X_v and Y_v.
5. Initialize arrays $X_{v'}$ and $Y_{v'}$ for each child v' of v as empty arrays.
6. Initialize pointer $k_{v'} = -1$ for arrays $X_{v'}$ and $Y_{v'}$ for each child v' of v.
7. **FOR** $k = 0 \rightarrow |X_v|$:
8. **LET** $u := X_v[k]$.
9. **FOR** each child u' of u:
10. **FOR** each child v' of v:
11. // Compute $D_1 = Y_v[k] + \delta$, i.e., $D_1 = D(v, u) + \delta$.
12. **IF** $\sigma(u, u') \neq \sigma(v, v')$ **THEN** $\delta = 1$ **ELSE** $\delta = 0$.
13. $D_1 = Y_v[k] + \delta$.
14. // Compute D_2. Check if $u' \in X_v$.
15. Increase k_v until $X_v[k_v] \geq u'$.
16. **IF** $X_v[k_v] = u'$ **THEN** $D_2 = Y_v[k_v] + 1$ **ELSE** $D_2 = \infty$.
17. // Compute D_3. Check if $u \in X_{v'}$.
18. Increase $k_{v'}$ until $X_{v'}[k_{v'}] \geq u$.
19. **IF** $X_{v'}[k_{v'}] = u$ **THEN** $D_3 = Y_{v'}[k_{v'}] + 1$ **ELSE** $D_3 = \infty$.
20. // Update neighbor lists.
21. $D(v', u') = \min\{D_1, D_2, D_3\}$.
22. **IF** $D(v', u') < t$ **THEN** append u' to $X_{v'}$; append $D(u', v')$ to $Y_{v'}$.
23. **END FOR**
24. **END FOR**
25. **END FOR**
26. **END FOR**

Figure 1 compares CAS and the pigeonhole-principle-based seeds. In this example, we want to find all mappings of a read AACCTTGG under an error tolerance threhsold of $t = 4$. Assume that we have verified that each of the two CAS seeds AACC and TTGG has confidence radii of $c_s = 2$. Therefore CAS can be guaranteed

to find all valid mappings with just these two seeds, as $\sum_s c_s = 4 \geq t$. Using the pigeonhole principle, however, a mapper needs to select $t = 4$ non-overlapping seeds. It forces the mapper to pick short and repetitive seeds, making the mapper perform more local alignments.

We establish the theoretical foundation of CAS and demonstrate that with CAS future mappers can map reads more efficiently using fewer, longer and less frequent seeds without losing valid mappings. We also propose a suffix-trie-based CAS database construction algorithm that builds a CAS database from T in linear time, based on which we design a greedy CAS seeding algorithm that extracts CAS from reads. We test the greedy CAS seeding algorithm against a state-of-the-art pigeonhole-principle-based seeding algorithm, Optimal Seed Solver (OSS), on an *E. coli* dataset.

Context-aware seeds

CAS reduces seed usage in read mapping by introducing a novel metric for seeds in T, the confidence radius. A seed s in T has a confidence radius c_s if all substrings in T whose edit distance is smaller than c_s must occur in T within a small window where s occurs. The size of the window equals to extending the length of s by $c_s - 1$ letter(s) at both ends. For example, seed AACC in T from Fig.1 has a confidence radius of 2. Any substring in T whose edit distance to AACC equals 1 (e.g., AAC, ACC, GAACC, AACCG) locates within the 1-letter extended window of AACC (GAACCG). The confidence radius of each possible seed in T can be computed by profiling T (see the next Section). CAS guarantees that all valid mappings of a read R can be located, as long as the seeds s extracted from R collectively have a confidence radius of $\sum_s c_s \geq t$. Below, we give the formal definition of CAS and prove the correctness of CAS.

Let s be a string in T and $[l_1, l_2]$ be a pair of locations in T. We say string $T[l_1, l_2]$ is in the *vicinity* of s under an integer c, if s appears in an interval $[l_{s1}, l_{s2}]$ in T $(T[l_{s1}, l_{s2}] = s)$, where $l_1 - c < l_{s1} < l_{s2} < l_2 + c$. Furthermore, let seed s be a substring of R at $[l_{r1}, l_{r2}]$ $(s = R[l_{r1}, l_{r2}])$ and let $T[l_1, l_2]$ be a valid mapping of R. We say $T[l_1, l_2]$ is in the *vicinity of s with regard to R* under c, if string $T[l_1 + l_{r1}, l_1 + l_{r2}]$ is in the vicinity of s under c. If a valid mapping $T[l_1, l_2]$ is in the vicinity of s with respect to R under t, then $T[l_1, l_2]$ can be discovered by locally aligning R against the surrounding text in T at each seed location of s.

The pigeonhole principle states that by dividing R into a set of t non-overlapping seeds, denoted by S, then for all intervals $[l_1, l_2]$ where $T[l_1, l_2]$ is a valid mapping of R, there must be $s \in S$ where $T[l_1, l_2]$ is in the vicinity of s with regard to R.

CAS seeks to retain the seed vicinity guarantee of the pigeonhole principle, where all valid mappings of a read R are in the vicinity of its seeds with regard to R under t, with fewer than t seeds. Given two substrings s and s' of T and an edit-distance threshold t, we say s' is a neighbor of s if $D(s, s') < t$. Assume that s' is a neighbor of s under t, CAS defines s' as a *trivial neighbor* of s, if and only if for any location interval $[l_1, l_2]$ where $T[l_1, l_2] = s'$ and $T[l_1, l_2]$ is in the vicinity of s under $D(s, s')$. Otherwise CAS defines s' as a *nontrivial neighbor* of s. Finally, CAS defines *the confidence radius* c_s of s as the minimum edit-distance between s and all nontrivial neighbors of s. Since a seed is trivial to itself and is at least 1-edit-distance away from any other string, we have $c_s \geq 1$ for any seed s.

We now give the central theorem of CAS, the theoretical foundation that enables seed-and-extend mappers to find all valid mappings using fewer than t seeds.

Theorem 1 *Let S be a set of non-overlapping seeds of a read R. Assume $\sum_{s \in S} c_s \geq t$. Then for any reference string $T[l_1, l_2]$ where $D(R, T[l_1, l_2]) < t$, there must be a seed $s \in S$, where $T[l_1, l_2]$ is in the vicinity of s with regard to R under t.*

Proof Assume that $T[l_1', l_2']$ is a valid mapping of R, where $D(R, T[l_1', l_2']) < t$. Further assume that $T[l_1', l_2']$ is not in the vicinity, with regard to R under t, of any $s \in S$. In the minimum-edit-distance alignment between R and $T[l_1', l_2']$, assume that the non-overlapping seeds s_1, s_2, ..., s_n of R are aligned to the non-overlapping segments s_{T1}, s_{T2}, ..., s_{Tn} of $T[l_1', l_2']$, with $n = |S|$. Since $T[l_1', l_2']$ is not in the vicinity, with regard to R under t, of any $s \in S$, none of the seeds s_i matches exactly to its counterpart sequence S_{Ti}; and none of the sequences s_{Ti} is a trivial neighbor under the confidence radius c_{si} of its counterpart seed s_i. Because c_{s_i} is the minimum edit-distance between s_i and any of its nontrivial neighbors, we have $D(R, T[l_1', l_2']) \geq \sum_i D(s_i, s_{T_i}) \geq \sum_s c_s \geq t$. $D(R, T[l_1', l_2']) \geq t$ contradicts the assumption that $T[l_1', l_2']$ is a valid mapping of R. Therefore such $T[l_1', l_2']$ does not exist.

