



# Quantum Meets Fine-Grained Complexity: Sublinear Time Quantum Algorithms for String Problems

François Le Gall<sup>1</sup> · Saeed Seddighin<sup>2</sup>

Received: 4 August 2022 / Accepted: 15 November 2022 / Published online: 5 December 2022  
© The Author(s) 2022

## Abstract

Longest common substring (LCS), longest palindrome substring (LPS), and Ulam distance (UL) are three fundamental string problems that can be classically solved in near linear time. In this work, we present sublinear time quantum algorithms for these problems along with quantum lower bounds. Our results shed light on a very surprising fact: Although the classic solutions for LCS and LPS are almost identical (via suffix trees), their quantum computational complexities are different. While we give an exact  $\tilde{O}(\sqrt{n})$  time algorithm for LPS, we prove that LCS needs at least time  $\tilde{\Omega}(n^{2/3})$  even for 0/1 strings.

**Keywords** String algorithms · Quantum algorithms · Sublinear-time algorithms · Fine-grained complexity

## 1 Introduction

Perhaps the earliest questions that were studied in computer science are the algorithmic aspects of string problems. *The edit distance*, *longest common substring*, and *longest palindrome substring* are some of the more famous problems in this category. Efforts to solve these problems led to the discovery of several fundamental techniques such as *dynamic programming*, *hashing algorithms*, and *suffix trees*. These algorithms have numerous applications in several fields including DNA-sequencing, social media, compiler design, anti-virus softwares, etc.

---

Throughout the paper, when we say an exact solution, we mean a solution that does not lose anything in the approximation. However, it may be possible that our algorithm succeeds with probability less than 1 in which case we explicitly mention that.

---

✉ François Le Gall  
legall@math.nagoya-u.ac.jp

<sup>1</sup> Graduate School of Mathematics, Nagoya University, Furocho, Chikusaku, Nagoya, Aichi 464-8602, Japan

<sup>2</sup> Toyota Technological Institute at Chicago, 6045 S Kenwood Avenue, Chicago, IL 60637, USA

All of the above problems have received significant attention in the classical setting (see, e.g., [1–9]). Efficient classical algorithms for these problems emerged early on in the 1960s [10]. Moreover, thanks to a series of recent developments in fine-grained complexity [11, 12], we now seem to have a clear understanding of the classical lower bounds as well. Unless plausible conjectures such as  $\text{SETH}^1$  are broken, we do not hope for a substantially better algorithm for edit distance. Failure to extend the fine-grained lower bounds to the quantum setting has left some very interesting open questions behind both in terms of quantum complexity and quantum lower bounds.

Despite a plethora of new quantum algorithms for various problems (e.g., [13–18]), not much attention is given to string problems. Until recently, the only non-trivial quantum algorithm for such problems was the  $\tilde{O}(\sqrt{n} + \sqrt{m})$  time algorithm of Ramesh and Vinay [19] for pattern matching where  $n$  and  $m$  are the sizes of the text and the pattern (we also mention the works [20, 21] that consider string problems with nonstandard queries in the quantum setting). Recently, Boroujeni, Ehsani, Ghodsi, Hajiaghayi, and Seddighin [5] made a clever use of the Grover’s search algorithm [22] to obtain a constant factor approximation quantum algorithm for edit distance in truly subquadratic time. Shortly after, it was shown by Chakraborty, Das, Goldenberg, Koucky, and Saks [6] that a similar technique can be used to obtain a classical solution with the same approximation factor. Several improved classical algorithms have been given for edit distance in recent years [23–25] though it is still an open question if a non-trivial quantum algorithm can go beyond what we can do classically for edit distance.

In this work, we give novel sublinear time quantum algorithms and quantum lower bounds for LCS, LPS, and a special case of edit distance namely *Ulam distance* (UL). All these problems require  $\tilde{\Omega}(n)$  time in the classical setting even if approximate solutions are desired. LCS and LPS can be solved in linear time via suffix trees [10] and there is an  $O(n \log n)$  time algorithm for Ulam distance [10]. Our results shed light on a very surprising fact: Although the classical solutions for LCS and LPS are almost identical, their quantum computational complexities are different. While we give an exact  $\tilde{O}(\sqrt{n})$  time quantum algorithm for LPS, we prove that any quantum algorithm for LCS needs at least time  $\tilde{\Omega}(n^{2/3})$  even for 0/1 strings. We accompany this with several sublinear time quantum algorithm for LCS. A summary of our results is given in Tables 1 and 2.

## 1.1 Related Work

Our work is very similar in spirit to for instance the work of Ambainis, Balodis, Iraids, Kokainis, Prusis, and Vihrovs [26], where Grover’s algorithm is cleverly combined with classical techniques to design more efficient quantum algorithms for dynamic programming problems. This is particularly similar to our approach since we obtain our main results by combining known quantum algorithms with new classical ideas. In the present work, however, we go beyond Grover’s algorithm and make use of several other quantum techniques such as element distinctness, pattern matching, amplitude

<sup>1</sup> The *strong exponential time hypothesis* states that no algorithm can solve the satisfiability problem in time  $2^{n(1-\epsilon)}$ .

**Table 1** Our algorithms are shown in this table

Problem	Solution	Runtime	Reference
Longest common substring (LCS)	Exact	$\tilde{O}(n^{5/6})$	Theorem 1
Longest palindrome substring (LPS)	Exact	$\tilde{O}(\sqrt{n})$	Theorem 2
Longest common substring (LCS)	$1 - \epsilon$ approximation	$\tilde{O}(n^{3/4})$	Theorem 3
LCS for non-repetitive strings	exact	$\tilde{O}(n^{3/4})$	Theorem 4
LCS for non-repetitive strings	$1 - \epsilon$ approximation	$\tilde{O}(n^{2/3})$	Theorem 5
Ulam distance	$1 + \epsilon$ approximation	$\tilde{O}(\sqrt{n})$	Theorem 6

**Table 2** This table includes the quantum lower bounds for the problems we study in this work

Problem	Approximation factor	Lower bound	Reference
LCS	$1 - \epsilon$ approximation	$\tilde{\Omega}(n^{2/3})$	Theorems 8 and 9
LPS	$1 - \epsilon$ approximation	$\tilde{\Omega}(\sqrt{n})$	Theorem 10
Ulam distance	$1 + \epsilon$ approximation	$\tilde{\Omega}(\sqrt{n})$	Theorem 11

amplification, and amplitude estimation to obtain our improvements. In addition to this, we also develop quantum walks that improve our more general results for some special cases. In particular, our quantum walk for obtaining a  $1 - \epsilon$  approximate solution for LCS is tight up to logarithmic factors due to a lower-bound we give in Sect. 7.

Another line of research which is closely related to our work is the study of quantum lower bounds for edit distance [27, 28]. While a SETH-based quadratic lower bound is known for the classical computation of edit distance, quantum lower bounds are not very strong. Recently, a quantum lower bound of  $\Omega(n^{1.5})$  was given by Buhrman, Patro, and Speelman [28] under a mild assumption. Still no quantum algorithm better than the state of art classical solution (which runs in time  $O(n^2 / \log^2 n)$  [1]) is known for edit distance. The reader can find more details in [27].

While not directly related to string algorithms, another investigation of SETH in the quantum setting is the recent work by Aaronson, Chia, Lin, Wang, and Zhang [29] that focuses on the quantum complexity of the closest pair problem, a fundamental problem in computational geometry. Interestingly, the upper bounds obtained in that paper also use an approach based on element distinctness and quantum walks. Despite this, the high-level ideas of our work are substantially different from [29] as we utilize several novel properties of LCS and LPS to design our algorithms.

In the classical sequential setting, LCS and LPS can be solved in linear time [10]. The solutions are almost identical: we first construct suffix trees for the input strings and then find *the lowest common ancestors* for the tree nodes. Ulam distance can also be solved exactly in time  $O(n \log n)$  [10] which is the best we can hope for via a comparison-based algorithm [30] or algebraic decision trees [31]. Approximation algorithms running in time  $\tilde{O}(n/d + \sqrt{d})$  where  $d$  denotes the Ulam distance of the two strings have also been developed [32, 33]. Finally, we notice that after the

present results have been made public, Akmal and Jin [34] have constructed an optimal  $\tilde{O}(n^{2/3})$ -time quantum algorithm for LCS.

## 1.2 Preliminaries

### *Description of LCS, LPS and UD*

In the longest common substring problem (LCS), the input consists of two strings and our goal is to find the longest substring<sup>2</sup> which is shared between the two strings. We denote the two input strings by  $A$  and  $B$ . We assume that  $A$  and  $B$  have the same length, which we denote by  $n$ . We use  $\Sigma$  to denote the alphabet of the strings. For any  $\epsilon \in [0, 1)$ , we say that an algorithm outputs a  $(1 - \epsilon)$ -approximation of the longest common substring if for any input strings  $A$  and  $B$ , it outputs a common substring of length at least  $(1 - \epsilon)d$ , where  $d$  is the length of the longest common substring of  $A$  and  $B$ .

In the longest palindrome substring problem (LPS), the goal is to find the longest substring of a given string  $A$  which reads the same both forward and backward. The length of  $A$  is also denoted by  $n$  and its alphabet by  $\Sigma$ . For any  $\epsilon \in [0, 1)$ , we say that an algorithm outputs a  $(1 - \epsilon)$ -approximation of the longest palindrome substring if for any input string  $A$ , it outputs a palindrome substring of length at least  $(1 - \epsilon)d$ , where  $d$  is the length of the longest palindrome substring of  $A$ .

We say that a string of length  $n$  over an alphabet  $\Sigma$  is *non-repetitive* if no character appears twice in the string (note that this can happen only if  $|\Sigma| \geq n$ ). The Ulam distance is a special case of the edit distance in which the input strings are non-repetitive. Let us now define the problem more formally. In Ulam Distance (UD) we are given two non-repetitive strings  $A$  and  $B$  of length  $n$ , and consider how to transform one of them to the other one. For this purpose we allow two basic operations *character addition* and *character deletion*, each at a unit cost and our goal is to minimize the total cost of the transformation<sup>3</sup>. We denote by  $\text{ud}(A, B)$  the minimum number of such operations needed to transform  $A$  into  $B$ . The goal is to compute  $\text{ud}(A, B)$ , either exactly or approximately. For any  $\epsilon \in [0, 1]$ , we say that an algorithm outputs a  $(1 + \epsilon)$ -approximation of  $\text{ud}(A, B)$  if it outputs some value  $r$  such that the inequality  $(1 - \epsilon) \cdot \text{ud}(A, B) \leq r \leq (1 + \epsilon) \cdot \text{ud}(A, B)$  holds.

### *General definitions and conventions*

Throughout the paper, we use notations  $\tilde{O}(\cdot)$  and  $\tilde{\Omega}(\cdot)$  that hide the polylogarithmic factors in terms of  $n$ . We always assume that the size of  $\Sigma$  is polynomial in  $n$  and that each character is encoded using  $O(\log n)$  bits. The size of  $\Sigma$  thus never appears explicitly in the complexity of our algorithms. We say that a randomized or a quantum algorithm solves a problem like LCS, LPS or UL with high probability if it solves the problem with probability at least  $9/10$  (this success probability can be easily amplified to  $1 - 1/\text{poly}(n)$  with a logarithmic overhead in the complexity).

<sup>2</sup> In a substring, the characters are next to each other. In other words, the positions of the characters of a substring should make an interval. This is in contrast to subsequence where the positions can be arbitrary.

<sup>3</sup> Another popular version allows character substitution as a third operation. These two definitions of the Ulam distance only differ by a factor at most 2.

For convenience, we often only compute/approximate the size of the solution as opposed to explicitly giving the solution. However, it is not hard to see that for LCS and LPS, the same algorithms can also give an explicit solution with a logarithmic overhead in the runtime. (a solution can be specified by two integers pointing at the interval of the input.)

For a string  $X$ , we denote by  $X[i, j]$  the substring of  $X$  that starts from the  $i$ -th character and ends at the  $j$ -th character. We say a string  $X$  is  $q$ -periodic if we have  $X_i = X_{i+q}$  for all  $1 \leq i \leq |X| - q$ . Moreover, the periodicity of a string  $X$  is equal to the smallest number  $q > 0$  such that  $X$  is  $q$ -periodic. We also call a non-repetitive string of length  $n$  over an alphabet of size  $n$  a *permutation* (it represents a permutation of the set  $\Sigma$ ).

**Quantum access to the inputs**

In the quantum setting, we suppose that the input strings  $A$  and  $B$  can be accessed directly by a quantum algorithm. More precisely, we have an oracle  $O_A$  that, for any  $i \in \{1, \dots, n\}$ , any  $a \in \Sigma$ , and any  $z \in \{0, 1\}^*$ , performs the unitary mapping  $O_A : |i\rangle|a\rangle|z\rangle \mapsto |i\rangle|a \oplus A[i]\rangle|z\rangle$ , where  $\oplus$  denotes an appropriate binary operation defined on  $\Sigma$  (e.g., bit-wise parity on the binary encodings of  $a$  and  $A[i]$ ). Similarly we have an oracle  $O_B$  that, for any  $i \in \{1, \dots, n\}$ , any  $b \in \Sigma$ , and any  $z \in \{0, 1\}^*$ , performs the unitary mapping  $O_B : |i\rangle|b\rangle|z\rangle \mapsto |i\rangle|b \oplus B[i]\rangle|z\rangle$ . Each call to  $O_A$  or  $O_B$  can be implemented at unit cost. This description corresponds to quantum random access (“QRAM access”) to the input, which is the standard model to investigate the complexity of sublinear time quantum algorithms.

**2 Results**

We present sublinear time quantum algorithms along with quantum lower bounds for LCS, LPS, and UL. For the most part, the novelty of our work is to make use of existing quantum algorithms to solve our problems. For this purpose, we introduce new classical techniques that significantly differ from the conventional methods. However for a special case of LCS, we design a novel quantum walk that leads to an improvement over our more general solution. We give a brief explanation of this technique later in the section. For now, we start by stating the quantum tools that we use in our algorithms.

**2.1 Quantum Components**

**Grover’s search** ([15, 35]). Given a function  $f : [n] \rightarrow \{0, 1\}$ , Grover’s algorithm can find an element  $x \in [n]$  such that  $f(x) = 1$  or verify if  $f(i) = 0$  for all  $i \in [n]$ . This quantum algorithm runs in time  $\tilde{O}(\sqrt{n} \cdot T(n))$  and succeeds with probability  $9/10$  (the success probability can be increased to  $1 - 1/\text{poly}(n)$  with only a logarithmic overhead). Here,  $T(n)$  represents the time complexity of computing  $f(i)$  for one given element  $i \in [n]$ . Additionally, distinguishing between the case where  $f(x) = 1$  holds for at least  $m$  elements (for some value  $1 \leq m \leq n$ ) and the case where  $f(i) = 0$  for all  $i \in [n]$  can be done in time  $\tilde{O}(\sqrt{n/m} \cdot T(n))$ .

**Pattern matching** ([19]). Let  $P$  and  $S$  be a pattern and a text of lengths  $n$  and  $m$  respectively. One can either verify that  $P$  does not appear as a substring in  $S$  or find the leftmost (rightmost) occurrence of  $P$  in  $S$  in time  $\tilde{O}(\sqrt{n} + \sqrt{m})$  via a quantum algorithm. The algorithm gives a correct solution with probability at least  $9/10$ .

**Element distinctness** ([13]). Let  $X$  and  $Y$  be two lists of size  $n$  and  $f: (X \cup Y) \rightarrow \mathbb{N}$  be a function that is used to compare the elements of  $X$  and  $Y$ <sup>4</sup>.

