



# Faster Approximate Elastic-Degenerate String Matching – Part B

Paweł Gawrychowski 

Institute of Computer Science, University of Wrocław, Poland

Adam Górkiewicz 

Institute of Computer Science, University of Wrocław, Poland

Pola Marciniak 

Institute of Computer Science, University of Wrocław, Poland

Solon P. Pissis 

CWI, Amsterdam, The Netherlands

Vrije Universiteit, Amsterdam, The Netherlands

Karol Pokorski 

Institute of Computer Science, University of Wrocław, Poland

---

## Abstract

We revisit the complexity of approximate pattern matching in an elastic-degenerate string. Such a string is a sequence of  $n$  finite sets of strings of total length  $N$ , and compactly describes a collection of strings obtained by first choosing exactly one string in every set, and then concatenating them together. This is motivated by the need of storing a collection of highly similar DNA sequences.

The basic algorithmic question on elastic-degenerate strings is pattern matching: given such an elastic-degenerate string and a standard pattern of length  $m$ , check if the pattern occurs in one of the strings in the described collection. Bernardini et al. [SICOMP 2022] showed how to leverage fast matrix multiplication to obtain an  $\tilde{O}(nm^{\omega-1}) + O(N)$ -time complexity for this problem, where  $\omega$  is the matrix multiplication exponent. However, from the point of view of possible applications, it is more desirable to work with approximate pattern matching, where we seek approximate occurrences of the pattern. This generalization has been considered in a few papers already, but the best result so far for occurrences with  $k$  mismatches, where  $k$  is a constant, is the  $\tilde{O}(nm^2 + N)$ -time algorithm presented in Part A [CPM 2025]. This brings the question whether increasing the dependency on  $m$  from  $m^{\omega-1}$  to quadratic is necessary when moving from  $k = 0$  to larger (but still constant)  $k$ .

We design an  $\tilde{O}(nm^{1.5} + N)$ -time algorithm for pattern matching with  $k$  mismatches in an elastic-degenerate string, for any constant  $k$ . To obtain this time bound, we leverage the structural characterization of occurrences with  $k$  mismatches of Charalampopoulos, Kociumaka, and Wellnitz [FOCS 2020] together with fast Fourier transform. We need to work with multiple patterns at the same time, instead of a single pattern, which requires refining the original characterization. This might be of independent interest.

**2012 ACM Subject Classification** Theory of computation → Pattern matching

**Keywords and phrases** ED string, approximate pattern matching, Hamming distance,  $k$  mismatches

**Digital Object Identifier** 10.4230/LIPIcs.CPM.2025.29

**Funding** *Paweł Gawrychowski*: Partially supported by the Polish National Science Centre grant number 2023/51/B/ST6/01505.

*Adam Górkiewicz*: Partially supported by the Polish National Science Centre grant number 2023/51/B/ST6/01505.

*Pola Marciniak*: Partially supported by the Polish National Science Centre grant number 2023/51/B/ST6/01505.

*Solon P. Pissis*: Partially supported by the PANGAIA and ALPACA projects that have received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No 872539 and 956229, respectively.



© Paweł Gawrychowski, Adam Górkiewicz, Pola Marciniak, Solon P. Pissis, and Karol Pokorski; licensed under Creative Commons License CC-BY 4.0

36th Annual Symposium on Combinatorial Pattern Matching (CPM 2025).

Editors: Paola Bonizzoni and Veli Mäkinen; Article No. 29; pp. 29:1–29:21

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

An *elastic-degenerate string* (ED-string, in short)  $\tilde{T}$  is a sequence  $\tilde{T} = \tilde{T}[0]\tilde{T}[1]\dots\tilde{T}[n-1]$  of  $n$  finite sets, where  $\tilde{T}[i]$  is a subset of  $\Sigma^*$  and  $\Sigma$  is an ordered finite alphabet. The *length* of  $\tilde{T}$  is defined as the length  $n = |\tilde{T}|$  of the associated sequence. The *size* of  $\tilde{T}$  is defined as  $N = N_\epsilon + \sum_{i=0}^{n-1} \sum_{S \in \tilde{T}[i]} |S|$ , where  $N_\epsilon$  is the total number of empty strings in  $\tilde{T}$ . The *cardinality* of  $\tilde{T}$  is defined as  $c = \sum_{i=0}^{n-1} |\tilde{T}[i]|$ . Every ED-string represents a collection of strings, each of generally different length. We formalize this intuition as follows. The *language*  $\mathcal{L}(\tilde{T})$  generated by  $\tilde{T}$  is defined as  $\mathcal{L}(\tilde{T}) = \{S_0S_1\dots S_{n-1} : S_i \in \tilde{T}[i], \text{ for all } i \in [0, n-1]\}$ .