We call Theorem 1 *the weighted Pigeonhole principle*. The confidence radius of each seed serves as the weight of the seed in computing the overall error tolerance of the seed set. □

Construction of confidence radius database

The confidence radius c_s of each seed s is stored in a table, called *the confidence radius database*. The confidence radius database only needs to be constructed once offline for a reference T.

Computing c_s of seed s involves finding the minimum edit distance to its nontrivial neighbors. Below we propose an algorithm that constructs the confidence radius database in $O(|\Sigma|^2 \cdot M)$ time, where Σ is the alphabet set of T and M is the total number of neighbors of all strings in T (up to length P). Note that c_s need not be a tight lower bound. Ordinary seeds can be perceived as CAS with confidence radii of $c_s = 1$. As will be discussed later, tighter bounds (greater confidence radii) take longer to compute. We limit the search space in computing the confidence radii of seeds by introducing a maximum allowed confidence radius threshold θ. Seeds whose exact confidence radii (tight bounds) that are greater than θ are set to $c_s = \theta$. In this section, we assume both P and θ are constants.

The confidence radius database is constructed in two steps: first, we construct a neighbor database, which stores all neighbors of all seeds (up to length P) under the confidence radius threshold θ; then, we find the confidence radius of each from its neighbors. We prove that both steps can be done in $O(|\Sigma|^2 \cdot M)$ time.

Construction of the neighbor database

To find all neighbors of all substrings in T (up to a maximum length P), we first build a P-level suffix trie of T, then find all neighbors of each seed in *Trie* by systematically traversing the suffix trie in a top-down manner. Formally, let $Trie = (V, E)$ be a suffix trie of T of a maximum depth of $P + \theta$. Let $r \in V$ be the root of *Trie*. Each node represents a substring in T, i.e., the string obtained by concatenating the letters on edges along the path from r to v. We denote the edit distance between these two substrings corresponding to u and v as $D(u, v)$. We aim to solve the following problem:

Problem 2 *Given a suffix trie* $Trie = (V, E)$ *and an integer* θ*, compute all pairs of nodes* $u, v \in V$ *such that* $D(u, v) \le \theta$.

For any $v \in V$, $p(u)$ denotes the parent node of v in *Trie*. $\sigma(p(v), v)$ denotes the letter on the edge between v and $p(v)$, i.e., $(p(v), v) \in E$. We have the following lemmas.

Lemma 1 *Let* $u, v \in V$. *Then* $D(u, v) \le \theta$ *only if* $D(p(u), p(v)) \le \theta$.

Proof Proved in Landau and Vishkin [7] by enumerating and validating all possible scenarios. □

Lemma 2 *Let* $u, v \in V$. *We have*

$$D(u, v) = \min \begin{cases} D(p(u), p(v)) + \delta_{uv} \\ D(p(u), v) + 1 \\ D(u, p(v)) + 1 \end{cases}$$

where $\delta_{uv} = 1$ *if* $\sigma(p(u), u) \neq \sigma(p(v), v)$ *and* $\delta_{uv} = 0$ *if* $\sigma(p(u), u) = \sigma(p(v), v)$.

Proof This follows from the dynamic programming algorithm for the edit distance problem. □

Lemma 1 shows that nodes are neighbors only if their parents are neighbors. Hence the neighbors of a child node must be the children of the neighbors of its parent node. Lemma 2 further shows that the edit distance between two children nodes can be computed in constant time, given the edit distances between one child and the parent node of the other child, as well as the edit distance between the two parent nodes.

We construct the neighbor database by traversing *Trie* as follows: First, assign each node in V an integral *rank* from $\{1, 2, \cdots, |V|\}$ following a top-down, left-to-right order. The root r of *Trie* has rank of 1, and then the children of children of r have ranks of $2, 3, \cdots$, from the leftmost child to the rightmost child, and so on. Nodes that are deeper in *Trie* rank higher. Among nodes of the same depth, children of a higher ranking parent node rank higher. A breadth-first-search traversal of *Trie* ranks all nodes.

For any $v \in V$, we define $X_v := \{u \in V \mid D(u, v) \le \theta\}$ as the set of neighbors of v, including v itself, and define $Y_v := \{D(u, v) \mid u \in X_v\}$ as the accompanying edit-distance set of X_v. For every neighbor node u in X_v, Y_v provides the edit distance between u and v. We compute X_v and Y_v for each node $v \in V$ from low ranking nodes to high ranking nodes. Both X_v and Y_v are implemented as arrays.

The algorithm for constructing the neighbor database is summarized in Algorithm 1. We iterate through all nodes by rank from low to high. For each node $v \in V$, we iterate through all children of v. For each children node v' of v, we compute $X_{v'}$ and $Y_{v'}$ of v' based on the previously computed X_v and Y_v of v. Figure 2 illustrates the process of validating a candidate neighbor u' of another

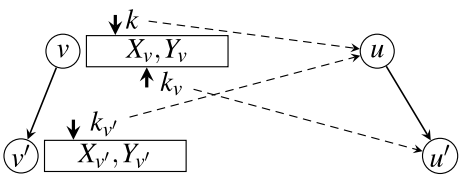

Fig. 2 Illustration of processing a single node v (i.e., line 4 to line 17 of Algorithm 1)

node v', based on the information of its parent node v and the neighbor u of v, where u is also the parent of u' (lines 4–25 in Algorithm 1). We prove that this algorithm maintains the following three invariants:

1 For any node $v \in V$, array X_v is always sorted according to their ranks, i.e., nodes that are added to X_v are always in ascending order *w.r.t.* their ranks.
2 Right before processing node v (i.e., before line 4 of Algorithm 1), X_v and Y_v are already computed and sorted *w.r.t.* their ranks.
3 Right after processing node v (i.e., after line 17 of Algorithm 1), $X_{v'}$ and $Y_{v'}$ are computed and sorted *w.r.t.* their ranks for each child v' of v.

The initialization step Algorithm 1 (line 2) computes X_r and Y_r for root node r. Its neighbors include all nodes whose depth in *Trie* is no greater than θ. The edit distance between r to a neighbor node u is simply the depth of u minus 1 (we assume that root r is at depth 1). Root r is also in X_r with $D(r, r) = 0$. Clearly, the first and the second invariant hold for root r.