There is a quantum algorithm that finds (if any) an  $(x, y)$  pair such that  $x \in X$ ,  $y \in Y$  and  $f(x) = f(y)$ . The algorithm succeeds with probability at least  $9/10$  and has running time  $\tilde{O}(n^{2/3} \cdot T(n))$ , where  $T(n)$  represents the time needed to answer to the following question: Given  $\alpha, \beta \in X \cup Y$  is  $f(\alpha) = f(\beta)$  and if not which one is smaller?

**Amplitude amplification** ([35]). Let  $Q$  be a decision problem and  $\mathcal{A}$  be a quantum algorithm that solves  $Q$  with one-sided error and success probability  $0 < p < 1$  (i.e., on a yes-instance  $\mathcal{A}$  always accepts, while on a no-instance  $\mathcal{A}$  rejects with probability  $p$ ). Let  $T$  be the runtime of  $\mathcal{A}$ . One can design a quantum algorithm for  $Q$  with runtime  $O(T/\sqrt{p})$  that solves  $Q$  with one-sided error and success probability at least  $9/10$ .

**Amplitude estimation** ([35]). Let  $\mathcal{A}$  be a quantum algorithm that outputs 1 with probability  $0 < p < 1$  and returns 0 with probability  $1 - p$ . Let  $T$  be the time needed for  $\mathcal{A}$  to generate its output. For any  $\alpha > 0$ , one can design a quantum algorithm with runtime  $O(T/(\alpha\sqrt{p}))$  that outputs with probability at least  $9/10$  an estimate  $\tilde{p}$  such that  $(1 - \alpha)p \leq \tilde{p} \leq (1 + \alpha)p$ .

## 2.2 LCS and LPS

In this section, we outline the ideas for obtaining sublinear time algorithms for LCS and LPS. We begin as a warm up by giving a simple exact algorithm for LCS that runs in sublinear time when the solution size is small. Next, we explain our techniques for the cases that the solution size is large (at a high-level, this part of the algorithm is very similar in both LCS and LPS). In our algorithms, we do a binary search on the size of the solution. We denote this value by  $d$ . Thus, by losing an  $O(\log n)$  factor in the runtime, we reduce the problem to verifying if a solution of size at least  $d$  exists for our problem instance.

**Exact quantum algorithm for LCS (small  $d$ ).** For small  $d$ , we use element distinctness to solve LCS. Let  $|\Sigma|$  be the size of the alphabet and  $v: \Sigma \rightarrow [0, |\Sigma| - 1]$  be a function that maps every element of the alphabet to a distinct number in range  $0 \dots |\Sigma| - 1$ . In other words,  $v$  is a perfect hashing for the characters. We extend this definition to substrings of the two strings so that two substrings  $t$  and  $t'$  are equal if and only if we have  $v(t) = v(t')$ . From the two strings, we then make two sets of numbers  $S_A$  and  $S_B$  each having  $n - d + 1$  elements. Element  $i$  of set  $S_A$  is a pair  $(A, i)$  whose value is equal to  $v(A[i, i + d - 1])$  and similarly element  $i$  of  $S_B$  is a pair  $(B, i)$  whose

<sup>4</sup> In the standard definition of element distinctness, we are given a single list of elements and the goal is to find out if two elements in the list are equal. The present definition, also known as claw finding, is slightly more general—for completeness we discuss how the upper bound  $\tilde{O}(n^{2/3} \cdot T(n))$  is obtained for our version as well in Sect. 4.1.

value is equal to  $v(B[i, i + d - 1])$ . The two sets contain elements with equal values if and only if the two strings have a common substring of size  $d$ . Therefore, by solving element distinctness for  $S_A$  and  $S_B$  we can find out if the size of the solution is at least  $d$ . Although element distinctness can be solved in time  $\tilde{O}(n^{2/3})$  when element comparison can be implemented in  $\tilde{O}(1)$  time, our algorithm needs more runtime since comparing elements takes time  $\omega(1)$ . More precisely, each comparison can be implemented in time  $\tilde{O}(\sqrt{d})$  in the following way: In order to compare two elements of the two sets, we can use Grover’s search to find out if the two substrings are different in any position and if so we can find the leftmost position in time  $\tilde{O}(\sqrt{d})$ . Thus, it takes time  $\tilde{O}(\sqrt{d})$  to compare two elements, which results in runtime  $\tilde{O}(n^{2/3} \sqrt{d})$ .

**1 -  $\epsilon$  approximation for LCS and LPS (large  $d$ )** We use another technique to solve LCS and LPS when the solution is large. While for LPS, this new idea alone gives an optimal solution, for LCS we need to combine it with the previous algorithm to make sure the runtime is sublinear. Let us focus on LCS first. For a constant  $0 < \epsilon < 1$ , we define  $d' = (1 - \epsilon)d$  and randomly draw a substring of length  $d'$  from  $A$ . We denote this substring by  $P$ . More precisely, we sample an  $1 \leq i \leq n - d' + 1$  uniformly at random and set  $P = A[i, i + d' - 1]$ . Assuming the solution size is at least  $d$ , it follows that  $P$  is part of a solution with probability at least  $\epsilon d/n$ . Then, by searching this substring in  $B$ , we can find a solution of size  $d'$ .

We use the pattern matching quantum algorithm of Ramesh and Vinay [19] to search  $P$  in  $B$ . This takes time  $\tilde{O}(\sqrt{n})$  since  $|P| \leq |B| = n$ . Moreover, the success probability of this algorithm is  $\Omega(d/n)$  and therefore by amplitude amplification, we can improve the success rate to 9/10 by only losing a factor of  $O(\sqrt{n/d})$  in the runtime. Thus, if the solution is at least  $d$ , we can obtain a solution of size at least  $(1 - \epsilon)d$  in time  $\tilde{O}(\sqrt{n/d} \cdot \sqrt{n}) = \tilde{O}(n/\sqrt{d})$ . Notice that the runtime is sublinear when  $d$  is large.

The same technique can be used to approximate LPS. Similarly, we define  $d' = (1 - \epsilon)d$  for some constant  $0 < \epsilon < 1$  and draw a random substring of size  $d'$  from  $A$ . With the same argument, provided that the solution size is at least  $d$ , the probability that  $P$  is part of an optimal solution is at least  $\Omega(d/n)$ . We show in Sect. 5 that by searching the reverse of  $P$  in its neighbourhood we are able to find a solution of at least  $d'$ . This step of the algorithm slightly differs from LCS in that we only search the reverse of  $P$  in the area at most  $d$  away from  $P$ . Thus, both the text and the pattern are of size  $O(d)$  and therefore the search can be done in time  $\tilde{O}(\sqrt{d})$ . By utilizing amplitude amplification, we can obtain an algorithm with runtime  $\tilde{O}(\sqrt{n/d} \cdot \sqrt{d}) = \tilde{O}(\sqrt{n})$  and approximation factor  $1 - \epsilon$ .

**From 1 -  $\epsilon$  approximation to exact solution.** We further develop a clever technique to obtain an exact solution with the above ideas. We first focus on LCS to illustrate this new technique. The high-level intuition is the following: After sampling  $P$  from  $A$  and searching  $P$  in  $B$ , if the pattern appears only once (or a small number of times) in  $B$  then by extending the matching parts of  $B$  from both ends we may find a common substring of size  $d$  (see Fig. 1).

Thus intuitively, the challenging case is when there are several occurrences of  $P$  in  $B$ . The key observation is that since  $|P|$  is large and there are several places that  $P$  appears in  $B$  then they must overlap (see Fig. 2). This gives a very convenient

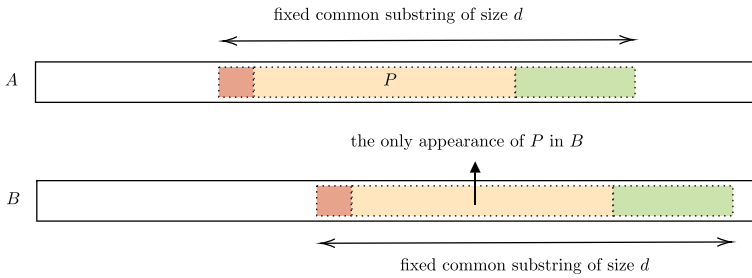


Fig. 1 When  $P$  appears in  $B$  only once

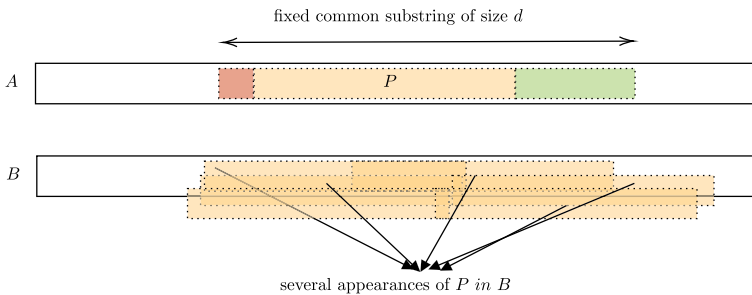


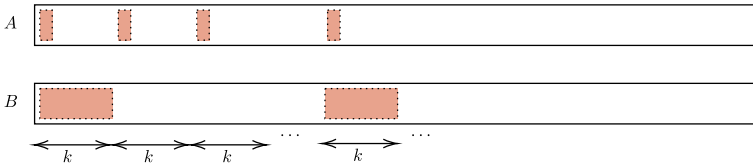
Fig. 2 When  $P$  appears several times in  $B$

approach to tackle the problem. Assume that  $P$  appears at positions  $i$  and  $j > i$  of  $B$  and these parts are overlapping. This implies that  $P$  is certainly  $(j - i)$ -periodic. To see how this enables us to solve the problem, assume that  $P$  is  $x$ -periodic and that a large continuous area of  $B$  is covered by occurrences of  $P$ . It follows that except for the boundary cases,  $P$  appears as parts of that interval that are exactly  $x$  units away. Therefore, by detecting such an interval and computing the periodicity of  $P$ , one can determine **all** occurrences of  $P$  in the interval almost as quickly as finding one occurrence of  $P$ .

Since the techniques are involved, we defer the details to Sect. 3 and here we just mention some intuition about its complexity analysis: It follows from the above observations that when several occurrences of  $P$  cover an entire interval  $I$  of  $B$ , then we only need to consider  $O(1)$  places in  $I$  that may correspond to an optimal solution. Moreover, the length of every such interval is at least  $d$ . Thus, there are only  $n/d$  places of  $B$  that we need to take into account in our algorithm. Therefore at a high-level, when  $d$  is large (say  $\Omega(n)$ ), the only non-negligible cost we pay is for the pattern matching which takes time  $\tilde{O}(\sqrt{n})$ . We show in Sect. 3 that as  $d$  becomes smaller the runtime increases; more precisely, when  $d = n/\alpha$  the runtime increases by a factor of  $O(\sqrt{\alpha})$  and thus the overall runtime of the algorithm is  $\tilde{O}(\sqrt{n} \cdot \sqrt{\alpha}) = \tilde{O}(n/\sqrt{d})$ . A very similar argument can be used to shave the  $1 - \epsilon$  approximation factor from the LPS algorithm as well.

To summarize, we obtain the following theorems. Theorem 1 combines the two algorithms we just described: if  $d$  is larger than  $n^{1/3}$  we use the second algorithm with runtime  $\tilde{O}(n/\sqrt{d})$  otherwise we use the first algorithm with runtime  $\tilde{O}(n^{2/3}\sqrt{d})$ .





**Fig. 3**  $k = \epsilon\sqrt{d}$ . Only the elements colored in red are included in the two sets (Color figure online)

**Theorem 1** *The longest common substring of two strings of size  $n$  can be computed with high probability by a quantum algorithm in time  $\tilde{O}(n^{5/6})$ .*

**Theorem 2** *The longest palindrome substring of a string of size  $n$  can be computed with high probability by a quantum algorithm in time  $\tilde{O}(\sqrt{n})$ .*

We accompany Theorems 1 and 2 with quantum lower bounds (all the lower bounds are proven in Sect. 7). Intuitively, obtaining a solution with time better than  $\tilde{O}(\sqrt{n})$  is impossible for either problem due to a reduction to searching unordered sets. This makes our solution for LPS optimal up to subpolynomial factors. For LCS, an improved lower bound of  $\tilde{\Omega}(n^{2/3})$  can be obtained via a reduction from element distinctness. However, the gap is still open between our upper bound of  $\tilde{O}(n^{5/6})$  and lower bound of  $\tilde{\Omega}(n^{2/3})$ . Thus, we aim to improve our upper bound by considering approximate solutions and special cases. In the following, we briefly explain these results.

**Improved  $1 - \epsilon$  approximation for LCS.**

One way to obtain a better algorithm for LCS is by considering  $1 - \epsilon$  approximation algorithms. Note that the quantum algorithm for element distinctness is based on quantum walks, and our solution for LCS is obtained via a reduction to element distinctness. The runtime of quantum walks can be improved when multiple solutions are present for a problem. If instead of a solution of size  $d$  which is exact, we resort to solutions of size  $(1 - \epsilon)d$ , we are guaranteed to have at least  $\epsilon d$  solutions. Thus, intuitively this should help us improve the runtime of our algorithm.

Although the intuition comes from the inner workings of the quantum techniques, we are actually able to improve the runtime with a completely combinatorial idea. For small  $d$ , instead of constructing two sets  $S_A$  and  $S_B$  with  $n - d + 1$  elements, we construct two sets of size  $O(n/\sqrt{d})$  and prove that if the two sets have an element in common, then there is a solution of size at least  $(1 - \epsilon)d$ . The construction of the two sets is exactly the same as the construction of  $S_A$  and  $S_B$  except that only some of the elements are present in the new subsets (see Fig. 3 for an illustration of the construction).

We prove that since each set now has size  $O(n/\sqrt{d})$ , if the two strings have an LCS of size  $d$  then a solution of size  $(1 - \epsilon)d$  can be found in time  $\tilde{O}((n/\sqrt{d})^{2/3}\sqrt{d}) = \tilde{O}(n^{2/3}d^{1/6})$ . This combined with the  $\tilde{O}(n/\sqrt{d})$  algorithm for large  $d$  gives us a  $1 - \epsilon$  approximation algorithm with runtime  $\tilde{O}(n^{3/4})$ .

**Theorem 3** *For any constant  $0 < \epsilon < 1$ , the longest common substring of two strings of size  $n$  can be approximated within a factor  $1 - \epsilon$  with high probability by a quantum algorithm in time  $\tilde{O}(n^{3/4})$ .*

### LCS for non-repetitive strings.

For all string problems, a special case which is of particular interest is when the characters are different. For instance, although DNA's consist of only 4 characters, one can make the representation more informative by assigning a symbol to every meaningful block of the sequence. This way, there is only little chance a character appears several times in the sequence. Similarly, in text recognition, one may rather represent every word with a symbol of the alphabet resulting in a huge alphabet and strings with low repetitions. These scenarios are motivating examples for the study of edit distance and longest common subsequence under this assumption, known respectively as Ulam distance [32, 33, 36] and longest increasing subsequence (which has been the target of significant research by the string algorithms community—see, e.g., [37] for references).

We thus consider LCS for input strings  $A$  and  $B$  that are non-repetitive (an important special case is when  $A$  and  $B$  are permutations). We show that there exists an  $\tilde{O}(n^{3/4})$ -time quantum algorithm for exact LCS and an  $\tilde{O}(n^{2/3})$ -time quantum algorithm for approximate LCS. This significantly improves the generic results of Theorems 1 and 3 and, for approximate LCS, this matches (up to possible polylogarithmic factors) the lower bound of Theorem 8.