The main motivation behind introducing ED-strings [19] was to encode a collection of highly similar DNA sequences in a compact form. Consider a multiple sequence alignment (MSA) of such a collection (see below). Maximal conserved substrings (conserved columns of the MSA) form singleton sets of the ED-string and the non-conserved ones form sets that list the corresponding variants. Moreover, the language of the ED-string consists of (at least) the underlying sequences of the MSA. Under the assumption that these underlying sequences are highly similar, the size of the ED-string is substantially smaller than the total size of the collection. ED-strings have been used in several applications: indexing a pangenome [8], on-line string matching in a pangenome [10], and pairwise comparison of pangenomes [15].

MSA	ED-string $\tilde{T}$
GTA <b>ACTGCCGT</b> --TG	$\left\{ \begin{matrix} \text{GTA} \\ \text{GTA} \\ \text{GTATCTCCC} \end{matrix} \right\} \cdot \left\{ \begin{matrix} \text{A} \\ \text{A} \\ \text{T} \end{matrix} \right\} \cdot \left\{ \begin{matrix} \text{CT} \\ \text{CT} \\ \text{C} \end{matrix} \right\} \cdot \left\{ \begin{matrix} \text{G} \\ \text{G} \\ \text{C} \end{matrix} \right\} \cdot \left\{ \begin{matrix} \text{CC} \\ \text{CC} \\ \text{C} \end{matrix} \right\} \cdot \left\{ \begin{matrix} \text{GT} \\ \text{GTAA} \\ \epsilon \end{matrix} \right\} \cdot \left\{ \begin{matrix} \text{TG} \\ \text{TG} \\ \text{C} \end{matrix} \right\}$
GTA <b>ACTGCCGTAA</b> TG	
GTAT <b>CTCCC</b> ----TG	

ED-strings are also interesting as a simplified model for string matching on node-labeled graphs [3]. The ED-string  $\tilde{T}$  can be viewed as a graph of  $n$  layers [21], where the  $c$  nodes are strings from  $\Sigma^*$ , such that from layer  $i$  to layer  $i+1$  all possible edges are present and the nodes in layer  $i$  are adjacent only to the nodes in layer  $i+1$ . As a simplified model, ED-strings offer important advantages, such as *on-line* (left to right) string matching algorithms whose running times have a *linear dependency on  $N$*  [2, 5, 17]. This linear dependency on  $N$  (without any multiplicative polylogarithmic factors) is highly desirable in applications because, nowadays, it is typical to encode a very large number of genomes (e.g., millions of SARS-CoV-2 genomes<sup>1</sup>) in a single representation resulting in huge  $N$  values.

In this work, we focus on the string matching (or pattern matching) task. In the *elastic-degenerate string matching* (EDSM) problem, we are given a string  $P$  of length  $m$  (known as the *pattern*) and an ED-string  $\tilde{T}$  (known as the *text*), and we are asked to find the occurrences of  $P$  in  $\mathcal{L}(\tilde{T})$ . Grossi et al. showed that EDSM can be solved in  $\mathcal{O}(nm^2 + N)$  time using combinatorial pattern matching tools [17]. Aoyama et al. improved this to  $\tilde{\mathcal{O}}(nm^{1.5}) + \mathcal{O}(N)$  time by employing fast Fourier transform [2]. Finally, Bernardini et al. improved it to  $\tilde{\mathcal{O}}(nm^{w-1}) + \mathcal{O}(N)$  time [5] by employing fast matrix multiplication, where  $w < 2.37134$  [1] is the matrix multiplication exponent. The authors of [1] also showed a conditional lower bound for combinatorial algorithms<sup>2</sup> for EDSM stating that EDSM cannot be solved in  $\mathcal{O}(nm^{1.5-\epsilon} + N)$  time, for any constant  $\epsilon > 0$ .

In the approximate counterpart of EDSM, we are also given an integer  $k > 0$ , and we are asked to find  $k$ -approximate occurrences of  $P$  in  $\mathcal{L}(\tilde{T})$ ; namely, the occurrences that are at Hamming or edit distance at most  $k$  from  $P$ . For Hamming distance, we call the problem

<sup>1</sup> <https://gisaid.org>

<sup>2</sup> The term “combinatorial” is not formally well-defined.