In the main loop (lines 3–26), for a node $v \in V$, Algorithm 1 iterates through all of its children. For a child v' of v, lines 4–25 compute $X_{v'}$ and $Y_{v'}$ of v'. Line 4–6 initialize the pointers that will be used to fetch the edit distances $D(v, u')$ and $D(v', u)$, which are stored in Y_v and the partially computed $Y_{v'}$, respectively. Because u ranks higher than u', by the time of computing $D(v', u')$, $D(v', u)$ is already computed and stored in $X_{v'}$. $D(v', u')$ is then computed according to Lemma 2. Specifically, pointer k tracks the position of u in array X_v (i.e., the index of u in array X_v); pointer k_v tracks the position of u' in array X_v; and pointer $k_{v'}$ tracks the position of u in array $X_{v'}$. Line 11 computes $D_1 := D(v, u) + \delta$, in which $D(v, u)$ is fetched from Y_v indexed by k. Line 12 computes $D_2 := D(v, u') + 1$, in which $D(v, u')$ is fetched from Y_v indexed by k_v. Line 13 computes $D_3 := D(v', u) + 1$, in which $D(v', u)$ is fetched from $Y_{v'}$ indexed by $k_{v'}$. Line 14 computes $D(v', u') := \min\{D_1, D_2, D_3\}$; adds u' to $X_{v'}$ and adds $D(v', u')$ to $Y_{v'}$ if $D(v', u') < t$.

Algorithm 1 maintains the first invariant. For each child v' of v, assuming X_v is sorted, then neighbors are also added to $X_{v'}$ in a sorted manner, as Algorithm 1 iterates through neighbors ordered by X_v. Since X_r is sorted for root r, given the inductive nature of Algorithm 1, we conclude that X_v must be sorted for any $v \in$ *Trie*.

Algorithm 1 maintains the third invariant. According to Lemma 1, a node $u' \in X_{v'}$ requires $u \in X_v$ for their parents u and v. Any node $\bar{u}'$ whose parent $\bar{u} \notin X_v$ results in $\bar{u}' \notin X_{v'}$. Algorithm 1 iterates through all u in X_v. Therefore, after line 17, all neighbors of child v' must have been found, assuming the second invariant holds. The second invariant holds because all neighbors of r are correctly defined during initialization. As the algorithm propagates, because of the inductive nature of Algorithm 1, the second invariant holds.

Let M denote the member size of set $\{(u, v) \mid D(u, v) \le \theta\}$. The complexity of Algorithm 1 is $O(|\Sigma|^2 \cdot M)$.

Theorem 2 *Algorithm 1 computes X_v and Y_v for each $v \in V$ in $O(|\Sigma|^2 \cdot M)$ time.*

Proof For each $v \in V$, lines 4–25 compute $X_{v'}$ and $Y_{v'}$ for each child v' of v in $O(|X_v| \cdot |\Sigma|^2 + \sum_{v':p(v')=v} |X_{v'}|)$ time. Since pointers of k_v and $k_{v'}$ can only move forward, lines 14–19 cost $|X_v| + \sum_{v':p(v')=v} |X_{v'}|$ operations. Operations in lines 11–22 cost constant time. Hence, lines 7–25 cost $O(|X_v| \cdot |\Sigma|^2)$ operations, as the number of children of each node is bounded by $|\Sigma|$. The overall run time of Algorithm 1 is thus bounded by $\sum_{v \in V} O(|X_v| \cdot |\Sigma|^2 + \sum_{v':p(v')=v} |X_{v'}|) = O(|\Sigma|^2 \cdot M)$. □

With $|\Sigma|$ being a small constant (for example $\Sigma = \{A, C, G, T\}$ for DNA analysis), Algorithm 1 finds all M neighbor pairs in *Trie* in $O(M)$ time.

Computing the confidence radius among nontrivial neighbors

The neighbor database stores both the trivial and nontrivial neighbors of each seed. However, CAS only requires the minimum edit distance to the nontrivial neighbors of each seed. In order to derive the confidence radius of each seed, we propose an augmentation to Algorithm 1, such that it computes the minimum edit distance to nontrivial neighbors while constructing the neighbor database. We prove that the augmentation does not increase the time complexity of Algorithm 1.

Within the neighbor array X_v of a node v, let the sub-array X_v^0 store all trivial neighbors and X_v^1 store all nontrivial neighbors, where $X_v = X_v^0 \cup X_v^1$. By definition, the confidence radius of v is computed as $c_v := \min_{u \in X_v^1} D(u, v)$. To compute c_v, instead of finding all nontrivial neighbors, X_v^1, we compute a subset $X_v^2 \subset X_v^1$, where $\min_{u \in X_v^2} D(u, v) = \min_{u \in X_v^1} D(u, v)$.

Let u be a neighbor of v; we say u is an *immediate* neighbor of v if u is a substring, or a superstring, or an overlapping string of v; otherwise we say u is a *non-immediate* neighbor of v (see Fig. 3 for examples). Immediate neighbors are not necessarily trivial neighbors. If u is a trivial neighbor of v, by definition, then u must be an immediate neighbor of v. However, the opposite is not necessarily true, i.e., u could be a substring of v (an immediate neighbor) yet u is nontrivial to v. Substring u may appear at more locations in T than v does. It is easier

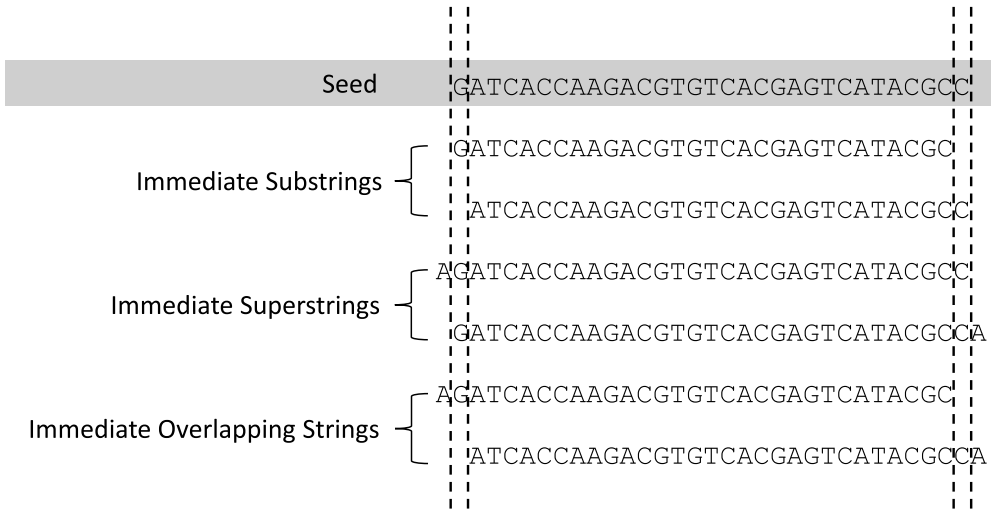

Fig. 3 Examples of trivial neighbors of a seed, including substrings, superstrings, and overlapping strings of this seed

to determine whether u is an immediate neighbor to v than whether u is a trivial neighbor to v.

Let X_v^2 be the set of non-immediate neighbors of a node v. The minimum edit distance from v to nontrivial neighbors of v equals to the minimum edit distance between v to neighbors in X_v^2. We prove this in Theorem 4. To prove Theorem 4, we first prepare the following two lemmas.