**Theorem 4** *The longest common substring of two non-repetitive strings of size  $n$  can be computed with high probability by a quantum algorithm in time  $\tilde{O}(n^{3/4})$ .*

**Theorem 5** *For any constant  $0 < \epsilon < 1$ , the longest common substring of two non-repetitive strings of size  $n$  can be approximated within a factor  $1 - \epsilon$  with high probability by a quantum algorithm in time  $\tilde{O}(n^{2/3})$ .*

The improvements for non-repetitive strings are obtained by improving the complexity of the first part of the algorithm used for general LCS from  $\tilde{O}(n^{2/3}\sqrt{d})$  to  $\tilde{O}(n^{2/3} + \sqrt{nd})$  for the case of exact LCS, and from  $\tilde{O}(n^{2/3}d^{1/6})$  to  $\tilde{O}(n^{2/3}/d^{1/3} + \sqrt{n})$  for approximate LCS. We now briefly describe how we achieve such improvements.

Let us first consider exact LCS. The  $\tilde{O}(n^{2/3}\sqrt{d})$ -time quantum algorithm described above was based on a “black-box reduction” to element distinctness, i.e., we used the quantum algorithm for element distinctness (which is based, as already mentioned, on a technique known as quantum walk) as a black-box. In comparison, our new algorithm is constructed by designing a quantum walk especially tailored for our problem. More precisely, we use the approach by Magniez, Nayak, Roland and Santha [38] to design a quantum walk over the Johnson graph (more precisely, we work with a graph defined as the direct product of two Johnson graphs, which is more convenient for our purpose).

We say that a pair  $(i, j) \in [n - d + 1] \times [n - d + 1]$  is marked if there is a common substring of length  $d$  that starts at position  $i$  in  $A$  and position  $j$  in  $B$ , i.e.,  $A[i, i + d - 1] = B[j, j + d - 1]$ . The goal of our quantum walk is to find a pair of subsets  $(R_1, R_2)$  where  $R_1$  and  $R_2$  are two subsets of  $[n - d + 1]$  of size  $r$  (for some parameter  $r$ ) such that there exists a marked pair  $(i, j)$  in  $R_1 \times R_2$ . Note that since the strings are non-repetitive, for two random subsets  $R_1, R_2$ , the expected number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  is roughly  $\Theta(r^2/n)$ . A simple but crucial observation is that only those pairs can be marked since marked pairs should agree on their first character. Thus for two random subsets  $R_1, R_2$  we can check if there exists

a marked pair in  $R_1 \times R_2$  in expected time  $\tilde{O}(\sqrt{r^2/n})$  using Grover search, since we have only  $\Theta(r^2/n)$  candidates for marked pairs. This significantly improves over the upper bound  $\tilde{O}(\sqrt{r^2})$  we would get without using the assumption that the strings are non-repetitive. This is how we can improve the complexity down to  $\tilde{O}(n^{2/3} + \sqrt{nd})$ . Note that a technical difficulty that we need to overcome is guaranteeing that the running time of the checking procedure is small not only for random subsets but for also all  $(R_1, R_2)$ , i.e., we need a guarantee on the worst-case running time of the checking procedure. We solve this issue by disregarding the pairs of subsets  $(R_1, R_2)$  that contain too many candidates—we are able to prove that the impact is negligible by using concentration bounds.

The improvement for approximate LCS uses a very similar idea. The main difference is that now we consider a pair  $(i, j) \in [n - d + 1] \times [n - d + 1]$  marked if there is a common substring of length  $\lceil(1 - \epsilon)d\rceil$  that starts at position  $i$  in  $A$  and position  $j$  in  $B$ . Since the fraction of marked pairs increases by a factor  $\epsilon d$  we obtain a further improvement. Analyzing the running time of the resulting quantum walk shows that we obtain overall time complexity  $\tilde{O}(n^{2/3}/d^{1/3} + \sqrt{nd})$ .

### 2.3 Ulam Distance

Finally, we present a sublinear time quantum algorithm that computes a  $(1 + \epsilon)$ -approximation of the Ulam distance (i.e., the edit distance for non-repetitive strings).

**Theorem 6** *For any constant  $\epsilon > 0$ , there exists a quantum algorithm that computes with high probability a  $(1 + \epsilon)$ -approximation of the Ulam distance of two non-repetitive strings in time  $\tilde{O}(\sqrt{n})$ .*

In comparison, classical algorithms require linear time even for computing a constant-factor approximation of the Ulam distance when the distance is small (see, e.g., [32] for a discussion of the classical lower bounds). Theorem 6 thus shows that while for general strings it is still unknown whether a quantum speed-up is achievable for the computation of the edit distance, we can obtain a quadratic speed-up for non-repetitive strings. In Sect. 7 we show a quantum lower bound (see Theorem 11) that matches (up to possible polylogarithmic factors) the upper bound of Theorem 6. Since it is easy to show that any quantum algorithm that computes the Ulam distance exactly requires  $\Omega(n)$  time, our results are thus essentially optimal.

Let us now describe briefly how the quantum algorithm of Theorem 6 is obtained. Our approach is based on a prior work by Naumovitz, Saks, and Seshadhri [33] that showed how to construct, for any constant  $\epsilon > 0$ , a classical algorithm that computes a  $(1 + \epsilon)$ -approximation of the Ulam distance and runs in sublinear time when  $\text{ud}(A, B)$  is large (the running time becomes linear when  $\text{ud}(A, B)$  is small). The core technique is a variant of the Saks-Seshadhri algorithm for estimating the longest increasing sequence from [39], which can be used to construct a binary “indicator” (we denote this indicator  $\text{UlamIndic}$  in Sect. 6) that outputs 1 with probability  $p$  and 0 with probability  $1 - p$ , for some value  $p$  that is related to the value of  $\text{ud}(A, B)$ . Conceptually, the approach is based on estimating the value of  $\text{ud}(A, B)$  from this indicator using a

hierarchy of gap tests and estimators, each with successively better run time. This results in a fairly complex algorithm.

At a high level, our strategy is applying quantum amplitude estimation on the classical indicator  $\text{UlamIndic}$  to estimate  $p$  and thus  $\text{ud}(A, B)$ . Several technical difficulties nevertheless arise since the indicator requires a rough initial estimation of  $\text{ud}(A, B)$  to work efficiently. To solve these difficulties, we first construct a quantum gap test based on quantum amplitude estimation that enables to test efficiently if the success probability of an indicator is larger than some given threshold  $q$  or smaller than  $(1 - \eta)q$  for some given gap parameter (see Proposition 3 in Sect. 6). We then show how to apply this gap test to the indicator  $\text{UlamIndic}$  with successively better initial estimates of  $\text{ud}(A, B)$  in order to obtain a  $(1 + \epsilon)$ -approximation of  $\text{ud}(A, B)$  in  $\tilde{O}(\sqrt{n})$  time.

### 3 Longest Common Substring

Recall that in LCS, we are given two strings  $A$  and  $B$  of size  $n$  and the goal is to find the largest string  $t$  which is a substring of both  $A$  and  $B$ . This problem can be solved in time  $O(n)$  via suffix tree in the classical setting. In this section, we give upper bounds for the quantum complexity of this problem (lower bounds are discussed in Sect. 7). We first begin by giving an algorithm for exact LCS that runs in time  $\tilde{O}(n^{5/6})$  in Sect. 3.1. We then give an algorithm for approximate LCS that runs in time  $\tilde{O}(n^{3/4})$  in Sect. 3.2.

#### 3.1 Quantum Algorithm for Exact LCS

In our algorithm, we do a binary search on the size of the solution. We denote this value by  $d$ . Thus, by losing an  $O(\log n)$  factor in the runtime, we reduce the problem to verifying if a substring of size  $d$  is shared between the two strings. Our approach is twofold; we design separate algorithms for small and large  $d$ .

##### 3.1.1 Quantum Algorithm for Small $d$

Our algorithm for small  $d$  is based on a reduction to element distinctness. Let  $|\Sigma|$  be the size of the alphabet and  $v : \Sigma \rightarrow [0, |\Sigma| - 1]$  be a function that maps every element of the alphabet to a distinct number in range  $0 \dots |\Sigma| - 1$ . In other words,  $v$  is a perfect hashing for the characters. We extend this definition to substrings of the two strings. For a string  $t$ , we define  $v(t)$  as follows:

$$v(t) = \sum_{i=1}^{|t|} v(t_i) |\Sigma|^{i-1}.$$

It follows from the definition that two strings  $t$  and  $t'$  are equal if and only if we have  $v(t) = v(t')$ .

From the two strings, we make two sets of numbers  $S_A$  and  $S_B$  each having  $n - d + 1$  elements. Element  $i$  of set  $S_A$  is a pair  $(A, i)$  whose value is equal to  $v(A[i, i + d - 1])$

and similarly element  $i$  of  $S_B$  is a pair  $(B, i)$  whose value is equal to  $v(B[i, i + d - 1])$ . The two sets contain elements with equal values if and only if the two strings have a common substring of size  $d$ . Therefore, by solving element distinctness for  $S_A$  and  $S_B$  we can find out if the size of the solution is at least  $d$ . Although element distinctness can be solved in time  $\tilde{O}(n^{2/3})$  when element comparison can be implemented in  $\tilde{O}(1)$  time, our algorithm needs more runtime since comparing elements takes time  $\omega(1)$ . More precisely, each comparison can be implemented in time  $\tilde{O}(\sqrt{d})$  in the following way: In order to compare two elements of the two sets, we can use Grover’s search to find out if the two substrings are different in any position and if so we can find the leftmost position in time  $\tilde{O}(\sqrt{d})$ . Thus, it takes time  $\tilde{O}(\sqrt{d})$  to compare two elements which results in runtime  $\tilde{O}(n^{2/3}\sqrt{d})$ . We summarize this result in the following lemma.

**Lemma 1** *There exists a quantum algorithm that runs in time  $\tilde{O}(n^{2/3}\sqrt{d})$  and verifies with probability 9/10 if there is a common substring of length  $d$  between the two strings.*

### 3.1.2 Quantum Algorithm for Large $d$

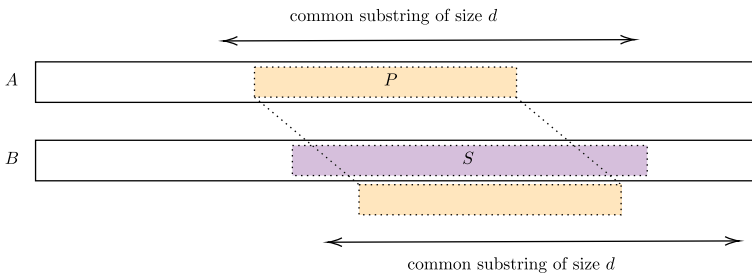
We now present a quantum algorithm that runs in time  $\tilde{O}(n/\sqrt{d})$  and verifies if there is a common substring of length  $d$  between the two strings.

**General description of the algorithm.** We say that a character of  $A$  is marked if it appears among the first  $\lfloor d/3 \rfloor$  characters of some substring of length  $d$  shared between  $A$  and  $B$ . For example, If there is exactly one common substring of length  $d$ , there are precisely  $\lfloor d/3 \rfloor$  marked characters but we may have more marked characters as the number of common subsequences of length  $d$  between  $A$  and  $B$  increases. In our algorithm, we sample a substring of length  $2\lfloor d/3 \rfloor$  of  $A$  and a substring of length  $d$  of  $B$ . We call the substring sampled from  $A$  the pattern and denote it by  $P$  and denote the substring sampled from  $B$  by  $S$ . We denote the intervals of  $A$  and  $B$  that correspond to  $P$  and  $S$  by  $[\ell_P, r_P]$  and  $[\ell_S, r_S]$  respectively. That is,  $A[\ell_P, r_P] = P$  and  $B[\ell_S, r_S] = S$ . We say that the pair  $(P, S)$  is good if the following conditions hold (see Fig. 4 for an illustration):

- There exists a pair  $(i, j)$  such that  $A[i, i + d - 1] = B[j, j + d - 1]$  is a common subsequence of size  $d$  between the two strings.
- $i \leq \ell_P$  and  $\ell_P - i < \lfloor d/3 \rfloor$ .
- $\ell_S \leq j - i + \ell_P \leq j - i + r_P \leq r_S$ .

It directly follows that if  $(P, S)$  is a good pair,  $\ell_P$  has to be a marked character. When we fix a good pair  $(P, S)$  and we refer to the optimal solution, we mean the common subsequence of size  $d$  made by  $A[i, i + d - 1]$  and  $B[j, j + d - 1]$ .

If there is no substring of length  $d$  shared between  $A$  and  $B$ , then obviously no good pair exists. Let us now compute the probability of sampling such  $P$  and  $S$  under the assumption that there is a common substring of length  $d$  shared between  $A$  and  $B$ . There are at least  $\lfloor d/3 \rfloor$  marked characters in  $A$  and thus with probability  $\Omega(d/n)$  we sample the right pattern. Moreover, the corresponding substring of solution in  $B$  has at least  $\lfloor d/3 \rfloor$  positions such that if we sample  $S$  starting from those positions the pair  $(P, S)$  is good. Thus, with probability  $\Omega(d^2/n^2)$  we sample a good pair  $(P, S)$ .



**Fig. 4** An example of a good pair  $(P, Q)$ . The orange part of  $S$  shows the part of  $S$  that matches with  $P$  (Color figure online)

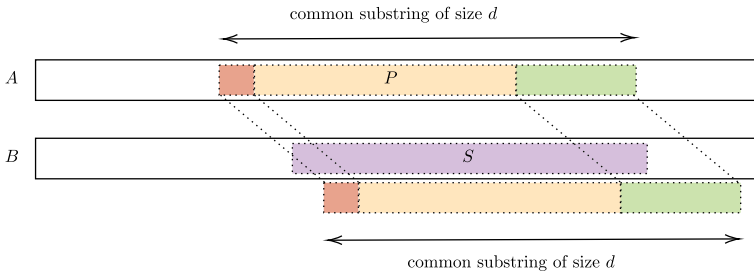
We describe below a quantum procedure that given a good pair  $(P, S)$ , constructs with probability at least  $1 - 1/\text{poly}(n)$  a common substring of length  $d$  in time  $\tilde{O}(\sqrt{d})$ . By first sampling  $(P, S)$  and then running this procedure, we get a one-sided procedure that verifies if there exists a common substring of length  $d$  shared between  $A$  and  $B$  with probability  $\Omega(d^2/n^2)$  in time  $\tilde{O}(\sqrt{d})$ . With amplitude amplification, we can thus construct a quantum algorithm that verifies if there exists a common substring of length  $d$  with high probability in time  $\tilde{O}(\sqrt{n^2/d^2} \cdot \sqrt{d}) = \tilde{O}(n/\sqrt{d})$ . We summarize this result in the following lemma.

**Lemma 2** *There exists a quantum algorithm that runs in time  $\tilde{O}(n/\sqrt{d})$  and verifies with probability  $9/10$  if there is a common substring of length  $d$  between the two strings.*

**Constructing a common substring from a good pair.** From here on, we assume that  $(P, S)$  is good and describe how to construct a common substring of length  $d$  in time  $\tilde{O}(\sqrt{d})$ .