EDSM WITH  $k$  MISMATCHES (see below); and for edit distance, EDSM WITH  $k$  ERRORS. The approximate EDSM problem was introduced by Bernardini et al. who showed a simple  $\mathcal{O}(kmc + kN)$ -time algorithm for EDSM WITH  $k$  MISMATCHES and an  $\mathcal{O}(k^2mc + kN)$ -time algorithm for EDSM WITH  $k$  ERRORS using combinatorial pattern matching tools [6]. In Part A, the dependency on  $k$  for both the Hamming and the edit distance metrics is improved, obtaining an  $\tilde{\mathcal{O}}(k^{2/3}mc + \sqrt{k}N)$ -time algorithm for EDSM WITH  $k$  MISMATCHES and an  $\tilde{\mathcal{O}}(kmc + kN)$ -time algorithm for EDSM WITH  $k$  ERRORS [22].

Pattern $P$	ED-string $\tilde{T}$	Matches
CTCTG	$\left\{ \begin{smallmatrix} \text{G} & \text{T} & \text{A} \\ \text{G} & \text{T} & \text{A} \end{smallmatrix} \right\} \cdot \left\{ \begin{smallmatrix} \text{A} \\ \text{T} \end{smallmatrix} \right\} \cdot \left\{ \begin{smallmatrix} \text{C} & \text{T} \\ \text{C} & \text{T} \end{smallmatrix} \right\} \cdot \left\{ \begin{smallmatrix} \text{G} \\ \text{C} \end{smallmatrix} \right\} \cdot \left\{ \begin{smallmatrix} \text{C} & \text{C} \\ \text{C} & \text{C} \end{smallmatrix} \right\} \cdot \left\{ \begin{smallmatrix} \text{GT} \\ \text{GTAA} \\ \epsilon \end{smallmatrix} \right\} \cdot \left\{ \begin{smallmatrix} \text{TG} \\ \text{TG} \end{smallmatrix} \right\}$	$\delta_H(\text{CTCTG}, \text{ATCTG}) = 1$ $\delta_H(\text{CTCTG}, \text{CCCTG}) = 1$

Unfortunately, the cardinality  $c$  of  $\tilde{T}$  in the above complexities is bounded only by  $N$ , so even for  $k = 1$ , the existing algorithms run in  $\Omega(mN)$  time in the worst case. Bernardini et al. [4] showed many algorithms for approximate EDSM for  $k = 1$  working in  $\tilde{\mathcal{O}}(nm^2 + N)$  time or in  $\mathcal{O}(nm^3 + N)$  time for both the Hamming and the edit distance metrics. In Part A, the results for  $k = 1$  (for both metrics) are improved to  $\mathcal{O}(nm^2 + N)$  time and extended to work for  $k > 1$  (for both metrics) in  $\tilde{\mathcal{O}}(nm^2 + N)$  time, for any constant  $k > 1$  [22].

Time complexity	Remarks	Reference
$\mathcal{O}(cm + N)$	$k = \mathcal{O}(1)$	[6]
$\mathcal{O}(nm^3 + N)$	$k = 1$	[4]
$\mathcal{O}(nm^2 + N \log m)$	$k = 1$	[4]
$\mathcal{O}(nm^2 + N)$	$k = 1$	Part A, [22]
$\tilde{\mathcal{O}}(nm^{1.5}) + \mathcal{O}(N)$	$k = 1$	<b>Theorem 1</b>
$\tilde{\mathcal{O}}(nm^2 + N)$	$k = \mathcal{O}(1)$	Part A, [22]
$\tilde{\mathcal{O}}(nm^{1.5} + N)$	$k = \mathcal{O}(1)$	<b>Theorem 2</b>

In this work, we consider the EDSM WITH  $k$  MISMATCHES problem with constant  $k \geq 1$ , and observe that all the existing algorithms have at best a *quadratic dependency on  $m$* , the length of the pattern, for this problem. This is in stark contrast to the case of  $k = 0$ , and brings the question of whether non-combinatorial methods could be employed to solve EDSM WITH  $k$  MISMATCHES for any constant  $k \geq 1$ , similar to EDSM ( $k = 0$ ) [2, 5].