Lemma 3 *If u is a superstring of v, then u is a trivial neighbor of v.*

Proof Since u is a superstring of v, for any location $[l_1, l_2]$ of u, there exists $[l_1, l_2]$ where $T[l_1, l_2] = v$ and $l_1 - D(u, v) \leq l_1 < l_2 \leq l_2 + D(u, v)$. By definition, u is a trivial neighbor of v. □

Lemma 4 *If u is a substring or an overlapping string of v and u is a nontrivial neighbor of v, then there exists $w \in Trie$, where w is neither an immediate neighbor nor a trivial neighbor of v, with $|w| = |v|$ and $D(v, w) \leq D(v, u)$.*

Proof Since u is a nontrivial neighbor of v, there exists $[l_1, l_2]$, where $T[l_1, l_2] = u$ but $T[l_1, l_2]$ is not in the $D(v, u)$-edit vicinity of v. We extract a substring w within $T[l_1 - D(u, v), l_2 + D(u, v)]$, where w locally and optimally aligns to v in $T[l_1 - D(u, v), l_2 + D(u, v)]$, with $|w| = |v|$, as shown in Fig. 4. Then w must be a nontrivial neighbor of v since $T[l_1, l_2]$ is not in the $D(u, v)$-edit vicinity of v. Because w is optimally aligned to v within $[l_1 - D(u, v), l_2 + D(u, v)]$, we have $D(w, v) \leq D(u, v)$. □

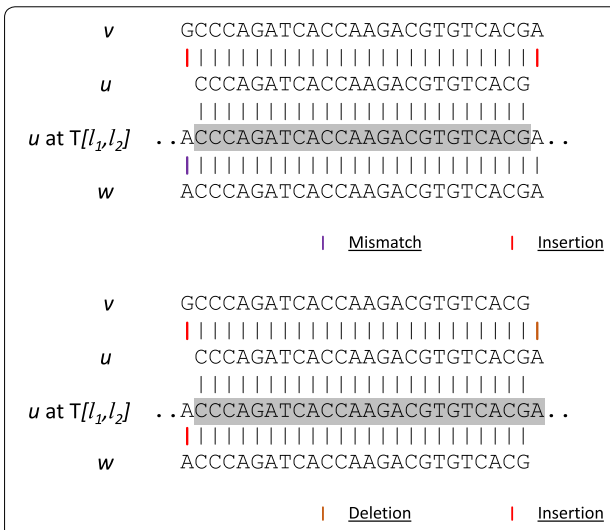

Fig. 4 Illustration of Lemma 4. The figure to the left shows an example where u is a substring of v, while the figure to the right shows an example where u is an overlapping string of v. Notice that in both figures, w is always optimally aligned to v

By combining Lemmas 3 and 4 we prove the following theorem.

Theorem 3 $c_v = \min_{u \in X_v^2} D(u, v)$, *where X_v^2 is the set of non-immediate neighbors of v.*

Proof Lemmas 3 and 4 state that for any nontrivial immediate neighbor u of seed v, there must exist a nontrivial and non-immediate neighbor w of v where $D(w, v) \leq D(u, v)$. Therefore, by definition, we have $c_v = \min_{u \in X_v^2} D(u, v)$. □

We find the immediate neighbors, X_v^3, of each node $v \in Trie$, by checking if a neighbor $u \in X_v$ is a immediate substring, superstring or overlapping string of v. We associate with each node v a new vector $Z_v := \{F(v,u) \mid u \in X_v\}$, where $F(v,u)$ stores the information of whether $u \in X_v^3$. With $X_v^2 = X_v \setminus X_v^3$, the updated workflow is illustrated in Fig. 5.

Computation of $F(v, u)$ can be piggybacked on top of computing $D(v, u)$ in Algorithm 1. Given u and v, $F(v, u)$ stores whether v and u possess any of the below *immediate conditions*:

1 u is a prefix of v.
2 u is a suffix of v.
3 v is a prefix of u.
4 v is a suffix of u.
5 u is neither a prefix nor a suffix but a substring of v.
6 v is neither a prefix nor a suffix but a substring of u.
7 A prefix of u is a suffix of v.
8 A suffix of u is a prefix of v.

From above immediate conditions, we deduce the immediate relationship between v and u. With conditions 1–6, we can infer the superstring-substring relationship. With Condition 7–8, we can infer the overlapping relationship. If v and u qualifies none of the above immediate conditions, then they must be non-immediate neighbors.

For simplicity, we initialize each node as satisfying immediate conditions 1, 2, 3 and 4 to itself. We initialize the root node r as a prefix to any of its neighbors; and any neighbors of r as a suffix to r. Finally, r is not an overlapping string or a substring of any neighbor.

$F(u, v)$ can be computed in constant time if $F(p(v), p(u))$, $F(p(v), u)$ and $F(v, p(u))$ are known. For example, in Fig. 5, $F(v', u')$ satisfies condition 1, only if (a) $v' = u'$ or (b) $F(v', u)$ satisfies condition 1. $F(v', u')$ satisfies condition 2, only if (a) $u' = v'$ or (b) $F(v, u)$ satisfies condition 2 and $\sigma(u, u') = \sigma(v, v')$. Conditions 3 and 4 are mirror cases of conditions 1 and 2, respectively with v and u, v' and u' trading places. $F(v', u')$ satisfies condition 5, only if (a) $F(v', u)$ satisfies condition 5 or (b) $F(v', u)$ satisfies condition 2, while $v' \neq u'$ and v is not root. Condition 6 is a mirror case of condition 5. $F(v', u')$ satisfies condition 7 only if (a) $F(v, u')$ satisfies condition 7 or (b) $F(v, u')$ satisfies condition 2, while $v \neq u'$ and v is not root. Condition 8 is a mirror case of condition 7.

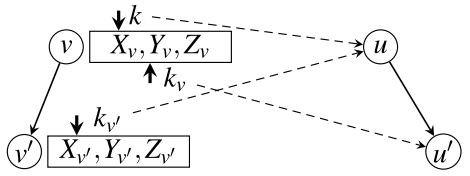

Fig. 5 Illustration of adding $Z_v := \{F(u,v) \mid u \in X_v\}$ to each node

The computation of $F(\cdot, \cdot)$ is piggybacked on top of the computation of $D(\cdot, \cdot)$, as both methods use dynamic programming. Both computations share the same structure: To compute $D(\cdot, \cdot)$ and $F(\cdot, \cdot)$ between two child nodes, we must aquire $D(\cdot, \cdot)$ and $F(\cdot, \cdot)$ between all child-parent and parent-parent node pairs; and then derive $D(\cdot, \cdot)$ and $F(\cdot, \cdot)$ between the two child nodes in constant time. As a result, piggybacking the computation of immediateness does not increase the complexity of Algorithm 1.