We aim to find positions of  $S$  that match with  $P$ . For this purpose, we use the string matching algorithm of Ramesh and Vinay [19] that searches  $P$  in  $S$  in time  $\tilde{O}(\sqrt{|S|} + \sqrt{|P|})$ . If there is no match, then we can conclude that  $S$  and  $P$  do not meet our property and therefore this should not happen. If there is one such position, then we can detect in time  $\tilde{O}(\sqrt{d})$  if by extending this matching from two ends we can obtain a common substring of size  $d$  as follows: We use Grover's search to find the left-most (up to  $d$  characters away) position where the two substrings differ when extending them from right. We do the same from the left and it tells us if this gives a common substring of size at least  $d$ . Each search takes time  $\tilde{O}(\sqrt{d})$ . We refer the reader to Figure 5 for a pictorial illustration.

More generally, the above idea works when there are only  $O(1)$  many positions of  $S$  that match with  $P$ . However, we should address the case that  $P$  appears many times in  $S$ . In the following we discuss how this can be handled. In time  $\tilde{O}(\sqrt{d})$  we first find the leftmost and rightmost positions of  $S$  that match with  $P$ . Let the starting index of the leftmost match be  $\ell$  and the starting index of the rightmost match be  $r$ . In other words,  $P = S[\ell, \ell + |P| - 1] = S[r, r + |P| - 1]$ . Since  $|S| = d$  and we have  $|P| = 2\lfloor d/3 \rfloor$  then the two substrings overlap. Therefore,  $P$  as well as the entire



**Fig. 5** If  $P$  matches  $S$  in only one position, by extending the two ends in the two strings we can find a common substring of length  $d$ . The orange part of  $S$  shows the part of  $S$  that matches with  $P$ . Also, the red and green parts and extensions from left and right for the matched parts (Color figure online)

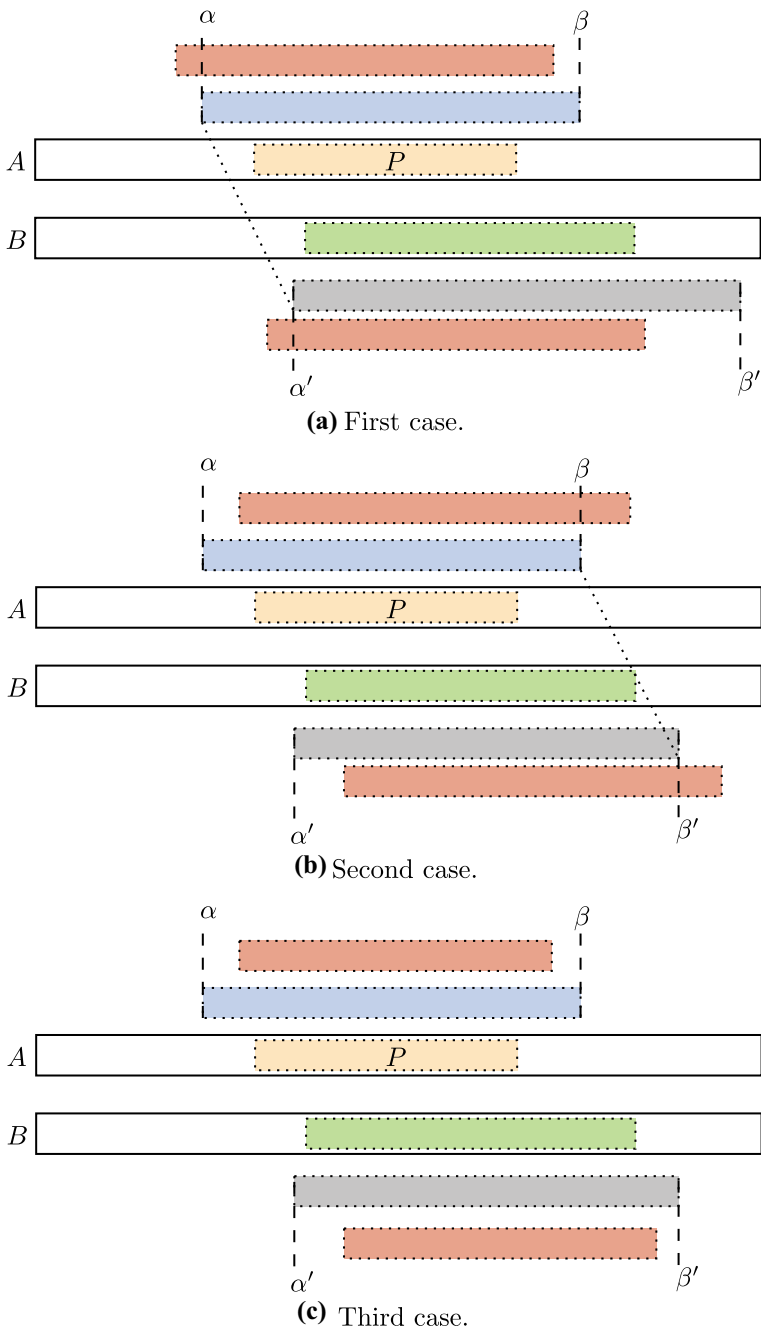
string  $S[\ell, r + |P| - 1]$  is  $(r - \ell)$ -periodic. This is the key property on which our algorithm will be based.

Let us denote the substring  $S[\ell, r + |P| - 1]$  by  $T$ . We extend both  $P$  and  $T$  from left and right up to a distance of  $2d$  in the following way: We increase the index of the ending point so long as the substring remains  $(r - \ell)$ -periodic. We also stop if we move the ending point by more than  $2d$ . We do the same for the starting point; we decrease the starting point so long as the substring remains  $(r - \ell)$ -periodic and up to at most a distance of  $2d$ . Since we bound the maximum change by  $2d$ , then this can be done in time  $\tilde{O}(\sqrt{d})$  via Grover’s search. Let us denote the resulting substrings by  $A[\alpha, \beta]$  and  $B[\alpha', \beta']$ . The following observation enables us to find the solution in time  $\tilde{O}(\sqrt{d})$ : one of the following three cases holds for the optimal solution.

1. The starting index of the corresponding optimal solution in  $A$  is smaller than  $\alpha$ . Since the matched parts of the solution are both in the periodic segments, this implies that the first non-periodic indices (when going backwards from the matched parts) in the two solution intervals are  $A[\alpha - 1]$  and  $B[\alpha' - 1]$ . As a consequence the corresponding character of  $A[\alpha]$  in  $B$  is  $B[\alpha']$  (Figure 6a).
2. The ending index of the corresponding optimal solution in  $A$  is larger than  $\beta$  and thus (with a similar argument as above) the corresponding character of  $A[\beta]$  in  $B$  is  $B[\beta']$  (Figure 6b).
3. The corresponding optimal solution in  $A$  is in the interval  $A[\alpha, \beta]$  and thus it is  $(r - \ell)$ -periodic (Figure 6c).

In all three cases, we can find the optimal solution in time  $\tilde{O}(\sqrt{d})$ . For the first two cases, we know one correspondence between the two common substrings of length  $d$ . More precisely, in Case 1 we know that  $A[\alpha]$  is part of the solution and this character corresponds to  $B[\alpha']$ . Thus, it suffices to do a Grover’s search from two ends to extend the matching in the two directions. Similarly in Case 2 we know that  $A[\beta]$  corresponds to  $B[\beta']$  and thus via a Grover’s search we find a common substring of length  $d$  in the two strings.

Finally, for Case 3 we point out that since the entire solution lies in intervals  $A[\alpha, \beta]$  and  $B[\alpha', \beta']$  then we have  $\beta \geq \alpha + d - 1$  and  $\beta' \geq \alpha' + d - 1$ . Notice that both intervals  $A[\alpha, \beta]$  and  $B[\alpha', \beta']$  are  $(r - \ell)$ -periodic. Let the starting position of  $P$  in  $A$  be  $x$ . For a fixed position  $y$  in the interval  $[\alpha', \beta']$  that matches with  $P$ , if we extend



**Fig. 6** The three cases considered. The red intervals correspond to the longest common substring. The yellow interval is  $P$  and the green interval is  $S[\ell, r + |P| - 1]$ . The blue and grey intervals are extensions of the yellow and green intervals so long as they remain  $(r - \ell)$ -periodic (Color figure online)



this matching from the two ends we obtain a solution of size  $\min\{x - \alpha, y - \alpha'\} + \min\{\beta - x + 1, \beta' - y + 1\}$ . This means that extending one of the following matchings between  $P$  and  $S$  gives us a common subsequence of size  $d$ :

- An optimal solution when  $x - \alpha \geq y - \alpha'$ : the rightmost matching in the interval  $B[\alpha', \alpha' + |P| + (x - \alpha)]$ , or
- An optimal solution when  $x - \alpha \leq y - \alpha'$ : the leftmost matching in the interval  $B[\alpha' + (x - \alpha), \beta']$ .

Each matching can be found in time  $\tilde{O}(\sqrt{d})$  and similar to the ideas explained above we can extend each matching to verify if it gives us a common substring of size  $d$  in time  $\tilde{O}(\sqrt{d})$ .

### 3.1.3 Combining the Two Algorithms

Combining Lemmas 1 and 2 gives us a solution in time  $\tilde{O}(n^{5/6})$ .

**Theorem 1** (repeated). *The longest common substring of two strings of size  $n$  can be computed with high probability by a quantum algorithm in time  $\tilde{O}(n^{5/6})$ .*

**Proof** We do a binary search on  $d$  (size of the solution). To verify a given  $d$ , we consider two cases: if  $d < n^{1/3}$  we run the algorithm of Lemma 1 and otherwise we run the algorithm of Lemma 2. Thus, the overall runtime is bounded by  $\tilde{O}(n^{5/6})$ .  $\square$

### 3.2 Quantum Algorithm for Approximate LCS

In the following, we show that if an approximate solution is desired then we can improve the runtime down to  $\tilde{O}(n^{3/4})$ . Similar to the above discussion, we use two algorithms for small  $d$  and large  $d$ . Our algorithm for large  $d$  is the same as the one we use for exact solution. For small  $d$  we modify the algorithm of Lemma 1 to improve its runtime down to  $\tilde{O}(n^{2/3}d^{1/6})$ . This is explained in Lemma 3.

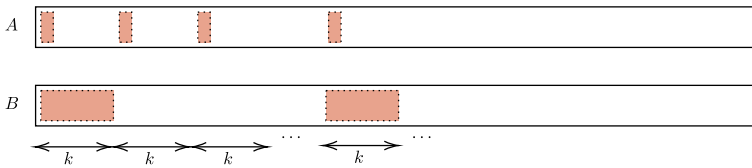
**Lemma 3** *For any constant  $0 < \epsilon < 1$ , there exists a quantum algorithm that runs in time  $\tilde{O}(n^{2/3}d^{1/6})$  and if the two strings share a common substring of length  $d$ , finds a common substring of length  $(1 - \epsilon)d$ .*

**Proof** Similar to Lemma 1, we use element distinctness for this algorithm. We make two sets  $S_A$  and  $S_B$  and prove that they share two equal elements if and only if their corresponding substrings of size  $(1 - \epsilon)d$  are equal. The difference between this algorithm and the algorithm of Lemma 1 is that here  $S_A$  and  $S_B$  contain  $O(n/\sqrt{\epsilon d})$  elements instead of  $n - d + 1$  elements.

Let  $k = \sqrt{\epsilon d}$ . We break both strings into blocks of size  $k$  and make the two sets in the following way (see Figure 7 for an illustration):

- For each  $i$  such that  $i \bmod k = 1$ , we put element  $(A, i)$  in set  $S_A$ .
- For each  $i$  such that  $\lceil i/k \rceil \bmod k = 1$  we put element  $(B, i)$  in set  $S_B$ .

It follows that among the  $\epsilon d$  first characters of the solution, there is one position which is included in  $S_A$  such that its corresponding position in  $B$  is also included



**Fig. 7**  $k = \sqrt{\epsilon d}$ . Only the elements colored in red are included in the two sets (Color figure online)

in  $S_B$ . Thus, if we define the equality between two elements  $(A, i)$  and  $(B, j)$  as  $A[i, i + (1 - \epsilon)d - 1] = B[j, j + (1 - \epsilon)d - 1]$  then by solving element distinctness for  $S_A$  and  $S_B$  we can find a desired solution. Since we have  $O(n/\sqrt{d})$  elements in each set then the overall runtime is bounded by  $\tilde{O}((n/\sqrt{d})^{2/3} \sqrt{d}) = \tilde{O}(n^{2/3} d^{1/6})$ .  $\square$

Lemmas 2 and 3 lead to a quantum algorithm for LCS with approximation factor  $1 - \epsilon$  and runtime  $\tilde{O}(n^{3/4})$ .

**Theorem 3** (repeated). *For any constant  $0 < \epsilon < 1$ , the longest common substring of two strings of size  $n$  can be approximated within a factor  $1 - \epsilon$  with high probability by a quantum algorithm in time  $\tilde{O}(n^{3/4})$ .*

**Proof** We do a binary search on  $d$ . To verify a given  $d$ , we consider two cases: if  $d < n^{1/2}$  we run the algorithm of Lemma 3 and otherwise we run the algorithm of Lemma 2. Thus, the overall runtime is bounded by  $\tilde{O}(n^{3/4})$ .  $\square$

## 4 Longest Common Substring for Non-repetitive Strings

In this section, we consider LCS for input strings  $A$  and  $B$  that are non-repetitive (e.g., permutations). We show that there exists an  $\tilde{O}(n^{3/4})$ -time quantum algorithm that computes an exact LCS and an  $\tilde{O}(n^{2/3})$ -time quantum algorithm that computes a  $(1 - \epsilon)$ -approximation of LCS. This significantly improves the generic results of the previous section.

### 4.1 Quantum Walks

We first review the quantum walk approach that solves element distinctness and several similar problems. We actually consider a variant of the standard approach that is especially well suited for our purpose: instead of considering a quantum walk on a Johnson graph we consider a quantum walk on the direct product of two Johnson graphs. While quantum walks on such graphs have also been used by Buhrman and Špalek in [14], their implementation used Szegedy's version of quantum walks [40]. One noteworthy aspect of our approach is that we rely on the stronger version of quantum walks developed by Magniez, Nayak, Roland and Santha [38] (this is crucial for the walk we present in Sect. 4.2 and 4.3; using Szegedy's version does not lead to an efficient algorithm because the checking cost of our quantum walk is too high).

Let  $S$  be a finite set and  $r$  be any integer such that  $r \leq |S|$ . Let us denote by  $\mathcal{T}_r$  the set of all couples  $(R_1, R_2)$  where both  $R_1$  and  $R_2$  are subsets of  $S$  of size  $r$ . The

elements of  $\mathcal{T}_r$  are called the states of the walk. Let  $\mathcal{T}_r^* \subseteq \mathcal{T}_r$  be some subset. The elements of  $\mathcal{T}_r^*$  are called the marked states. The quantum walk approach associates to each  $(R_1, R_2) \in \mathcal{T}_r$  a data structure  $D(R_1, R_2)$ . Three types of costs are associated with  $D$ . The setup cost  $s(r)$  is the cost to set up the data structure  $D(R_1, R_2)$  for any  $(R_1, R_2) \in \mathcal{T}_r$ . The update cost  $u(r)$  is the cost to update  $D(R_1, R_2)$  for any  $(R_1, R_2) \in \mathcal{T}_r$ : converting  $D(R_1, R_2)$  to  $D((R_1 \setminus a) \cup \{a'\}, (R_2 \setminus b) \cup \{b'\})$  for any  $a \in R_1, a' \in S \setminus R_1, b \in R_2$  and  $b' \in S \setminus R_2$ . The checking cost  $c(r)$  is the cost of checking, given  $D(R_1, R_2)$  for any  $(R_1, R_2) \in \mathcal{T}_r$ , if  $(R_1, R_2) \in \mathcal{T}_r^*$ . By combining amplitude amplification and quantum walks, we can find with high probability one marked state  $(R_1, R_2)$ . We will use in this paper the following statement (we refer to [38] for details) of this approach.