► **Theorem 1.** *Given a pattern  $P$  of length  $m$  and an ED-string  $\tilde{T}$  of length  $n$  and size  $N$ , EDSM WITH  $k$  MISMATCHES, for  $k = 1$ , can be solved in  $\mathcal{O}(nm^{1.5} \text{polylog } m + N)$  time.*

► **Theorem 2.** *Given a pattern  $P$  of length  $m$  and an ED-string  $\tilde{T}$  of length  $n$  and size  $N$ , EDSM WITH  $k$  MISMATCHES, for any constant  $k \geq 1$ , can be solved in  $\mathcal{O}((nm^{1.5} + N) \text{polylog } m)$  time.*

**Other Approaches.** The key ingredient of [4] and [22] to achieve the  $m^2$  term for  $k = 1$  is a new counterpart of the  $k$ -errata tree [11], where the copied nodes of the input trie are explicitly inserted into the tree. This counterpart is an actual trie, and hence it allows to apply standard tree traversal algorithms. Since for  $k = 1$ , the constructed trie for  $\tilde{T}[i]$  and the suffixes of  $P$  has  $\mathcal{O}(m \log(N + m))$  nodes originating from  $P$ , bitwise  $\mathcal{O}(m / \log(N + m))$ -time operations per such node result in the desired complexity. The main tool in [22] for extending to a constant  $k > 1$  is also  $k$ -errata trees; however, the authors of [22] manage to apply  $k$ -errata trees as a black-box. We stress that those algorithms are *combinatorial* (they do not use fast Fourier transform or fast matrix multiplication) and work also for *edit distance*.

**Our Approach.** As in the previous works on elastic-degenerate string matching, we work with the so-called active prefixes extension problem. In this problem, we are given a text of length  $m$ , an input bitvector of length  $m$ , and a collection of patterns of total length  $N$ . The goal is to produce an output bitvector, also of length  $m$ . Informally, whenever there is an occurrence of some pattern starting at position  $i$  and ending at position  $j$ , we can propagate a one from the input bitvector at position  $i - 1$  to a one in the output bitvector at position  $j$ . For approximate pattern matching, we have  $k + 1$  input and output bitvectors, corresponding to matching prefixes with different (i.e., the corresponding) number of mismatches.

The previous solutions can be seen as propagating the information from every  $i - 1$  to every  $j$  explicitly. This, of course, cannot achieve time complexity better than quadratic in  $m$ . Instead, we leverage the following high-level idea: if a given pattern occurs very few times in the text, then we can afford to iterate through all of its occurrences and propagate the corresponding information. Otherwise, its occurrences are probably somewhat structured. More concretely, exact occurrences of a pattern of length  $\ell$  in a text of length  $1.5\ell$  form a single arithmetic progression. This has been extended by Bringmann, Künnemann, and Wellnitz [7] to occurrences with  $k$  mismatches, and further refined (and made effectively computable) by Charalampopoulos, Kociumaka, and Wellnitz [9]: either there are few occurrences, or they can be represented by a few arithmetic progressions (where few means polynomial in  $k$ ). Further, a representation of all the occurrences can be computed effectively.

To implement this high-level idea, we first apply some relatively standard preliminary steps that allow us to handle short patterns effectively with  $k$ -errata trees. Further, we show how to reduce a given instance to multiple instances in which the pattern and the text are roughly of the same length. Then, we handle patterns with only a few occurrences naively. For the remaining patterns, we obtain a compact representation (as a few arithmetic progressions) of their occurrences. We cannot afford to process each progression separately, but we observe that, because we have restricted the length of the text, their differences are in fact equal to the same  $q$  for every remaining pattern. Now, if  $q$  is somewhat large (with the exact threshold to be chosen at the very end to balance the complexities), we can afford to process every occurrence of every pattern naively. Otherwise, we would like to work with every remainder modulo  $q$  separately, leveraging fast Fourier transform to process all progressions starting at the positions with that remainder together. As a very simplified example, if the progressions were the same for all the patterns, we only need to compute the sumset of the set of starting positions with a one in the input bitvector (restricted to positions with a specific remainder modulo  $q$ ) with the set of lengths of the patterns. This can be indeed done with a single fast Fourier transform. However, the structural characterization of Charalampopoulos et al. [9] only says that, for every pattern, we have  $\mathcal{O}(k^2)$  arithmetic progressions with the same period  $q$ . However, the progressions are possibly quite different for different patterns. Our new insight is that, in fact, we can group the progressions into only a few classes, more specifically  $\mathcal{O}(k^2)$  irrespectively of the number of patterns, and then process each class together. This requires looking more carefully at the structural characterization of Charalampopoulos et al. [9], and might be of independent interest.

**Structure of the Paper.** In Section 2, we provide some preliminaries and problem definitions. In Section 3, we discuss how the EDSM WITH  $k$  MISMATCHES problem can be solved via the APE WITH  $k$  MISMATCHES problem, the auxiliary problem used also in previous solutions [2, 4, 5] and in Part A [22]. In Section 4, we present our algorithms for three different cases: very short in Section 4.1; short in Section 4.2; and, finally, long in Section 4.3, which is the most interesting case. We conclude with balancing the thresholds in Section 4.4.