Finally, the confidence radius of node v equals $\min D(v, u)$ where $u \in X_v^2$, where $F(v, u)$ does not satisfy any of the immediate conditions. The confidence radius of a node can be found by simply scanning its neighbor array, which finishes in linear time. The overall complexity of constructing the confidence radius database is still $O(|\Sigma|^2 \cdot M)$.

Maintaining the confidence radius metadata

The confidence radii of seeds is stored in a $|T|$-by-P table, where P is user-provided. Fig. 6 demonstrates an example confidence radius table. The $[x, y]$ entry of the table stores the confidence radius of seed $T[x, x + y]$. Overall, the confidence radius table has a space complexity of $O(|T| \cdot P)$, with $|T| \gg P$ in practice.

In read mapping, it is not always necessary to maintain the confidence radii for every seed. Instead, it is often sufficient to just maintain the confidence radii of seeds at fixed-length intervals. For instance, instead of recording the confidence radii for seeds of all lengths between 1 to P at every location x in T, the confidence radius database can selectively record the confidence radii just for seeds of lengths at every 10-bp interval: $T[x, x + 10]$, $T[x, x + 20]$, ..., $T[x, x + 10 \times \lceil \frac{P}{10} \rceil]$. We call this method *interval sampling*. Let I denote the interval length for seed sampling, interval sampling reduces the space complexity of the CAS database to $O(|T| \cdot \lceil \frac{P}{I} \rceil)$. An example confidence radius database with a length of $I = 4$ is provided in Fig. 7.

For seeds that are not sampled by the confidence radius database, their confidence radii may still be indirectly inferred during mapping. Given a seed u and its substring v, we have the following theorem:

Theorem 4 *Let seed v be a substring of seed u. Let occ(u) and occ(v) denote the frequency of u and v in T. Then* $c_u \geq c_v$, *if* $occ(u) = occ(v)$.

Proof Assume $c_u < c_v$. Then there exists a string $w \in T$ where $D(u, w) = c_u$ while w is not in the vicinity of u. Since v is a substring of u, there must exist w', a substring of w, where $D(w', v) \leq c_u$. However, since

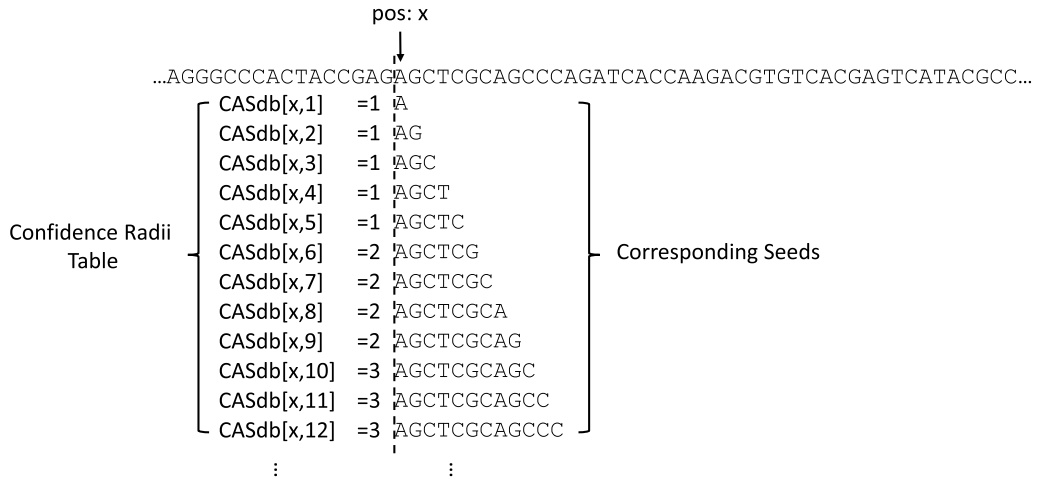

Fig. 6 An example confidence radius table

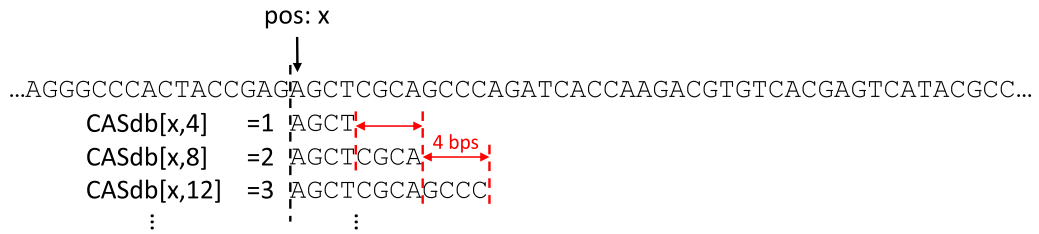

Fig. 7 An example confidence radius table with a sampling interval of $I = 4$

$c_v > c_u \geq D(w', v)$, w' must not be a non-trivial neighbor of v. Hence w' must be in the vicinity of v. Because w is not in the vicinity of u but w' is in the vicinity of v, we conclude that v occurs in at least one more location in T than u does (in the vicinity of w'). This contradicts the fact that $occ(u) = occ(v)$. □

Theorem 4 suggests that for a seed s that is not sampled by the confidence radius database, as long as there exists a confidence-radius-database-covered substring s' of s that shares the same occurrence frequency with s, then $c_s \geq c_{s'}$. Since the confidence radius is just a lower bound, and need not be an exact measurement of the minimum edit distance towards the non-trivial neighbors of a seed, we can set $c_s = c_{s'}$.

In mapping, upon encountering a seed s with a length that is not a multiple of the sampling interval I, the mapper may enumerate all substrings of s with lengths of $I \cdot \lfloor \frac{|s|}{I} \rfloor$. Unlike s, these substrings are supported by the confidence radius database. The mapper then checks the frequency of each substring in T and removes the substrings whose occurrence frequencies do not match s. The confidence radius of s is set to the maximum confidence radius among the remaining frequency-matching substrings. This method works only if there exists at least one frequency-matching substring of s. When there is none, the mapper is recommended to give up s, and directly use a database-supported substring of s instead, preferably the one with the minimum occurrence frequency or the maximum confidence radius, which maximizes the mapping efficiency.

A seeding scheme with context-aware seeds

While the major goal of this paper is to establish the theoretical framework of CAS, to test the effectiveness of CAS, we propose a greedy seed selection method, referred to as greedy CAS seeding. Greedy CAS seeding selects consecutive Maximum Exact Matching substrings (MEMs, which are seeds that cannot be further extended without bumping into errors) from a read as seeds. At the end of each MEM, greedy CAS seeding heuristically skips the next two base pairs, in an effort to skip potential errors. Greedy CAS seeding sorts seeds by their frequency from low to high, into S_{raw}. Then selects the minimum number of seeds S from S_{raw} in sequential order such that $\sum_{s \in S} c_s \geq t$. In the rare cases where there is insufficient number of CAS seeds such that there does not exist a set of seeds S with $\sum_{s \in S} c_s \geq t$, greedy CAS

seeding reverts back to using the pigeonhole principle, by dividing the read into t non-overlapping seeds.