**Proposition 1** *Consider any value  $r \in \{1, \dots, |S|\}$  and any constant  $\gamma > 0$ . Assume that either  $\mathcal{T}_r^* = \emptyset$  or  $\frac{|\mathcal{T}_r^*|}{|\mathcal{T}_r|} \geq \delta$  holds for some known value  $\delta > 0$ . There exists a quantum algorithm that always rejects if  $\mathcal{T}_r^* = \emptyset$ , outputs a set  $(R_1, R_2) \in \mathcal{T}_r^*$  with probability at least  $1 - \frac{1}{n^\gamma}$  otherwise, and has complexity*

$$\tilde{O}\left(s(r) + \sqrt{1/\delta}(\sqrt{r} \times u(r) + c(r))\right).$$

Let us now explain how to solve element distinctness using this framework. Let  $x_1, \dots, x_n$  denote the  $n$  elements of the first list, and  $y_1, \dots, y_n$  denote the  $n$  elements of the second list of the input of element distinctness. Let us first discuss the complexity in the standard access model, where we assume that each element is encoded using  $O(\log n)$  bits and that given  $i \in [n]$  we can obtain all the bits of  $x_i$  and  $y_i$  in  $O(\log n)$  time. We set  $S = \{1, \dots, n\}$  and say that a state  $(R_1, R_2) \in \mathcal{T}_r$  is marked if there exists a pair  $(i, j) \in R_1 \times R_2$  such that  $x_i = y_j$ . Easy calculations show that the fraction of marked states is  $\Omega(r^2/n^2)$ . By using an appropriate data structure that allows insertion, deletion and lookup operations in polylogarithmic time (see Section 6.2 of [13] for details about how to construct such a data structure) to store the elements  $x_i$  for all  $i \in R_1$ , and using another instance of this data structure to store the elements  $y_j$  for all  $j \in R_2$ , we can perform the setup operation in time  $s(r) = \tilde{O}(r)$  and perform each update operation in time  $u(r) = \tilde{O}(1)$ . Each checking operation can be implemented in time  $c(r) = \tilde{O}(\sqrt{r})$  as follows: perform a Grover search over  $R_1$  to check if there exists  $i \in R_1$  for which  $x_i = y_j$  for some  $j \in R_2$  (once  $i$  is fixed the later property can be checked in polylogarithmic time using the data structure representing  $R_2$ ).<sup>5</sup> From Proposition 1, the overall time complexity of the quantum walk is thus

$$\tilde{O}\left(r + \frac{n}{r}(\sqrt{r} \times 1 + \sqrt{r})\right).$$

By taking  $r = n^{2/3}$  we get time complexity  $\tilde{O}(n^{2/3})$ . As discussed in Section 6.1 of [13], the above data structure, and thus the whole quantum walk can also be implemented in the same way in the (weaker) comparison model: if given any pair of indices

<sup>5</sup> The checking operation can actually be implemented more efficiently but the cost  $c(r) = \tilde{O}(\sqrt{r})$  is enough for our purpose.

$(i, j) \in [n] \times [n]$  we can decide in  $T(n)$  time whether  $x_i < y_j$  or  $x_i \geq y_j$ , then the time complexity of the implementation is  $\tilde{O}(n^{2/3}T(n))$ .

#### 4.2 Quantum Algorithm for Exact LCS of Non-repetitive Strings

We set  $S = \{1, \dots, n - d + 1\}$ . Let us write  $\alpha = 54 \log n$ . For any state  $(R_1, R_2) \in \mathcal{T}_r$  we define the data structure  $D(R_1, R_2)$  as follows:

- $D(R_1, R_2)$  records all the values  $A[i]$  and  $B[j]$  for all  $i \in R_1$  and  $j \in R_2$  using the same data structure as the data structure considered above for element distinctness;
- additionally,  $D(R_1, R_2)$  records the number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$ , and stores explicitly all these pairs (using a history-independent data structure updatable in polylogarithmic time, which can be constructed based on the data structure from [13]).

We define the set of marked states  $\mathcal{T}_r^*$  as follows. We say that  $(R_1, R_2)$  is marked if the following two conditions hold:

- there exists a pair  $(i, j) \in R_1 \times R_2$  such that  $A[i, i + d - 1] = B[j, j + d - 1]$ ;
- the number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  is at most  $\lceil \alpha(r^2/n + 1) \rceil$ .

The following lemma is easy to show.

**Lemma 4** *If the two non-repetitive strings  $A$  and  $B$  have a common substring of length  $d$ , then the fraction of marked states is  $\Omega(r^2/n^2)$ .*

**Proof** The fraction of states  $(R_1, R_2)$  for which there exists a pair  $(i, j) \in R_1 \times R_2$  such that  $A[i, i + d - 1] = B[j, j + d - 1]$  is at least

$$\frac{\binom{n-2}{r-2}}{\binom{n}{r}} = \frac{r(r-1)}{n(n-1)}.$$

The fraction of  $(R_1, R_2)$  such that Condition (i) holds is thus  $\Omega(r^2/n^2)$ .

Let  $X$  be the random variable representing the number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  when  $(R_1, R_2)$  is taken uniformly at random in  $\mathcal{T}_r$ . We will show that  $X \leq \lceil \alpha(r^2/n + 1) \rceil$  with high probability. Let us prove this upper bound for the worst possible case: input strings for which the number of pairs  $(i, j) \in S \times S$  such that  $A[i] = B[j]$  is  $|S|$ .

Assume that we have fixed  $R_1$ . For each  $i \in R_1$  there exists a unique index  $j \in S$  such that  $A[i] = B[j]$ . Let  $Y$  be the random variable representing the total number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  when choosing  $R_2$ . Note that  $Y$  follows an hypergeometric distribution of mean  $r^2/n$ . From standard extensions of Chernoff's bound (see, e.g., Section 1.6 of [41]), we get

$$\Pr \left[ Y \geq \alpha \left( \frac{r^2}{n} + 1 \right) \right] \leq \exp \left( - (3 \log n) \cdot \left( \frac{r^2}{n} + 1 \right) \right) \leq \frac{1}{n^3}.$$

Thus the fraction of  $(R_1, R_2)$  such that Condition (ii) does not hold is at most  $1/n^3$ .

The statement of the lemma then follows from the union bound. □

Let us now analyze the costs corresponding to this quantum walk. The setup and update operations are similar to the corresponding operations of the quantum walk for element distinctness described in Sect. 4.1. The only difference is that we need to keep track of the pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$ . Note that for each  $i \in [n]$ , there is at most one index  $j \in R_2$  such that  $A[i] = B[j]$ , since the strings are non-repetitive. Moreover this index can be found in polylogarithmic time using the data structure associated with  $R_2$ . Similarly, for each  $j \in [n]$ , there is at most one index  $i \in R_1$  such that  $A[i] = B[j]$ , which can be also found in polylogarithmic time. Thus the time complexities of the setup and update operations are  $s(r) = \tilde{O}(r)$  and  $u(r) = \tilde{O}(1)$ , respectively, as in the quantum walk for element distinctness described in Sect. 4.1. The checking operation first checks if the number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  is at most  $\lceil \alpha(r^2/n + 1) \rceil$  and then, if this condition is satisfied, performs a Grover search on all pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  stored in  $D(R_1, R_2)$ , in order to check if there exists a pair  $(i, j) \in R_1 \times R_2$  such that  $A[i, i + d - 1] = B[j, j + d - 1]$ . The time complexity is

$$c(r) = \tilde{O}(1 + \sqrt{\lceil \alpha(r^2/n + 1) \rceil} \cdot \sqrt{d}).$$

Using Proposition 1, we get that the overall time complexity of the quantum walk is

$$\tilde{O}\left(r + \frac{n}{r}(\sqrt{r} \times 1 + \sqrt{\lceil \alpha(r^2/n + 1) \rceil} \cdot \sqrt{d})\right) = \tilde{O}\left(r + \frac{n}{\sqrt{r}} + \frac{n\sqrt{d}}{r} + \sqrt{nd}\right).$$

For  $d \leq n^{1/3}$ , this expression is minimized for  $r = n^{2/3}$  and gives complexity  $\tilde{O}(n^{2/3})$ . For  $d \geq n^{1/3}$ , the complexity is dominated by the last term  $\tilde{O}(\sqrt{nd})$ .

By combining the above algorithm with the  $\tilde{O}(n/\sqrt{d})$ -time algorithm of Lemma 2, we get overall complexity  $\tilde{O}(n^{3/4})$ .

**Theorem 4 (repeated).** *The longest common substring of two non-repetitive strings of size  $n$  can be computed with high probability by a quantum algorithm in time  $\tilde{O}(n^{3/4})$ .*

**Proof** We do a binary search on  $d$ . To verify a given  $d$ , we consider two cases: if  $d < \sqrt{n}$  we run the  $\tilde{O}(\sqrt{nd})$ -time quantum algorithm we just described and otherwise we run the algorithm of Lemma 2. Thus, the overall runtime is bounded by  $\tilde{O}(n^{3/4})$ . □

### 4.3 Quantum Algorithm for Approximate LCS of Non-repetitive Strings

We set again  $S = \{1, \dots, n - d + 1\}$  and, for any  $(R_1, R_2) \in \mathcal{T}_r$ , we define the data structure  $D(R_1, R_2)$  exactly as above. This time, we say that  $(R_1, R_2)$  is marked if the following two conditions hold:

- (i) there exists a pair  $(i, j) \in R_1 \times R_2$  such that  $A[i, i + \lceil (1 - \epsilon)d \rceil - 1] = B[j, j + \lceil (1 - \epsilon)d \rceil - 1]$ ;

(ii) the number of pairs  $(i, j) \in R_1 \times R_2$  such that  $A[i] = B[j]$  is at most  $\lceil \alpha r^2/n \rceil$ ; where we use the same value  $\alpha = 54 \log n$  as above. We now show the following lemma.

**Lemma 5** *Assume that  $r \leq n/d$ . If  $A$  and  $B$  have a common subsequence of length  $d$ , then the fraction of marked sets is  $\Omega\left(\frac{dr^2}{n^2}\right)$ .*

**Proof** Let us write  $m = d - \lceil(1 - \epsilon)d\rceil$  and say that a pair  $(i, j) \in S \times S$  is good if  $A[i, i + \lceil(1 - \epsilon)d\rceil - 1] = B[j, j + \lceil(1 - \epsilon)d\rceil - 1]$ . There are at least  $m$  good pairs in  $S \times S$ . The probability that  $R_1$  does not contain any element involved in a good pair is

$$\begin{aligned} \frac{\binom{n-m}{r}}{\binom{n}{r}} &= \frac{(n-m)!r!(n-r)!}{r!(n-m-r)!n!} \\ &= \frac{(n-m)(n-m-1)\cdots(n-m-r+1)}{n(n-1)\cdots(n-r+1)} \\ &\leq \left(\frac{n-m}{n}\right)^r = \left(1 - \frac{m}{n}\right)^r \leq \exp\left(-\frac{rm}{n}\right) \\ &\leq 1 - \frac{rm}{2n}, \end{aligned}$$

where we used the fact that  $\exp(-x) \leq 1 - x/2$  on  $x \in [0, 1]$  to derive the last inequality. Thus with probability at least  $rm/(2n)$ , the set  $R_1$  contains at least one element involved in a good pair.

Assuming that  $R_1$  contains at least one element involved in a good pair, when choosing  $R_2$  at random with probability at least  $r/n$  it contains a good pair. Thus the overall probability that  $R_1 \times R_2$  contains a good pair (and thus satisfies Condition (i)) is at least  $\Omega(dr^2/n^2)$ .

The fraction of  $(R_1, R_2)$  such that Condition (ii) does not hold is at most  $1/n^3$  (the proof of this claim is exactly the same as for Lemma 4). The statement of the lemma then follows from the union bound.  $\square$

We thus get a quantum walk with running time

$$\tilde{O}\left(r + \frac{n}{r\sqrt{d}}(\sqrt{r} \times 1 + \sqrt{\lceil \alpha(r^2/n + 1) \rceil} \cdot \sqrt{d})\right) = \tilde{O}\left(r + \frac{n}{\sqrt{rd}} + \frac{n}{r} + \sqrt{n}\right).$$

If  $d \leq \sqrt{n}$  then the expression is minimized for  $r = n^{2/3}/d^{1/3}$  and the complexity is  $\tilde{O}(n^{2/3}/d^{1/3})$ . (Note that with this value of  $r$  the condition of Lemma 5 is satisfied.) If  $d > \sqrt{n}$  we obtain complexity  $\tilde{O}(d)$  by taking  $r = n/d$ .<sup>6</sup>

By combining the above algorithm with the  $\tilde{O}(n/\sqrt{d})$ -time algorithm of Lemma 2, we get overall complexity  $\tilde{O}(n^{2/3})$ .

<sup>6</sup> For  $d > \sqrt{n}$ , we can actually obtain the better upper bound  $\tilde{O}(\sqrt{n})$  by slightly modifying the algorithm. The bound  $\tilde{O}(d)$  is nevertheless sufficient for our purpose.

**Theorem 5** (repeated). *For any constant  $0 < \epsilon < 1$ , the longest common substring of two non-repetitive strings of size  $n$  can be approximated within a factor  $1 - \epsilon$  with high probability by a quantum algorithm in time  $\tilde{O}(n^{2/3})$ .*

**Proof** We do a binary search on  $d$ . To verify a given  $d$ , we consider two cases: if  $d < n^{2/3}$  we run the  $\tilde{O}(n^{2/3}/d^{1/3} + d)$ -time quantum algorithm we just described and otherwise we run the algorithm of Lemma 2. Thus, the overall runtime is bounded by  $\tilde{O}(n^{2/3})$ .  $\square$

## 5 Longest Palindrome Substring

In this section, we present a quantum algorithm that solves the longest palindrome substring problem (LPS) in time  $\tilde{O}(\sqrt{n})$ . In Sect. 7, we complement this result by a tight lower bound of  $\tilde{\Omega}(\sqrt{n})$  for quantum algorithms even for the special case of 0/1 strings and even for approximating the solution.

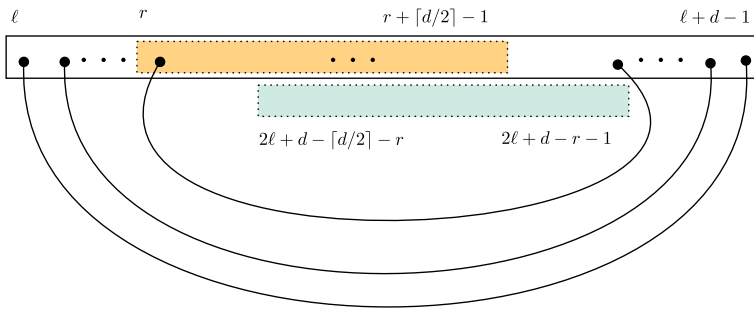
### 5.1 General Description of Our Quantum Algorithm

Our algorithm starts with a binary search on the size of the solution. We use  $l$  and  $u$  as a lower bound and an upper bound on the size of the solution. Initially we set  $l = 1$  and  $u = n$ . Each time we set  $d = \lceil (l + u)/2 \rceil$  and try to find a solution of size at least  $d$ . If such a solution exists, then we continue on by setting  $l = d$ . Otherwise we set  $u = d - 1$ . After  $O(\log n)$  steps we have  $l = u$  which is equal to the size of the solution. Thus, in the following, we assume that an integer  $d$  is given and our goal is to find out if a solution of size at least  $d$  exists in the string.