**Computational Model.** We assume the standard Word RAM model with words consisting of  $\Omega(\log n)$  bits, where  $n$  is the size of the input. Basic operations on such words, such as indirect addressing and arithmetic operations, are thus assumed to take constant time.

## 2 Preliminaries

**Strings.** Let  $\Sigma$  be a finite ordered alphabet of size  $|\Sigma| = \sigma$ . We will usually assume that  $\Sigma = \{1, 2, \dots, \text{poly}(n)\}$ , where  $n$  is the size of the input, which is called the *polynomial alphabet*. The elements of  $\Sigma$  are called *characters*. A sequence of characters from  $\Sigma$ ,  $X[0]X[1] \dots X[n-1]$ , is called a (classic) *string*  $X$ . We call  $n$  the *length* of  $X$ , and denote it by  $|X|$ . The empty string is denoted by  $\varepsilon$ . By  $X[i..j] = X[i]X[i+1] \dots X[j]$ , we denote a *fragment* of  $X$  (starting at position  $i$  and ending at position  $j$ ), which equals  $\varepsilon$  when  $i > j$ . Fragments of the form  $X[..j] := X[0..j]$  and  $X[i..] := X[i..n-1]$  are called *prefixes* and *suffixes* of  $X$ , respectively. A fragment of  $X$  (or its prefix/suffix) is called *proper*, if it is not equal to  $X$ . Strings that are fragments of  $X$  (for some  $i$  and  $j$ ) are called *substrings* of  $X$ . We also write  $X[i..j)$  to denote  $X[i..j-1]$ . By  $XY$ , we denote the *concatenation* of strings  $X$  and  $Y$ , i.e., the string  $X[0] \dots X[|X|-1]Y[0] \dots Y[|Y|-1]$ . String  $X$  is a *cyclic shift* of  $Y$  when  $X = X_1X_2$  and  $Y = X_2X_1$ , and then we call  $X$  and  $Y$  *cyclically equivalent*. We say that  $X$  has an *occurrence* in  $Y$  (at position  $t$ ), if  $Y = AXB$  for some strings  $A$  and  $B$  such that  $|A| = t$ . Finally,  $X^r$  is the *reversal* of  $X$ , i.e., the string  $X[n-1]X[n-2] \dots X[0]$ .

**Elastic-Degenerate Strings.** We study the following extensions of classic strings.

A *symbol* (over alphabet  $\Sigma$ ) is an unordered subset of (classic) strings from  $\Sigma^*$ , different from  $\{\varepsilon\}$  and  $\emptyset$ . Note that symbols may contain  $\varepsilon$  but not as their only element. The *size* of a symbol is the total length of all strings in the symbol (with the additional assumption that the empty string is counted as if it had length 1). The *Cartesian concatenation* of two symbols  $X$  and  $Y$  is defined as  $X \otimes Y := \{xy \mid x \in X, y \in Y\}$ .

An *elastic-degenerate string* (or *ED-string*, in short)  $\tilde{X} = \tilde{X}[0]\tilde{X}[1] \dots \tilde{X}[n-1]$  (over alphabet  $\Sigma$ ) is a sequence of symbols (over  $\Sigma$ ). We use  $|\tilde{X}|$  to denote the *length* of  $\tilde{X}$ , i.e., the length of the associated sequence (the number of its symbols). The *size* of  $\tilde{X}$  is the sum of the sizes of the symbols in  $\tilde{X}$ . As for classic strings, we denote a *fragment* of  $\tilde{X}$  by  $\tilde{X}[i..j] = \tilde{X}[i]\tilde{X}[i+1] \dots \tilde{X}[j]$ . We similarly denote *prefixes* and *suffixes* of  $\tilde{X}$ . The *language* of  $\tilde{X}$  is  $\mathcal{L}(\tilde{X}) = \tilde{X}[0] \otimes \tilde{X}[1] \otimes \dots \otimes \tilde{X}[n-1]$ .

Given a (classic) string  $P$  and an ED-string  $\tilde{T}$ , we say that  $P$  *matches* the fragment  $\tilde{T}[i..j]$  (or that an *occurrence* of  $P$  starts at position  $i$  and ends at position  $j$  of  $\tilde{T}$ ), if  $i = j$  and  $P$  is a fragment of at least one of the strings of  $\tilde{T}[i]$  (the whole pattern is fully contained in one of the symbols), or if  $i < j$  and there is a sequence of strings  $(P_i, P_{i+1}, \dots, P_j)$ , such that:  $P = P_i P_{i+1} \dots P_j$ ;  $P_i$  is a suffix of one of the strings of  $\tilde{T}[i]$ ,  $P_k \in \tilde{T}[k]$ , for all  $i < k < j$ ; and  $P_j$  is a prefix of one of the strings of  $\tilde{T}[j]$  ( $P$  uses parts of at least two symbols).