Figure 8 compares the seed extraction results of greedy CAS seeding against the state-of-the-art, pigeonhole-principle-based seeding method, the Optimal Seed Solver (OSS) [8]. OSS has been previously shown to generate the least frequent seeds, when compared to other pigeonhole-principle-based seeding methods, such as flexible-placement k-mers or spaced seeds. Figure 8 demonstrates both seeding methods in action under $t = 4$. Greedy CAS seeding is shown in the upper half while OSS is shown in the lower half. Compared to OSS, which uses a total of $t = 4$ seeds, greedy CAS seeding uses only two seeds. As a result, greedy CAS seeding can afford longer and less frequent seeds.

When interval sampling is enabled, greedy CAS seeding follows the seeding mechanism introduced above. If a seed s has a length that is not supported by the confidence radius database and it has no frequency-matching substrings, then s is withdrawn and the database-supported substring of s with the maximum confidence radius is selected as a substitute.

Greedy CAS seeding has a maximum complexity of $O(|R| + |S| \log(|S|))$. We use a Burrows-Wheeler Transformation (BWT) array to index seeds. With the BWT array, it takes $O(|s|)$ operations to access the seed database for seed s and locate all seed locations of s. Given that $\sum_{s \in S} |s| \leq |R|$, and $|S| \leq t \ll |R|$, we conclude that the maximum complexity of greedy CAS seeding is $O(|R| + t \log(t))$.

Experiments

We benchmark greedy CAS seeding against OSS and naïve seeding on the *E. coli* genome. Naïve seeding selects consecutive 12-bp seeds following the Pigeonhole principle. We benchmark both seeding schemes on a 22-million, 100-bp *E. coli* read set from EMBL-EBI, ERX008638-1. We build a confidence radius database for *E. coli* genome with a maximum confidence radius threshold $\theta = 5$ and a max seed length $P = 60$. We measure the effectiveness of both approaches by comparing the average total seed frequency of selected seeds under different edit distance thresholds $t = \{1, 2, 3, 4, 5\}$. The average total frequency is the sum of seed frequencies extracted from each read, averaged over all reads in the read set.

Figure 9 shows the average total seed frequency comparison between the two approaches. OSS has slightly smaller total seed frequency (averaged over all reads) under $t = 1$, but it quickly increases, exceeding CAS at $t > 1$. OSS outperforms CAS under $t = 1$ because greedy CAS seeding extracts seeds sequentially; while OSS scans through all possible MEM placements in a read and picks the least frequent placement. CAS achieves significantly lower seed frequencies with $t > 1$. When t gets larger, OSS is pressured to use more seeds, which leads to using shorter and more frequent seeds. To the contrary, greedy

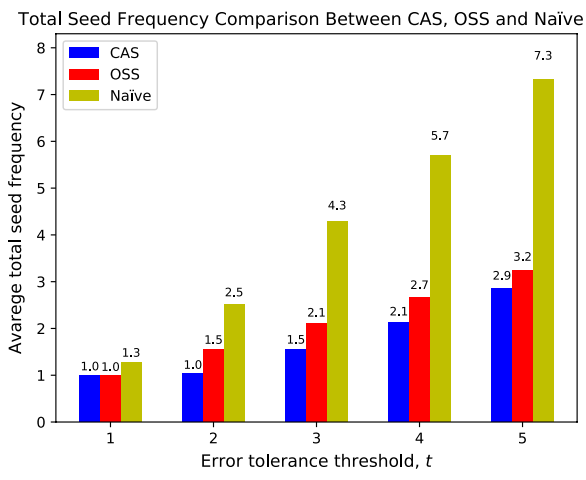

Fig. 9 Comparison between CAS and OSS in terms of total seed frequency, with variate edit distance thresholds t

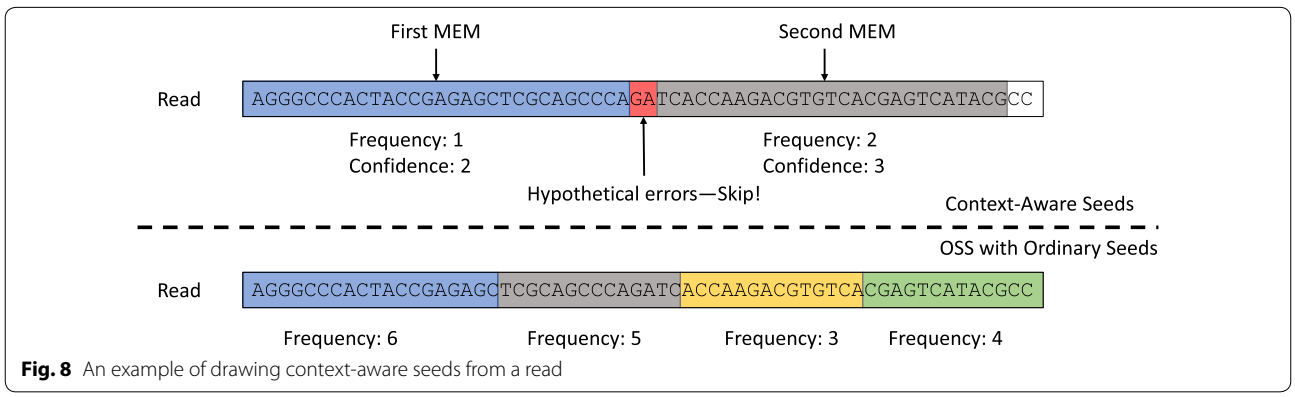

Fig. 8 An example of drawing context-aware seeds from a read

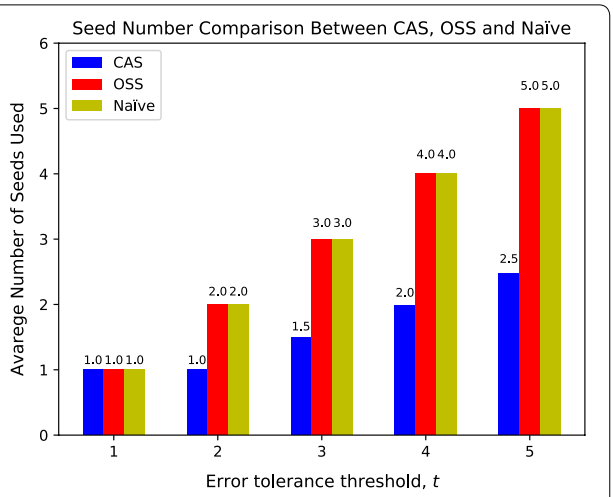

Fig. 10 Comparison between CAS and OSS in terms of average number of seeds used, with variate edit distance thresholds t

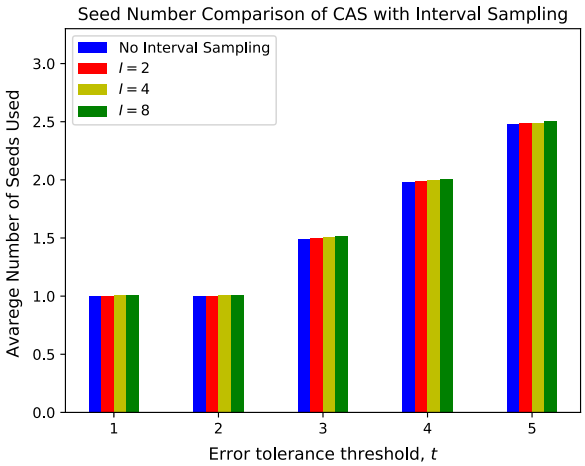

Fig. 12 Comparison of CAS with variate sampling intervals, in terms of average number of seeds used, with variate edit distance thresholds t

CAS seeding often uses fewer than t seeds, as shown in Fig. 10, which lets it use longer and less frequent seeds.