One thing to keep in mind is that if the solution size is larger than  $d$ , we may not necessarily have a palindrome substring of size  $d$  but in that case, we certainly have a palindrome substring of size  $d + 1$ . Thus, all it takes to verify if the solution size is at least  $d$  is to look for a solution of size either  $d$  or  $d + 1$ , and we can simply consider each case separately. From here on, we assume that our string contains a palindrome substring of size exactly  $d$  and we wish to find such a substring.

For each palindrome substring of length  $d$ , we mark the first  $\lfloor d/2 \rfloor$  characters of this substring. We will show in Sect. 5.2 how to check if a character is marked or not in time  $\tilde{O}(\sqrt{d})$ . Since we may have several palindrome substrings of length  $d$  we may have more than  $d/2$  marked characters in our string. However, since there is at least one solution of size  $d$ , then the number of marked characters is at least  $\Omega(d)$ . Our quantum algorithm applies Grover algorithm to find one marked character. The running time is  $\tilde{O}(\sqrt{n/d} \cdot \sqrt{d}) = \tilde{O}(\sqrt{n})$ . We thus obtain the following result.

**Theorem 2** (repeated). *The longest palindrome substring of a string of size  $n$  can be computed with high probability by a quantum algorithm in time  $\tilde{O}(\sqrt{n})$ .*



**Fig. 8** If  $A[\ell, \ell + d - 1]$  is a palindrome, then the blue substring is equal to the reverse of the orange substring (Color figure online)

## 5.2 Description of the Checking Procedure

We now describe the main idea of the procedure that checks whether a given position  $r$  is a marked character or not. This is the main technical contribution of the whole section.

Define the center of a palindrome substring as the average of its starting position and its ending position (it is not integer if the length of the palindrome substring is even). Note that if we fix a center  $c$ , we can via Grover's search determine in time  $\tilde{O}(\sqrt{d})$  whether the size of the corresponding palindrome substring is  $d$  or not.

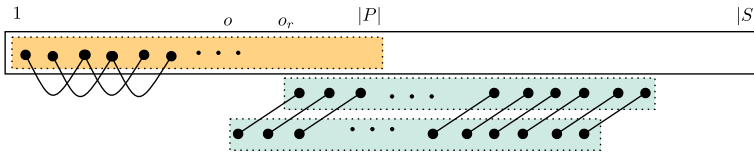
We now bring a crucial observation on which the checking procedure is built. Assume that position  $r$  of the input string is marked. Let  $\ell$  be the leftmost character of a corresponding solution, i.e., the substring  $A[\ell, \ell + d - 1]$  of the input array is palindrome and position  $r$  is one of the first  $\lceil d/2 \rceil$  characters of this interval. Since  $A[\ell, \ell + d - 1]$  is palindrome then we have that  $A[r, r + \lceil d/2 \rceil - 1]$  is equal to the reverse of  $A[2\ell + d - \lceil d/2 \rceil - r, 2\ell + d - r - 1]$ . See Figure 8 for an illustration.

Let  $S$  be the substring of length  $d$  that spans the interval  $[r, r + d - 1]$  of the input string. The main strategy of our checking procedure is to identify positions of  $S$  that are candidates for the center of the solution. For this purpose we also define a pattern  $P$  of length  $\lceil d/2 \rceil$  which is equal to the reverse of interval  $[r, r + \lceil d/2 \rceil - 1]$ . We say a position  $i$  of  $S$  matches with pattern  $P$  if  $S[i, i + |P| - 1] = P$ .

If  $r$  is marked, then it is in the left half of some palindrome substring  $A[\ell, \ell + d - 1]$ . Moreover, the size of  $S$  is equal to  $d$  and by the way we construct  $P$  it is equal to the reverse of  $A[r, r + \lceil d/2 \rceil - 1]$ . Thus,  $A[2\ell + d - \lceil d/2 \rceil - r, 2\ell + d - r - 1]$  is equal to  $P$  and completely lies inside  $S$ . Thus, by finding all positions of  $S$  that match with  $P$ , we definitely find  $A[2\ell + d - \lceil d/2 \rceil - r, 2\ell + d - r - 1]$ . In particular, if there is only one position of  $S$  that matches with  $P$ , this enables to identify the center of  $A[\ell, \ell + d - 1]$ .

In Sect. 5.2.1 below we show how to identify efficiently all the positions of  $S$  that match with pattern  $P$ . These positions give a list of candidates for the center of a solution. At least one of them is the center of a solution. If there is only one candidate (or a constant number of candidates), we can then via Grover's search determine in time  $\tilde{O}(\sqrt{d})$  whether the size of the corresponding palindrome substring is  $d$  or not.





**Fig. 9** The green substring is equal to pattern  $P$  and the orange substring is the reverse of  $P$ . Since  $P$  is  $(o_r - o)$ -periodic, then the entire substring of  $S[1, o_r + |P| - 1]$  is also  $(o_r - o)$ -periodic (Color figure online)

In Sect. 5.2.2, we explain how to deal with the case where there are many candidates for the center.

### 5.2.1 Identifying the Patterns

Below we describe how to identify all positions of the  $S$  that match with pattern  $P$  by considering the periodicity of  $P$ . We show that this can be done in time  $\tilde{O}(\sqrt{d})$ .

To do so, we first find the rightmost occurrence of the  $P$  in  $S$  using the quantum algorithm of Ramesh and Vinay [19]. This takes time  $\tilde{O}(\sqrt{d})$  since  $|S| + |P| = O(d)$ . If  $P$  does not appear in  $S$  at all or only once, we are done. Otherwise, let  $o_r$  be the rightmost position of  $S$  that matches with  $P$ . That is  $S[o_r, o_r + |P| - 1] = P$ . Moreover, let  $o$  be the position of the second rightmost occurrence of  $P$  in  $S$ . We argue that  $T = \{o_r - i(o_r - o) \mid 0 \leq i \leq \lfloor (o_r - 1)/(o_r - o) \rfloor\}$  is the set of all occurrences of  $P$  in  $S$ . More precisely, every element of  $T$  is the starting position of an occurrence of  $P$  in  $S$ . Since  $|S| \leq 2|P|$ , there is one special case in which  $o = o_r - |P|$ . In this case  $|T|$  contains exactly two elements and the proof is trivial. Thus, we assume in the following that the two rightmost occurrences of  $P$  in  $S$  overlap.

Remember that we say a string  $s$  is  $q$ -periodic if we have  $s_i = s_{i+q}$  for all  $1 \leq i \leq |s| - q$ . Since  $P$  appears as a substring of  $S$  at both locations  $o_r$  and  $o$ , the pattern  $P$  is  $(o_r - o)$ -periodic. Moreover,  $P$  is not  $i$ -periodic for any  $1 < i < o_r - o$  since otherwise  $o$  would not be the position of the second rightmost occurrence of  $P$  in  $S$ . We first prove that each element of  $T$  is a position in  $S$  that matches with  $P$ . Recall that by the way we construct  $S$  and  $P$ ,  $P$  is equal to the reverse of the first  $|P|$  characters of  $S$ . Moreover,  $|S| \leq 2|P|$  and thus the entire interval  $S[1, o_r + |P| - 1]$  is  $(o_r - o)$ -periodic (we refer to Figure 9 for an illustration). This implies that every element of  $T$  is the beginning of one occurrence of  $P$  in  $S$ .

In order to show that  $T$  contains all occurrences of  $P$  in  $S$ , assume for the sake of contradiction that there is a position  $p$  of  $S$  that is the beginning of one occurrence of  $P$  but  $p$  is not included in  $T$ . This implies that there is one element  $e \in T$  such that  $|e - p| < o_r - o$ . Thus, with the same argument  $P$  is  $|e - p|$ -periodic. This is in contradiction with what we proved previously.

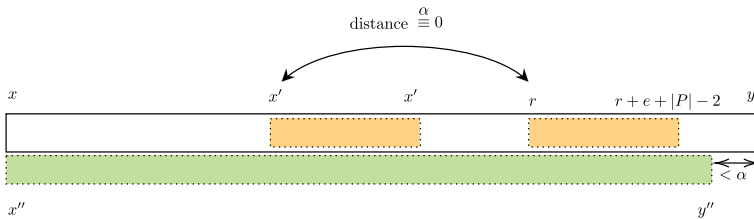


Fig. 10 Illustration for the proof of Lemma 6. All the colored intervals are palindrome (Color figure online)

### 5.2.2 Checking the Patterns

We have shown that the set  $T$  computed in Sect. 5.2.1 is the set of the starting positions of each occurrence of  $P$  in  $S$ . We define

$$C = \{r - 1 + (e + \lceil d/2 \rceil)/2 \mid e \in T\},$$

which is the set of all possible centers for the solution. Since we consider all occurrences of  $P$  in  $S$ , one of the elements in  $C$  is equal to the actual center of our solution. If  $|C| = O(1)$ , one can iterate over all elements of  $C$  and verify in time  $O(\sqrt{d})$  if each one is a center for a palindrome substring of size  $d$ . In the following, we show that even if  $|C|$  is large, we can narrow down the search to a constant number of elements in  $C$ , which makes it possible to find efficiently a palindrome substring of length  $d$ .

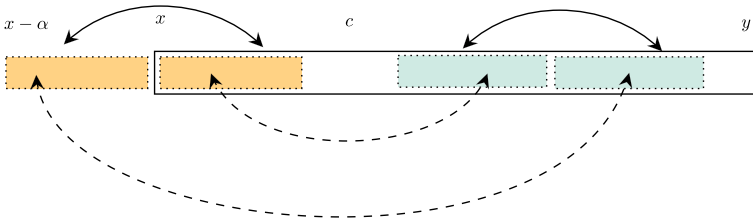
We showed in Sect. 5.2.1 that  $C$  is large only if  $P$  is periodic with small periodicity. Let  $\alpha$  be the periodicity of  $P$ . As we discuss in Sect. 5.2.1, every consecutive pair of elements in  $T$  have distance  $\alpha$ . Thus, each pair of consecutive elements in  $C$  have distance  $\alpha/2$ . Let  $y$  be the largest index of  $A$  such that  $A[r, y]$  is  $\alpha$ -periodic. Also, let  $x$  be the smallest index of  $A$  such that  $A[x, y]$  is  $\alpha$ -periodic. We prove two things in the following:

- 1) If  $y - x + 1 \geq d + \alpha$  then we already have a palindrome substring of size  $d$  in the interval  $[x, y]$  that can easily be found in time  $\tilde{O}(\sqrt{d})$ ;
- 2) Otherwise the center of our solution has a distance of at most  $\alpha/2$  from  $(x + y)/2$ , and thus we only need to consider a constant number of elements in  $C$ , which enables us to solve the problem in time  $\tilde{O}(\sqrt{d})$  as well.

Here is the proof of the first statement (we refer to Fig. 10 for an illustration).

**Lemma 6** *If  $y - x + 1 \geq d + \alpha$  then one can find a palindrome substring of size at least  $d$  in time  $\tilde{O}(\sqrt{d})$ .*

Let  $e \in T$  be an arbitrary element of  $T$ . This means that  $P = A[r + e - 1, r + e - 1 + |P| - 1]$ . Since  $A[r, r + |P| - 1]$  is equal to the reverse of  $P$  and the two intervals  $[r, r + |P| - 1]$  and  $[r + e - 1, r + e - 1 + |P| - 1]$  overlap then  $A[r, r + e - 1 + |P| - 1]$  is a palindrome substring. Recall that  $P$  is  $\alpha$ -periodic and thus so is the entire substring  $A[r, r + e - 1 + |P| - 1]$ . This implies that interval  $[r, r + e - 1 + |P| - 1]$  is inside interval  $[x, y]$ . We start by  $x' = r$  and  $y' = r + e - 1 + |P| - 1$ . Next, we shift  $[x', y']$  within  $[x, y]$  by multiples of  $\alpha$  until  $||x - x'| - |y - y'|| \leq \alpha$ . Since  $A[x, y]$  is  $\alpha$ -periodic, we



**Fig. 11** Illustration for the proof of Lemma 7. Dashed arrows show that the two strings are the reverse of each other. Solid arrows show that the two strings are equal

still have  $A[x', y'] = A[r, r + e - 1 + |P| - 1]$  and therefore  $A[x', y']$  is palindrome. Finally, we set  $x'' = x' - \min\{|x - x'|, |y - y'|\}$  and  $y'' = y' + \min\{|x - x'|, |y - y'|\}$ . It follows from the periodicity of  $A[x, y]$  that  $A[x'', y'']$  is palindrome. Moreover, the size of  $[x'', y'']$  is at most  $\alpha$  smaller than the size of  $[x, y]$ . Thus,  $A[x'', y'']$  is a desired substring.

It is easy to find such a solution in time  $\tilde{O}(\sqrt{d})$ . The only part of the proof which is non-trivial from a computational standpoint is finding  $[x, y]$ . If  $y - x = O(d)$  then one can find it via Grover’s search in time  $\tilde{O}(\sqrt{d})$ . Otherwise, instead of finding  $x$  and  $y$ , we extend  $r$  from both ends up to a distance of  $2d$  so long as it remains  $\alpha$ -periodic. This can be done in time  $\tilde{O}(\sqrt{d})$  and then with the same analysis we can find a palindrome substring of size at least  $d$ . □

The proof of the second statement relies on the following lemma (we refer to Fig. 11 for an illustration).

**Lemma 7** *If  $y - x + 1 < d + \alpha$  then the center of the solution cannot be more than  $2\alpha$  away from  $(x + y)/2$ .*

We denote the optimal solution by  $A[\ell, \ell + d - 1]$ . Let  $c = \ell + (d - 1)/2$  be the center of the solution. Assume for the sake of contradiction that  $c < (x + y)/2 - 2\alpha$ . This means that  $\ell < x - \alpha$ . Thus, the interval  $A[x - \alpha, x - 1]$  is exactly equal to the interval  $A[x, x + \alpha - 1]$  and thus  $A[x - \alpha, y]$  is also  $\alpha$ -periodic. This is in contradiction with the maximality of  $[x, y]$ . A similar proof holds for the case that  $c > (x + y)/2 + 2\alpha$ . □

Lemma 7 implies that we only need to consider all candidates in  $C$  that are within the range  $[(x + y)/2 - 2\alpha, (x + y)/2 + 2\alpha]$ . Moreover, the distance between every pair of candidates in  $C$  is  $\alpha/2$  and therefore we only need to consider a constant number of elements. Therefore, we can find the solution in time  $\tilde{O}(\sqrt{d})$ .

### 6 (1 + ε)-Approximation of the Ulam Distance

In this section we prove the following theorem.

**Theorem 6** (repeated). *For any constant  $\epsilon > 0$ , there exists a quantum algorithm that computes with high probability a  $(1 + \epsilon)$ -approximation of the Ulam distance of two non-repetitive strings in time  $\tilde{O}(\sqrt{n})$ .*

## 6.1 Classical Indicator for the Ulam Distance

Naumovitz, Saks and Seshadhri [33] showed how to construct, for any constant  $\epsilon > 0$ , a classical algorithm that computes a  $(1 + \epsilon)$ -approximation of the Ulam distance of two non-repetitive strings  $A, B$  of length  $n$  in time  $\tilde{O}(n/\text{ud}(A, B) + \sqrt{n})$ . Their algorithm is complex: it consists of nine procedures that form a hierarchy of gap tests and estimators, each with successively better run time. The core technique, which lies at the lowest layer of the hierarchy and gives a very good—but slow—estimation of the Ulam distance, is a variant of the Saks–Seshadhri algorithm for estimating the longest increasing sequence from [39].