**Hamming Distance.** Given two (classic) strings  $X$  and  $Y$  of the same length over alphabet  $\Sigma$ , their *Hamming distance*  $\delta_H(X, Y)$  is defined as the number of *mismatches* (i.e., the positions  $i$  such that  $X[i] \neq Y[i]$ ). We use  $\text{Mis}(X, Y)$  to denote the set of mismatches.

Given two (classic) strings  $X$  and  $Y$  over alphabet  $\Sigma$ , we say that  $X$  is an *approximate fragment* (with at most  $k$  mismatches) of  $Y$  if there is a string  $X'$  with  $\delta_H(X, X') \leq k$ , such that  $X'$  is a substring of  $Y$ . We similarly define *approximate prefixes* and *approximate suffixes*. We write  $\text{Occ}_k^H(X, Y)$  to denote the set of all  $k$ -mismatch approximate occurrences of  $X$  in  $Y$ , i.e., all positions  $i$  in  $Y$ , such that  $\delta_H(X, Y[i..i+|X|]) \leq k$ .

Given a string  $P$ , an ED-string  $\tilde{T}$  and an integer  $k \geq 1$ , we say that  $P$  *approximately matches* the fragment  $\tilde{T}[i..j]$  (with at most  $k$  mismatches) of  $\tilde{T}$ , or that an approximate occurrence of  $P$  starts at position  $i$  and ends at position  $j$  of  $\tilde{T}$ , if there is a string  $P'$  such that  $\delta_H(P, P') \leq k$  and  $P'$  matches  $\tilde{T}[i..j]$ . We stress that, as in the case of exact occurrences, each approximate occurrence of  $P$  in  $\tilde{T}$  is of the following forms: either  $P$  has Hamming distance at most  $k$  to a fragment of a string in a symbol of  $\tilde{T}$ ; or it uses parts of at least two symbols of  $\tilde{T}$ . In the latter case, a prefix of  $P$  is an approximate (possibly empty) suffix of a string in  $\tilde{T}[i]$ , a suffix of  $P$  is an approximate (possibly empty) prefix of a string in  $\tilde{T}[j]$ , and the remaining fragments of the pattern are approximate matches of a string in all other used symbols of  $\tilde{T}$  (except the first and the last one).

**Periodicity.** For a string  $X$ , we write  $X^\infty$  to denote the string  $X$  concatenated infinitely many times with itself. We call a string  $X$  *primitive* when it cannot be represented as  $Y^h$ , for some string  $Y$  and some integer  $h \geq 2$ . We say that a string  $X$  is a *d-period with offset  $\alpha$*  of some other string  $Y$  when  $\delta_H(Y, X^\infty[\alpha.. \alpha + |Y|]) \leq d$  for some  $d, \alpha \in \mathbb{Z}_{\geq 0}$ ; and we call the elements of  $\text{Mis}(Y, X^\infty[\alpha.. \alpha + |Y|])$  the *periodic mismatches*. If  $d = 0$ , then  $X$  is an *exact period* of  $Y$ , or just a period. Note that all cyclic shifts of  $X$  are also (approximate or exact) periods, but with different offsets.

**ED-string Matching.** Similarly as in Part A and in previous works (cf. [4]), we define the following problem – we assume integer  $k$  to be a fixed constant and not part of the input.

**Problem:** ELASTIC DEGENERATE STRING MATCHING (EDSM) WITH  $k$  MISMATCHES  
**Input:** A string  $P$  of length  $m$  and an ED-string  $\tilde{T}$  of length  $n$  and size  $N \geq m$ .  
**Output:** All positions in  $\tilde{T}$  where at least one approximate occurrence (with at most  $k$  mismatches) of  $P$  ends.

In the above problem, we call  $P$  the *pattern* and  $\tilde{T}$  the *text*.

**Active Prefixes Extension.** Similarly as in Part A and in previous works (cf. [4]), we solve EDSM WITH  $k$  MISMATCHES through the following auxiliary problem.