We also benchmarked the performance of CAS with variate sampling intervals. The results are summarized in Figs. 11, 12. As shown in both figures, while the average seed frequency increases as the interval increases, the magnitude of the increase is very small. Compared to CAS without interval sampling, CAS with a sampling interval of $I = 8$ only increases the average seed frequency by 6%, while reducing the storage footprint by 8×. Overall, interval sampling significantly reduces the space complexity of the CAS database, with minor performance loss.

CAS is expected to perform better on larger genomes. The *E. coli* genome is a small genome, which has only around 4.6 million base pairs. In comparison, the human genome has more than 3 billion base pairs. For small genomes, seeds become less frequent by nature. Therefore short seeds become acceptable as they are not as frequent as they are in larger genomes. We therefore expect CAS to perform better in larger genomes. However, due to practical (not theoretical) limitations in scaling up the construction of the confidence radius database on larger genomes (further elaborated in the Discussion section), we only demonstrate CAS on the *E. coli* genome.

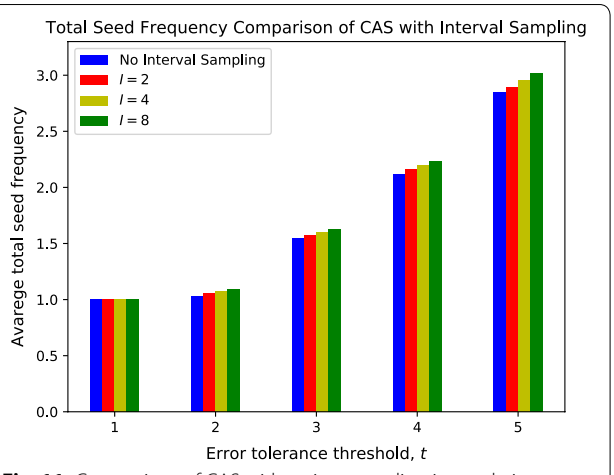

Fig. 11 Comparison of CAS with variate sampling intervals, in terms of total seed frequency, with variate edit distance thresholds t

While the focus of this paper is to establish the theoretical foundation of CAS, instead of providing a complete read mapping solution, it is worth mentioning that greedy CAS seeding (only the seeding mechanism) is more practical than OSS. OSS requires scanning through all substrings of R, which has a total size of $O(|R|^2)$, for seed frequencies. Combined with BWT, it takes at least $O(|R|^2)$ operations to collect all seed frequencies with OSS. Greedy CAS seeding, to the contrary, finishes in $O(|R| + t\log(t))$ time with $t \ll |R|$.

Discussion

Although Algorithm 1 finishes in $O(|\Sigma|^2 \cdot M)$ time, in practice, M could be on the scale of trillions or more, for large and complex genomes. This is because for large genomes, the suffix trie is close to full in the first ten to twenty levels, where almost every permutation of letters exists. Nodes in these levels have large numbers of neighbors: the number of neighbors of a node v, equals to the number of unique strings formed by editing the string of v with up to t edits. After each edit, the resulting string is guaranteed to appear in *Trie*. Figure 13 demonstrates the average number of neighbors (both trivial

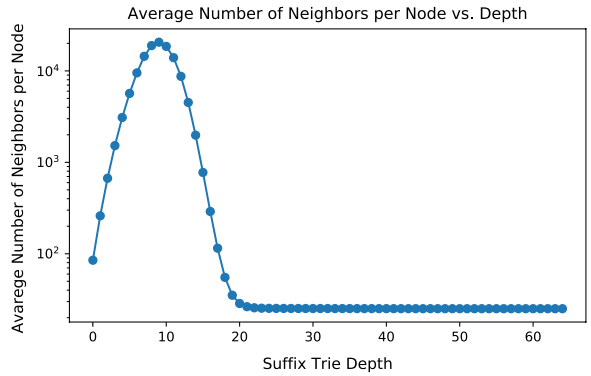

Fig. 13 Average number of neighbors of each node at variate depths in the *E. coli* suffix trie

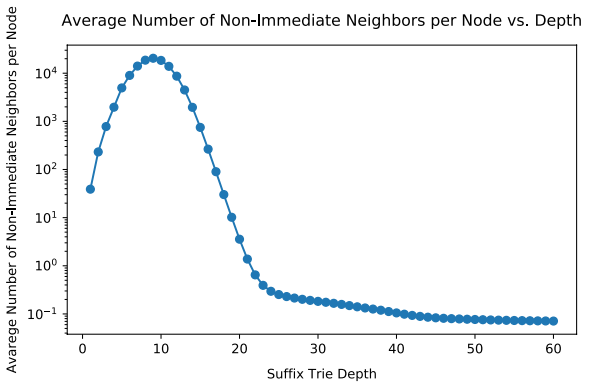

Fig. 15 Average number of non-immediate neighbors of each node at variate depths in the *E. coli* suffix trie

and non-trivial) per seed at different depths in a *Trie*. The *Trie* in Fig. 13 is built with *E. coli* genome with $\theta = 4$. As the depth grows, the average number of neighbors per node first increases, then peaks at 9 and then quickly decreases. Such a trajectory is closely related to the trend of the average number of children per node in *Trie*. Figure 14 presents the average number of children per node at variate depths in the *E. coli Trie*. Between depths 1-9, the *E. coli Trie* is close to a full 4-nary tree, with most nodes having 4 children. Between depths 10-14, the average number of children per node quickly decreases and stabilizes at slightly above 1. As a result, *Trie* becomes increasingly sparse after depth 15. Between depths 1-9, as *Trie* is dense, the number of neighbors of a seed approaches to the number of unique permutations of the seed, with up to t edits. As *Trie* quickly becomes sparse after depth 9, the number of neighbors of a seed quickly diminishes. Finally, the average count of neighbors stabilizes at around 25, which is roughly the average number of immediate neighbors of a seed. As the depth increases, fewer nodes have non-immediate neighbors. Figure 15 illustrates the average number of non-immediate neighbors of seeds at different depths in *Trie*. As shown in the figure, the average number of non-immediate neighbors per node decreases to slightly above 0 after depth 20. Hence, the majority of seeds in *E. coli* with lengths greater than 20 have no non-immediate neighbors and, therefore, have confidence radii of 4.