For our purpose, we will only need the following result from [33].<sup>7</sup>

**Proposition 2** ([33]) *Let  $\delta > 0$  be any constant. For any two non-repetitive strings  $A, B$  of length  $n$  and any integer parameter  $t' \geq c \cdot \text{ud}(A, B)$  for some constant  $c$  depending only on  $\delta$ , there exists a  $\tilde{O}(\sqrt{t'})$ -time classical algorithm  $\text{UlamIndic}(A, B, \delta, t')$  that outputs 1 with some probability  $p$  and 0 with probability  $(1 - p)$ , for some probability  $p$  such that*

$$p \in \left[ (1 - \delta) \cdot \frac{\text{ud}(A, B)}{n + t'}, (1 + \delta) \cdot \frac{\text{ud}(A, B)}{n + t'} + \frac{\delta}{n + t'} \right].$$

The procedure of Proposition 2 lies at one of the lowest layers of the hierarchy in [33], and works by applying the Saks–Seshadhri algorithm from [39] on randomly chosen substrings of  $A$  and  $B$  (of size roughly  $t$ ).<sup>8</sup>

## 6.2 Quantum Algorithm for the Ulam Distance

To prove Theorem 6, the basic idea is to apply quantum amplitude estimation on the classical algorithm  $\text{UlamIndic}$  from Proposition 2. Let us explain this strategy more precisely and show the technical difficulties we need to overcome. Assume that we know that  $\text{ud}(A, B)$  is in the interval  $[D_1, D_2]$  for some values  $D_1$  and  $D_2$ . We can then use  $\text{UlamIndic}(A, B, \delta, t)$  with  $t = c \cdot D_2$ , and apply the quantum amplitude estimation algorithm of Theorem 7 to estimate the probability it outputs 1. If we use  $k = \Theta(\sqrt{(n+t)/D_1}) = \Theta(\sqrt{n/D_1})$  in Theorem 7 and  $\delta$  small enough, we will get a good approximation of the quantity  $\text{ud}(A, B)/(n + t)$ , and thus a good approximation of  $\text{ud}(A, B)$ . The complexity of this strategy is  $\tilde{O}(\sqrt{n/D_1} \cdot \sqrt{D_2})$ , which is  $\tilde{O}(\sqrt{n})$  when  $D_1 \approx D_2$ .

<sup>7</sup> The precise statement of this result appears in Lemma 8.5 and Table 1 in [33]. The procedure is denoted  $\text{XL}_2$  in [33]. Note that the original statement is actually more general: it considers strings  $A, B$  of different lengths and gives an indicator for a slightly different quantity (called  $\text{Xloss}(A, B)$  in [33]). When  $A$  and  $B$  have the same length we have  $\text{Xloss}(A, B) = \frac{1}{2}\text{ud}(A, B)$ , which gives the statement of Proposition 2.

<sup>8</sup> We stress that this procedure is still too slow to get a  $\tilde{O}(n/\text{ud}(A, B) + \sqrt{\text{ud}(A, B)})$ -time classical algorithm. Indeed, getting a good approximation of  $\text{ud}(A, B)$  based only on this procedure would require repeating it  $\Omega((n+t)/\text{ud}(A, B))$  times, which would require  $\Omega(\sqrt{\text{ud}(A, B)} \cdot (n+t)/\text{ud}(A, B)) = \Omega(n/\sqrt{\text{ud}(A, B)})$  time. This is why several additional techniques (which lead to several additional layers in the hierarchy) are used in [33].

**Fig. 12** Quantum gap test

1. Apply the algorithm of Theorem 7 with  $k = \frac{20}{\eta\sqrt{q}}$ . Let  $\tilde{p}$  denote the output.
2. Output LARGE if  $\tilde{p} \geq (1 - \eta/2)q$  and output SMALL if  $\tilde{p} < (1 - \eta/2)q$ .

The main issue is that we do not know such tight upper and lower bounds on  $ud(A, B)$ . Concerning the upper bound, we overcome this difficulty by simply successively trying  $D_2 = n, D_2 = (1 - \eta)n, D_2 = (1 - \eta)^2n, \dots$  for some small constant  $\eta$  (for technical reasons we actually start from  $D_2 = (1 - \eta)(1 - \epsilon)\sqrt{n}/c$ , and deal with the case of larger  $D_2$  using the classical algorithm from [33]). For the lower bound, on the other hand, we cannot simply start with  $D_1 = 1$  and iteratively increase this value, since the cost would be too high: in order to achieve an overall running time of  $\tilde{O}(\sqrt{n})$  we need to keep  $D_1 \approx D_2$ . Instead of estimating the probability that UlamIndic outputs 1 using Theorem 7, which is too costly, we thus design a gap test (see Proposition 3 below) that enables us to check if this probability is larger than  $D_2$  or smaller than  $(1 - \eta)D_2$  much more efficiently, in time  $\tilde{O}(\sqrt{n}/D_2)$ .

We now present the details of our quantum algorithm. Let us first present the gap test. This test relies on quantum amplitude amplification. Here is the precise statement of quantum amplitude amplification that we will use.

**Theorem 7** (Theorem 12 in [35]) *Let  $\mathcal{A}$  be a classical algorithm that runs in time  $T$ , outputs 1 with probability  $p$  and outputs 0 with probability  $1 - p$ , for some (unknown) value  $p \in [0, 1]$ . For any integer  $k \geq 1$ , there exists a quantum algorithm that runs in time  $\tilde{O}(kT)$  and outputs with probability at least  $8/\pi^2$  an estimate  $\tilde{p}$  such that*

$$|p - \tilde{p}| \leq 2\pi \frac{\sqrt{p(1-p)}}{k} + \frac{\pi^2}{k^2}.$$

The gap test is described in the following proposition.

**Proposition 3** *Let  $\mathcal{A}$  be a classical algorithm that runs in time  $T$ , outputs 1 with probability  $p$  and outputs 0 with probability  $1 - p$ , for some (unknown) value  $p$ . For any  $q \in (0, 1]$  and any constant  $\eta \in (0, 1]$ , there exists a quantum algorithm denoted  $QTest(\mathcal{A}, q, \eta)$  that runs in time  $\tilde{O}(T/\sqrt{q})$  and with probability at least  $1 - \text{poly}(n)$  outputs LARGE if  $p \geq q$  and SMALL if  $p \leq (1 - \eta)q$ .*

**Proof** Figure 12 describes our main quantum gap test. The complexity of this test is  $\tilde{O}(T/\sqrt{q})$ , from Theorem 7. We show below that its success probability is at least  $8/\pi^2$ . The success probability can then be increased to  $1 - 1/\text{poly}(n)$  by repeating the test  $\Theta(\log n)$  times and using majority voting.

Let us analyze the success probability of the quantum gap test of Figure 12. Note that Theorem 7 guarantees that with probability at least  $8/\pi^2$  the value  $\tilde{p}$  satisfies the following two inequalities:

$$\tilde{p} \leq p + 2\pi\eta \frac{\sqrt{p(1-p)}\sqrt{q}}{20} + \pi^2\eta^2 \frac{q}{400}, \tag{1}$$

$$\tilde{p} \geq p - 2\pi\eta \frac{\sqrt{p(1-p)}\sqrt{q}}{20} - \pi^2\eta^2 \frac{q}{400}. \tag{2}$$

Let us first consider the case  $p \leq (1 - \eta)q$ . In that case Inequality (1) implies

$$\tilde{p} \leq (1 - \eta)q + 2\pi\eta \frac{q}{20} + \pi^2\eta^2 \frac{q}{400} < (1 - \eta/2)q$$

since  $2\pi/20 + \pi^2/400 < 1/2$ . The probability that the algorithm outputs LARGE is thus at most  $1 - 8/\pi^2$ .

Now consider the case  $p \geq q$ . In that case Inequality (2) implies

$$\begin{aligned} \tilde{p} &\geq p - 2\pi\eta \frac{\sqrt{p}\sqrt{q}}{20} - \pi^2\eta^2 \frac{q}{400} \\ &= \sqrt{p} \left( \sqrt{p} - 2\pi\eta \frac{\sqrt{q}}{20} \right) - \pi^2\eta^2 \frac{q}{400} \\ &\geq q - 2\pi\eta \frac{q}{20} - \pi^2\eta^2 \frac{q}{400} > (1 - \eta/2)q. \end{aligned}$$

The probability that the algorithm outputs SMALL is thus at most  $1 - 8/\pi^2$ . □

We are now ready to give the proof of Theorem 6.

**Proof of Theorem 6** If  $\text{ud}(A, B) \geq \frac{1-\epsilon}{c}\sqrt{n}$ , where  $c$  is the constant from Proposition 2, then  $\text{ud}(A, B)$  can already be computed in  $\tilde{O}(\sqrt{n})$  time by the classical algorithm from [33]. We thus assume below that  $\text{ud}(A, B) \leq \frac{1-\epsilon}{c}\sqrt{n}$ . We also assume that  $\text{ud}(A, B) > 0$ , since otherwise the two strings are identical, which can be checked in  $\tilde{O}(\sqrt{n})$  time using Grover search.

Our quantum algorithm is described in Figure 13. The algorithm invokes  $\text{UlamIndic}(A, B, \delta, \frac{c}{1-\epsilon}t)$  with several values of  $t$  such that  $t \leq \frac{1-\epsilon}{c}\sqrt{n}$ . The analysis of the correctness done below will show that with probability at least  $1 - 1/\text{poly}(n)$ , all calls to  $\text{UlamIndic}$  performed by the algorithm satisfy the condition  $t' \geq c \cdot \text{ud}(A, B)$  required in the statement of Proposition 3 (in our case  $t' = \frac{c}{1-\epsilon}t$ ). Now assume that the condition is satisfied and write  $p_t$  the probability that  $\text{UlamIndic}(A, B, \delta, \frac{c}{1-\epsilon}t)$  outputs 1. Observe that Proposition 2 guarantees that

$$p_t \in \left[ (1 - \delta) \left( 1 - \frac{1}{\sqrt{n}} \right) \cdot \frac{\text{ud}(A, B)}{n}, (1 + \delta) \cdot \frac{\text{ud}(A, B)}{n} + \frac{\delta}{n} \right] \tag{3}$$

since we have  $\frac{c}{1-\epsilon}t \leq \sqrt{n}$ .

Then observe that for  $r = \left\lceil \frac{\log((1-\delta)(\sqrt{n}-1)/n)}{\log(1-\eta)} \right\rceil$ , the inequality

$$(1 - \delta) \left( 1 - \frac{1}{\sqrt{n}} \right) \cdot \frac{\text{ud}(A, B)}{n} \geq (1 - \eta)^r / \sqrt{n}$$

holds, since we are assuming that  $\text{ud}(A, B) \geq 1$ . From Proposition 3 combined with (3), this means that with probability at least  $1 - 1/\text{poly}(n)$ , the algorithm will never

**Fig. 13** Quantum algorithm computing a  $(1 + \epsilon)$ -approximation of the Ulam distance of two strings  $A$  and  $B$

```

1. Set  $\eta = \epsilon/3$  and  $\delta = \epsilon/3$  and  $r = \left\lceil \frac{\log((1-\delta)(\sqrt{n}-1)/n)}{\log(1-\eta)} \right\rceil$ .
2. For  $i$  from 1 to  $r$  do:
  2.1. Set  $t = (1 - \eta)^i \cdot \frac{1-\epsilon}{c} \sqrt{n}$  and  $q = t/n$ .
  2.2. Apply Algorithm QTest( $A, q, \eta$ ) with  $\mathcal{A} = \text{UlamIndic}(A, B, \delta, \frac{c}{1-\epsilon}t)$ .
  2.3. If the output is LARGE, then stop and return  $t$ .
3. Return ERROR.
    
```

reach Step 3 (and thus does not output ERROR). In the remaining of the proof, we assume that the algorithm stops before reaching Step 3. Let  $i^*$  denote the value of  $i$  during the last iteration of the loop of Step 2, and write  $t^* = (1 - \eta)^{i^*} \cdot \frac{1-\epsilon}{c} \sqrt{n}$  and  $q^* = t^*/n$ . Observe that the output of the algorithm is  $t^*$ .

Assume first that  $i^* = 1$ . The output is thus  $t^* = (1 - \eta) \frac{1-\epsilon}{c} \sqrt{n}$ . Observe that Algorithm QTest then necessarily outputted LARGE for  $i = 1$ . Proposition 3 guarantees that with probability at least  $1 - 1/\text{poly}(n)$  the inequality  $p_{t^*} \geq (1 - \eta)q^*$  holds (otherwise it would have outputted SMALL). This inequality combined with (3) implies that

$$(1 + \delta) \cdot \frac{\text{ud}(A, B)}{n} + \frac{\delta}{n} \geq (1 - \eta)q^*$$

and thus the output  $t^* = nq^*$  of the algorithm satisfies the following inequalities:

$$\begin{aligned}
 t^* &\leq ((1 + \delta) \cdot \text{ud}(A, B) + \delta)/(1 - \eta) \\
 &\leq (1 + 2\delta) \cdot \text{ud}(A, B)/(1 - \eta) \\
 &\leq (1 + 2\delta)(1 + 2\eta) \cdot \text{ud}(A, B) \\
 &\leq (1 + \epsilon) \cdot \text{ud}(A, B),
 \end{aligned}
 \tag{4}$$

where we used the inequality  $1/(1 - x) \leq 1 + 2x$ , which holds for any  $x \in [0, 1/2]$ . Also note that the inequality

$$t^* = (1 - \eta) \cdot \frac{1 - \epsilon}{c} \sqrt{n} \geq (1 - \epsilon) \cdot \text{ud}(A, B)$$

is trivially satisfied since we are assuming  $\text{ud}(A, B) \leq \frac{1-\epsilon}{c} \sqrt{n}$  and  $\eta < \epsilon$ . The output is thus a  $(1 + \epsilon)$ -approximation of  $\text{ud}(A, B)$ .

Assume now that  $i^* \geq 2$ . Observe that Algorithm QTest then necessarily outputted LARGE for  $i = i^*$  and SMALL for  $i = i^* - 1$ . Since it outputted LARGE for  $i = i^*$ , Inequalities (4) hold with probability at least  $1 - 1/\text{poly}(n)$ , from exactly the same argument as above. Since the output was SMALL for  $i = i^* - 1$ , Proposition 3 guarantees that with probability at least  $1 - 1/\text{poly}(n)$  the inequality  $p_{t^*/(1-\eta)} \leq q^*/(1 - \eta)$  holds (otherwise it would have outputted LARGE). Since  $t^* = nq^*$ , this

inequality combined with (3) implies:

$$\begin{aligned}
 t^* &\geq n(1-\eta)p_{t^*/(1-\eta)} \\
 &\geq (1-\eta)(1-\delta)\left(1-\frac{1}{\sqrt{n}}\right) \cdot \text{ud}(A, B) \\
 &= (1-\epsilon/3)(1-\epsilon/3)\left(1-\frac{1}{\sqrt{n}}\right) \cdot \text{ud}(A, B) \\
 &\geq (1-\epsilon) \cdot \text{ud}(A, B).
 \end{aligned}$$

The output of the algorithm is thus a  $(1+\epsilon)$ -approximation of  $\text{ud}(A, B)$  as well.