**Problem:** ACTIVE PREFIXES EXTENSION (APE) WITH  $k$  MISMATCHES  
**Input:** A string  $T$  of length  $m$ ,  $k + 1$  bitvectors  $U_0, U_1, \dots, U_k$  of size  $m$  each, and strings  $P_1, P_2, \dots$  of total length  $N$ .  
**Output:**  $k + 1$  bitvectors  $V_0, V_1, \dots, V_k$  of size  $m$  each, where  $V_{d'}[j'] = 1$  if and only if there is a string  $P_i$  and  $j \in \{0, 1, \dots, m - 1\}$  with  $U_d[j] = 1$  such that  $j' = j + |P_i|$  and  $d' \geq d + \delta_H(P_i, T[j..j'])$ .

In the above problem, we call  $T$  the *text*, and  $P_1, P_2, \dots$  the *patterns*.

### 3 EDSM with $k$ Mismatches via APE with $k$ Mismatches

We begin by showing how to reduce EDSM WITH  $k$  MISMATCHES to multiple instances of APE WITH  $k$  MISMATCHES. This does not require any new ideas, and it proceeds similarly as in Part A and in previous works (cf. [4]), so we only state it here for completeness.



As we mentioned before, each approximate occurrence of pattern  $P$  in ED-string  $\tilde{T}$  is:

1. either an approximate fragment of a string of a symbol; or
2. crossing the boundary between two consecutive symbols.

We explain how to detect the occurrences of each form separately.

**Approximate Fragments of Symbols.** To check if the pattern is an approximate fragment of a string of a symbol, we test each symbol of  $\tilde{T}$  separately (cf. [6]). To this end, we apply the technique of Landau and Vishkin [20], informally referred to as the *kangaroo jumps*. First, we preprocess the concatenation of all the symbols of  $\tilde{T}$  and the pattern with the following.

► **Lemma 3** (suffix tree [12] with LCA queries [18]). *A string  $T$  over a polynomial alphabet can be preprocessed in  $\mathcal{O}(|T|)$  time to allow computing the longest common prefix of any two suffixes  $T[i..]$  and  $T[j..]$  of  $T$  in constant time.*

Recall that pattern  $P$  is of length  $m$ . For a symbol  $X = \{X_1, X_2, \dots, X_t\}$  with  $t$  strings, we consider each string  $X_i$  and, for every  $j = 0, 1, \dots, |X_i| - m$ , check if  $\delta_H(X_i[j..j+m], P) \leq k$  in  $\mathcal{O}(k)$  time by repeatedly computing the longest common prefix of the remaining suffix of  $X_i$  and  $P$ . This takes  $\mathcal{O}(m + kN)$  total time.

**Crossing the Boundary between two Consecutive Symbols.** To check if  $P$  approximately matches a fragment  $\tilde{T}[i..i']$ , for some positions  $i < i'$ , we reduce the problem to multiple instances of APE WITH  $k$  MISMATCHES. We iterate through the symbols of  $\tilde{T}$  left-to-right and maintain  $k + 1$  bitvectors  $B_0, B_1, \dots, B_k$ , each of size  $m$ , such that  $B_d[j] = 1$  when the prefix  $P[.j - 1]$  of  $P$  is a  $d$ -approximate suffix of the current  $\tilde{T}[..i]$ , for  $d = 0, 1, \dots, k$ . Let  $N_i$  denote the size of  $\tilde{T}[i]$ , i.e., the total length of all strings in  $\tilde{T}[i]$ .

To proceed to the next iteration, and compute the bitvectors for  $\tilde{T}[..i + 1]$  from the bitvectors for  $\tilde{T}[..i]$ , we need to consider two possibilities. First, to consider the case when the  $d$ -approximate suffix is fully within  $\tilde{T}[i + 1]$ , for every  $d = 0, 1, \dots, k$ , we find all  $d$ -approximate prefixes of  $P$  that are suffixes of  $\tilde{T}[i + 1]$ . This is done in  $\mathcal{O}(kN_{i+1})$  time by iterating over all strings in  $\tilde{T}[i + 1]$ , considering for each of them every sufficiently short prefix of  $P$ , and computing the number of mismatches (if they do not exceed  $k$ ) using kangaroo jumps in  $\mathcal{O}(k)$  time. Second, to consider the case when the  $d$ -approximate suffix crosses the boundary between  $\tilde{T}[i]$  and  $\tilde{T}[i + 1]$ , we create and solve an instance of APE WITH  $k$  MISMATCHES with the bitvectors representing the results for  $\tilde{T}[..i]$  and the strings in  $\tilde{T}[i + 1]$ . We take as the new bitvectors the bitwise-OR of the bitvectors corresponding to both cases.