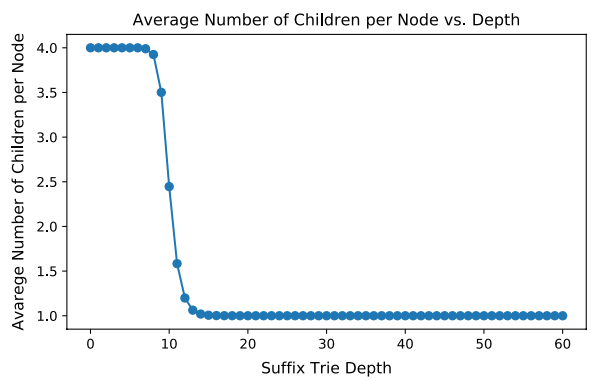

Fig. 14 Average number of children of each node at variate depths in the *E. coli* suffix trie

The suffix trie maintains a similar structure in complex genomes. At top, the trie is very dense, close to a full 4-nary tree and the suffix trie expands exponentially. After a certain depth (9 for *E. coli*), the suffix trie quickly becomes sparse, with the average number of children per node stabilizing at a little bit over 1. Similarly, the average number of neighbors first expands super-linearly. Then it peaks at around the depth where the suffix trie turns sparse and quickly decreases to just the average number of trivial neighbors per node.

However, unlike *E.coli*, the suffix tries in complex genomes have much greater scales. In human genomes, there are more than one billion unique 15-base-pair suffixes, compared to around one million in *E. coli*. This means that for human genomes, with $\theta = 4$, there could be more than 1 trillion total neighbors just for 15-base-pair suffixes. Maintaining metadata at such scale vastly exceeds the capacity of our currently available computational power. From our experiment, it takes around 300 CPU hours to compute the confidence radius database for the *E. coli* genome with $\theta = 4$ and $P = 60$ on a multi-CPU, mechanical hard drive system. However, it is worth noting that as a theoretical study, the database construction program is not fully optimized for speed and is currently I/O-bound due to frequently reading and writing neighbor information into neighbor arrays of nodes in *Trie*.

While there are many nodes (long suffixes) with fewer neighbors, given that Algorithm 1 traverses *Trie* in a top-down manner, it is unavoidable to track the massive number of neighbors for short suffixes. This is an interesting algorithmic problem for future work.

CAS may be applied to situations other than NGS read mapping. For example, the idea of context-aware seeds may improve long-read mapping. Long reads suffer from high error rates [9–11]. Finding error-free seeds for long reads is very challenging [12]. CAS can serve as a metric measuring the likelihood of seeds having errors: if there exists a seed, s, with high confidence radius, it is highly likely that s is free of errors. The likelihood of obtaining a reference-matching seed through many accidental errors is small.

Finally, CAS can be applied to develop probes for DNA and RNA identification. When designing probe sequences, it is important to make certain that the target sequence is unique in the genome [13–15]. It prevents probes from accidentally annealing to a similar sequences. CAS checks the existence of similar sequences by consulting the confidence radius database.

Conclusion

In this work, we proposed a new seeding framework, context-aware seeds (CAS). CAS extends the pigeonhole principle and guarantees finding all valid mappings with fewer seeds. CAS associates each seed s with a confidence radius c_s, defined as a lower bound of edit distances towards nontrivial neighbors of s. We proved that the CAS can find all valid mappings of any read R, as long as its seeds s satisfy $\sum c_s \geq t$.

We proposed a linear-time algorithm for constructing the confidence radius database. It computes the confidence radii of seeds by traversing the suffix trie of a reference. We experimented with CAS on the *E. coli* genome and compared it against the state-of-the-art pigeonhole-principle-based seeding scheme, OSS, and showed that CAS outperforms OSS by significantly reducing the sum of seed frequencies.

This paper focuses on the theoretical aspects of CAS, especially how it extends the pigeonhole principle into using fewer seeds. Composing a practical solution of Algorithm 1 on larger genomes is an interesting-yet-separate problem for future work.

Acknowledgements

This research is funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through grant GBMF4554 to CK, by the U.S. National Science Foundation (CCF-1319998) and by the U.S. National Institutes of Health (R01GM122935). This work was partially funded by the Shurl and Kay Curci Foundation. This project is funded, in part, by a grant (4100070287) from the Pennsylvania Department of Health. The department specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Authors' contributions

HX developed the theoretical formalism. HX and MS performed the analytic calculations. HX engineered the program and performed the experiments. Both HX, MS and CK contributed to the writting of the manuscript. CK supervised the project. All authors read and approved the final manuscript.

Competing interests

C.K. is co-founder of Ocean Genomics, Inc.

Author details

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh 15213, USA. [2] Department of Computer Science and Engineering, Pennsylvania State University, State College 16801, USA. [3] Computational Biology Department, Carnegie Mellon University, Pittsburgh 15213, USA. [4] Present Address: UM-SJTU Joint Institute, Shanghai Jiaotong University, Shanghai 200240, China.

Received: 13 December 2019 Accepted: 15 May 2020
Published online: 23 May 2020

References

1. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357.
2. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:1303.3997.
3. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
4. Xin H, Lee D, Hormozdiari F, Yedkar S, Mutlu O, Alkan C. Accelerating read mapping with FastHASH, vol. 14. London: BioMed Central; 2013. p. 13.
5. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–93.
6. Tran NH, Chen X. AMAS: optimizing the partition and filtration of adaptive seeds to speed up read mapping. IEEE/ACM Transact Comput Biol Bioinf. 2016;13(4):623–33.
7. Landau GM, Vishkin U. Fast parallel and serial approximate string matching. J Algorith. 1989;10(2):157–69.
8. Xin H, Nahar S, Zhu R, Emmons J, Pekhimenko G, Kingsford C, Alkan C, Mutlu O. Optimal seed solver: optimizing seed selection in read mapping. Bioinformatics. 2015;32(11):1632–42.
9. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. F1000 Research. 2017;6:100.
10. Haghshenas E, Hach F, Sahinalp SC, Chauve C. Colormap: correcting long reads by mapping short reads. Bioinformatics. 2016;32(17):545–51.
11. Haghshenas E, Sahinalp SC, Hach F. lordFAST: sensitive and fast alignment search tool for long noisy read sequencing data. Bioinformatics. 2018;35(1):20–7.
12. Jain C, Dilthey A, Koren S, Aluru S, Phillippy AM. A fast approximate algorithm for mapping long reads to large reference databases. In: International Conference on Research in Computational Molecular Biology, pp. 66–81; 2017. Springer.
13. Wang JS, Zhang DY. Simulation-guided DNA probe design for consistently ultraspecific hybridization. Nat Chem. 2015;7(7):545.
14. Dugat-Bony E, Peyretaillade E, Parisot N, Biderre-Petit C, Jaziri F, Hill D, Rimour S, Peyret P. Detecting unknown sequences with DNA microarrays: explorative probe design strategies. Environ Microbiol. 2012;14(2):356–71.
15. Li Q, Luan G, Guo Q, Liang J. A new class of homogeneous nucleic acid probes based on specific displacement hybridization. Nucleic Acids Res. 2002;30(2):5.

Publisher's Note