Since the complexity of applying Algorithm UlamIndic at Step  $i$  is  $\tilde{O}(\sqrt{(1-\eta)^i \sqrt{n}})$ , the overall complexity of the algorithm is

$$\tilde{O}\left(\sum_{i=1}^r \frac{1}{\sqrt{(1-\eta)^i \sqrt{n}}} \sqrt{(1-\eta)^i \sqrt{n}}\right) = \tilde{O}(\sqrt{n}),$$

as claimed.  $\square$

## 7 Lower Bounds

In this section we prove lower bounds for the problems considered in this paper.

We start by an easy lower bound for the longest common substring over a large alphabet.

**Theorem 8** *For any constant  $c \in (0, 1]$ , any quantum algorithm that computes with high probability a  $c$ -approximation of the longest common substring of two strings of length  $n$  over an alphabet of size  $2n$  requires  $\Omega(n^{2/3})$  time. This lower bound also holds for non-repetitive strings.*

**Proof** Consider the following version of the element distinctness problem: given a list  $L$  of  $m$  characters in an alphabet of size  $m$  such that either all the characters of  $L$  are distinct, or only one character occurs twice in  $L$  (i.e., the other  $m-2$  characters are distinct), decide which of the two cases holds. A  $\Omega(m^{2/3})$ -query lower bound is known for this problem in the quantum setting [42–44].

Let us take  $m = 2n$ . Construct a string  $A$  of length  $n$  by taking  $n$  elements from  $L$  uniformly at random, and construct another string  $B$  of length  $n$  using the remaining  $n$  elements from  $L$ . Observe that if all the characters in  $L$  are distinct then  $A$  and  $B$  are non-repetitive and have no common substring. On the other hand, if the characters in  $L$  are not all distinct, then with probability at least  $1/2$  the two strings  $A$  and  $B$  are non-repetitive and have a common substring of length 1. This gives a (randomized) reduction from element distinctness problem to our problem since a  $c$ -approximation of the longest common substring enables us to distinguish between the two cases.  $\square$

We now present a lower bound for LCS that holds even for binary strings.



**Theorem 9** For any constant  $c \in (0, 1]$ , any quantum algorithm that computes with high probability a  $c$ -approximation of longest common substring of two binary strings of length  $n$  requires  $\Omega(n^{2/3})$  time.

**Proof** The main idea is simple: by dividing the input strings into blocks of size  $\Theta(\log n)$ , we can reduce the case with alphabet of size  $2n$  to the case of binary alphabets, and thus use the lower bound of Theorem 8 with only a logarithmic overhead.

We now give more details of the reduction. Let  $A$  and  $B$  denote strings of length  $n$  over an alphabet of size  $2n$ . Each character of the alphabet is encoded by a random binary string of length  $s$  (i.e., a binary string of length  $s$  such that each bit is 1 with probability  $1/2$ ). Using easy arguments from probability theory (see, e.g., Section 2 of [45]), for any constant  $\alpha > 0$  we can guarantee that the following property holds with probability at least  $9/10$  if we take  $s \geq d_\alpha \cdot \log n$  for some constant  $d_\alpha$  that depends only on  $\alpha$ : the length of the longest common substring of the encodings of any two distinct characters is at most  $\alpha s$ . Let us choose  $\alpha = c/3$  and consider  $s = \lceil d_\alpha \cdot \log n \rceil$ . Below we assume that the above property holds (with happens with probability at least  $9/10$ ).

If the longest common substring of  $A$  and  $B$  has length zero then the longest common substring of their binary versions has length at most  $2\alpha s = 2cs/3 < cs$ . If the longest common substring of  $A$  and  $B$  has length at least one, on the other hand, then the longest common substring of their binary versions has length at least  $s$ . Thus a  $c$ -approximation of the longest common substring of the binary versions enables us to distinguish, with high probability, between the two cases.  $\square$

We now give our lower bounds for LPS and for the computation of the Ulam distance.

**Theorem 10** For any constant  $c \in (0, 1]$ , any quantum algorithm that computes with high probability a  $c$ -approximation of the longest palindrome of a binary string of length  $n$  requires  $\Omega(\sqrt{n})$  time.

**Proof** Let  $m \geq 3$  be an integer. Let  $S_1 \subseteq \{0, 1\}^m$  denote the set of all  $m$ -bit strings of Hamming weight one in which the first and last characters are both zero, and let  $S_0 \subseteq \{0, 1\}^m$  be the set containing only the all-zero string. Distinguishing between strings in  $S_0$  and  $S_1$  requires  $\Omega(\sqrt{m})$  queries in the quantum setting (see, e.g., [46]).

Let us write  $k = \lceil 3/c \rceil$ . Given a string  $x \in S_0 \cup S_1$  and any  $r \in \{1, \dots, k\}$ , let  $x^r$  be the binary string of length  $km$  obtained from  $x$  by replacing each 0 in  $x$  by  $0^k$  and each 1 in  $x$  by  $1^r 0^{k-r}$ .

Take  $n = k^2 m$ . We now consider the string  $A \in \{0, 1\}^n$  obtained by concatenating the strings  $x^1, x^2, \dots, x^k$ . Each  $x^i$ , for  $i \in \{1, \dots, k\}$ , is called a block of  $A$ . Observe that if  $x \in S_0$  (i.e., if  $x$  is the all-zero string), then  $A$  is also the all-zero strings. In this case the length of the LPS of  $A$  is  $n$ . On the other hand, if  $x \in S_1$  then no palindrome of  $A$  can include two full blocks (since the numbers of repeated 1s in distinct blocks do not match), and thus the length of the LPS of  $A$  is at most  $km + 2(km - 1) \leq 3km < cn$ . Thus computing a  $c$ -approximation of the LPS of  $A$  enables us to distinguishing between strings in  $S_0$  and  $S_1$ . This gives the lower bound  $\Omega(\sqrt{m}) = \Omega(\sqrt{n})$  on the complexity of computing a  $c$ -approximation of the LPS.  $\square$

**Theorem 11** For any  $\epsilon \in [0, 1)$ , any quantum algorithm that computes with high probability a  $(1 + \epsilon)$ -approximation of the Ulam distance of two non-repetitive strings requires  $\Omega(\sqrt{n})$  time.

**Proof** Let us consider the alphabet  $\{1, 2, \dots, n\}$ . Let  $A$  be the all-increasing string, i.e.,  $A = 123 \cdots n$ . Let  $B$  either the all-increasing string or the string obtained by permuting the  $\ell$ -th position and the  $(\ell + 1)$ -position of the all-increasing string for some unknown  $\ell \in \{1, \dots, n - 1\}$ . Note that in former case the Ulam distance of  $A$  and  $B$  is zero, while in the second case the Ulam distance is two. Computing a  $(1 + \epsilon)$ -factor approximation of the Ulam distance of  $A$  and  $B$  thus requires distinguishing between the two cases, which requires  $\Omega(\sqrt{n})$  queries from the lower bound on Grover search [46].  $\square$

**Acknowledgements** The authors are grateful to Michael Saks and C. Seshadhri for helpful correspondence.

**Funding** FLG was supported by the JSPS KAKENHI Grants Nos. JP16H01705, JP19H04066, JP20H00579, JP20H04139, JP20H05966, JP21H04879 and by the MEXT Quantum Leap Flagship Program (MEXT Q-LEAP) Grant No. JPMXS0120319794. SS was supported in part by an Adobe Research Award and a Google Research Gift.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Masek, W.J., Paterson, M.S.: A faster algorithm computing string edit distances. *J. Comput. Syst. Sci.* **20**(1), 18–31 (1980)
2. Bar-Yossef, Z., Jayram, T., Krauthgamer, R., Kumar, R.: Approximating edit distance efficiently. In: *Proceedings of the 45th Annual IEEE Symposium on the Foundations of Computer Science*, pp. 550–559 (2004)
3. Batu, T., Ergun, F., Sahinalp, C.: Oblivious string embeddings and edit distance approximations. In: *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 792–801 (2006)
4. Andoni, A., Krauthgamer, R., Onak, K.: Polylogarithmic approximation for edit distance and the asymmetric query complexity. In: *Proceedings of the 51st Annual IEEE Symposium on the Foundations of Computer Science*, pp. 377–386 (2010)
5. Boroujeni, M., Ehsani, S., Ghodsi, M., HajiAghayi, M., Seddighin, S.: Approximating edit distance in truly subquadratic time: quantum and MapReduce. In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1170–1189 (2018)
6. Chakraborty, D., Das, D., Goldenberg, E., Koucky, M., Saks, M.: Approximating edit distance within constant factor in truly sub-quadratic time. In: *Proceedings of the 59th Annual IEEE Symposium on the Foundations of Computer Science*, pp. 979–990 (2018)
7. Farach, M.: Optimal suffix tree construction with large alphabets. In: *Proceedings of the 38th Annual IEEE Symposium on Foundations of Computer Science*, pp. 137–143 (1997)
8. Amir, A., Charalampopoulos, P., Pissis, S.P., Radoszewski, J.: Longest common substring made fully dynamic. In: *Proceedings of the 27th Annual European Symposium on Algorithms*, pp. 6–1617 (2018)
9. Kociumaka, T., Starikovskaya, T., Vildhøj, H.W.: Sublinear space algorithms for the longest common substring problem. In: *Proceedings of the 22th Annual European Symposium on Algorithms*, pp. 605–617 (2014)

10. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, Cambridge (2009)
11. Backurs, A., Indyk, P.: Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). *SIAM J. Comput.* **47**(3), 1087–1097 (2018)
12. Abboud, A., Backurs, A., Williams, V.V.: Tight hardness results for LCS and other sequence similarity measures. In: Proceedings of the 56th Annual IEEE Symposium on the Foundations of Computer Science, pp. 59–78 (2015)
13. Ambainis, A.: Quantum walk algorithm for element distinctness. *SIAM J. Comput.* **37**(1), 210–239 (2007)
14. Buhrman, H., Špalek, R.: Quantum verification of matrix products. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 880–889 (2006)
15. Grover, L.K.: A fast quantum mechanical algorithm for database search. In: Proceedings of the 28th Annual ACM Symposium on the Theory of Computing, pp. 212–219 (1996)
16. Harrow, A.W., A.H., Lloyd, S.: Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **103**, 150502 (2006)
17. Magniez, F., Santha, M., Szegedy, M.: Quantum algorithms for the triangle problem. *SIAM J. Comput.* **37**(2), 413–424 (2007)
18. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Comput.* **26**(5), 1484–1509 (1997)
19. Ramesh, H., Vinay, V.: String matching in  $\tilde{O}(\sqrt{n} + \sqrt{m})$  quantum time. *J. Discrete Algorithms* **1**(1), 103–110 (2003)
20. Ambainis, A., Montanaro, A.: Quantum algorithms for search with wildcards and combinatorial group testing. *Quantum Inf. Comput.* **14**(5–6), 439–453 (2014)
21. Cleve, R., Iwama, K., Le Gall, F., Nishimura, H., Tani, S., Teruyama, J., Yamashita, S.: Reconstructing strings from substrings with quantum queries. In: Proceedings of the 13th Scandinavian Symposium and Workshops on Algorithm Theory, pp. 388–397 (2012)
22. Kaye, P., Laflamme, R., Mosca, M., et al.: An Introduction to Quantum Computing. Oxford University Press, Oxford (2007)
23. Andoni, A., Nosatzki, N.S.: Edit distance in near-linear time: it’s a constant factor. In: Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (2020)
24. Koucký, M., Saks, M.: Constant factor approximations to edit distance on far input pairs in nearly linear time. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pp. 699–712 (2020)
25. Brakensiek, J., Rubinfeld, A.: Constant-factor approximation of near-linear edit distance in near-linear time. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pp. 685–698 (2020)
26. Ambainis, A., Balodis, K., Iraids, J., Kokainis, M., Prūsis, K., Vihrovs, J.: Quantum speedups for exponential-time dynamic programming algorithms. In: Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1783–1793 (2019)
27. Rubinfeld, A.: Quantum DNA sequencing and the ultimate hardness hypothesis. <https://theorydish.blog/2019/12/09/quantum-dna-sequencing-the-ultimate-hardness-hypothesis/> (2019)
28. Buhrman, H., Patro, S., Speelman, F.: A framework of quantum strong exponential-time hypotheses. In: Proceedings of the 38th International Symposium on Theoretical Aspects of Computer Science. LIPIcs, vol. 187 (article 19), pp. 1–19 (2021)
29. Aaronson, S., Chia, N., Lin, H., Wang, C., Zhang, R.: On the quantum complexity of closest pair and related problems. In: Proceedings of the 35th Computational Complexity Conference. LIPIcs, vol. 169 (article 16), pp. 1–43 (2020)
30. Fredman, M.L.: On computing the length of longest increasing subsequences. *Discrete Math.* **11**(1), 29–35 (1975)
31. Ramanathan, P.: Tight  $\Omega(n \lg n)$  lower bound for finding a longest increasing subsequence. *Int. J. Comput. Math.* **65**(3–4), 161–164 (1997)
32. Andoni, A., Nguyen, H.L.: Near-optimal sublinear time algorithms for Ulam distance. In: Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 76–86 (2010)
33. Naumovitz, T., Saks, M.E., Seshadhri, C.: Accurate and nearly optimal sublinear approximations to Ulam distance. In: Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 2012–2031 (2017). Revised version available at <https://users.soe.ucsc.edu/sesh/publication.html>

34. Akmal, S., Jin, C.: Near-optimal quantum algorithms for string problems. In: Naor, J., Buchbinder, N. (eds.) Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms (SODA 2022), pp. 2791–2832. SIAM (2022). <https://doi.org/10.1137/1.9781611977073.109>
35. Brassard, G., Hoyer, P., Mosca, M., Tapp, A.: Quantum amplitude amplification and estimation. *Contemp. Math.* **305**, 53–74 (2002)
36. Charikar, M., Krauthgamer, R.: Embedding the Ulam metric into  $\ell_1$ . *Theory Comput.* **2**(11), 207–224 (2006)
37. Mitzenmacher, M., Seddighin, S.: Dynamic algorithms for LIS and distance to monotonicity. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pp. 671–684 (2020)
38. Magniez, F., Nayak, A., Roland, J., Santha, M.: Search via quantum walk. *SIAM J. Comput.* **40**(1), 142–164 (2011)
39. Saks, M.E., Seshadhri, C.: Estimating the longest increasing sequence in polylogarithmic time. In: Proceedings of the 51th Annual IEEE Symposium on Foundations of Computer Science, pp. 458–467 (2010)
40. Szegedy, M.: Quantum speed-up of markov chain based algorithms. In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, pp. 32–41 (2004)
41. Dubhashi, D.P., Panconesi, A.: Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge University Press, Cambridge (2009)
42. Aaronson, S., Shi, Y.: Quantum lower bounds for the collision and the element distinctness problems. *J. ACM* **51**(4), 595–605 (2004)
43. Ambainis, A.: Polynomial degree and lower bounds in quantum complexity: collision and element distinctness with small range. *Theory Comput.* **1**(1), 37–46 (2005)
44. Kutin, S.: Quantum lower bound for the collision problem with small range. *Theory Comput.* **1**(1), 29–36 (2005)
45. Arratia, R., Waterman, M.S.: An Erdős-Rényi law with shifts. *Adv. Math.* **55**(1), 13–23 (1985)
46. Bennett, C.H., Bernstein, E., Brassard, G., Vazirani, U.V.: Strengths and weaknesses of quantum computing. *SIAM J. Comput.* **26**(5), 1510–1523 (1997)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.