Before proceeding to the next iteration, we need to detect an occurrence that crosses the boundary between  $\tilde{T}[i]$  and  $\tilde{T}[i + 1]$ . To this end, we consider each string  $T \in \tilde{T}[i + 1]$ . Then, for every  $d = 0, 1, \dots, k$  and  $j = 0, 1, \dots, m - 1$  such that  $B_d[j] = 1$  and  $m - j \leq |T|$ , we check if  $P[j..]$  is a  $(k - d)$ -approximate prefix of  $\tilde{T}[i + 1]$  using kangaroo jumps in  $\mathcal{O}(k)$  time, and if so, report position  $i + 1$  as a  $k$ -approximate occurrence. Because we only need to consider  $|T| + 1$  possibilities for  $j$ , this takes  $\mathcal{O}(kN_{i+1})$  time.

We summarize the complexity of the reduction in the following lemma.

► **Lemma 4.** *Assume that APE WITH  $k$  MISMATCHES can be solved in  $f_k(m, N)$  time. Then EDSM WITH  $k$  MISMATCHES can be solved in  $\mathcal{O}(m + k \sum_i N_i + \sum_i f_k(m, N_i))$  time.*

## 4 Faster APE with $k$ Mismatches

We now move to designing efficient algorithms for APE WITH  $k$  MISMATCHES, separately for  $k = 1$  and then any constant  $k$ . Combined with the reduction underlying Lemma 4, this will result in Theorem 1 and Theorem 2. Recall that the input to an instance of APE WITH

$k$  MISMATCHES consists of a string  $T$  (called the text) of length  $m$  and a collection of strings  $P_1, P_2, \dots$  (called the patterns) of total length  $N$ .

For  $k = 1$ , the strings are partitioned depending on their lengths and parameters  $B'$  and  $B$  depending on  $m$ :

1. Very Short Case: the length of each string is  $\leq B'$ ,
2. Short Case: the length of each string is  $> B'$  and  $\leq B$ ,
3. Long Case: the length of each string is  $> B$ .

We separately solve the three obtained instances of APE WITH  $k$  MISMATCHES and return the bitwise-OR of the obtained bitvectors. For an arbitrary constant  $k$ , we have two cases:

1. Short Case: the length of each string is  $\leq B$ ,
2. Long Case: the length of each string is  $> B$ .

#### 4.1 Very Short Case (for $k = 1$ )

An efficient algorithm for this case can be obtained by using suffix trees and separately considering exact occurrences and occurrences with one mismatch. For completeness, we provide the proof in the appendix. An alternative algorithm is given in Part A, Lemma 18.

► **Theorem 5.** *An instance of APE WITH  $k$  MISMATCHES where  $k = 1$  and the length of each pattern is at most  $B'$  can be solved in  $\mathcal{O}(m(B')^2 + m(\log \log m)^2 + N)$  time.*

#### 4.2 Short Case

Recall that an instance of APE WITH  $k$  MISMATCHES consists of the patterns  $P_1, P_2, \dots, P_d$  of total length  $N$  and the text  $T[0..m-1]$ . Further, in this case, the length of each  $P_i$  is at least  $B'$  but at most  $B$ . We start with observing that, after  $\mathcal{O}(m + N)$ -time preprocessing, we can assume that  $\log d = \mathcal{O}(k \log m)$ , because we only need to keep patterns  $P_i$  such that  $\delta_H(T[j..j'], P_i) \leq k$ , for some fragment  $T[j..j']$ . This relates the number  $d$  of the patterns to  $k$  and  $m$ . Then, we state another known tool, and finally provide the algorithm.

**Reducing the Number of Patterns.** Recall that we only need to keep patterns  $P_i$  such that  $\delta_H(T[j..j'], P_i) \leq k$  for some fragment  $T[j..j']$ . If  $d = m^{\mathcal{O}(k)}$  then there is nothing to do. Otherwise, we consider all fragments of the text. For every such  $T[j..j']$ , we choose up to  $k$  positions where there is a mismatch, and for each of them we either choose a special character not occurring in  $T$ , represented by 0, or a character occurring somewhere in  $T$ . Overall, we have at most  $m^2 \cdot m^k \cdot (m+1)^k = m^{\mathcal{O}(k)}$  possibilities. For each of them, we construct the corresponding candidate string. Then, we sort the obtained candidate strings together with the patterns  $P_i$  with radix sort in  $\mathcal{O}(N + m^{\mathcal{O}(k)}) = \mathcal{O}(N)$  time. We scan the obtained sorted list and only keep patterns  $P_i$  equal to some candidate string.

**The  $k$ -errata Trie.** Cole, Gottlieb, and Lewenstein [11] considered the problem of preprocessing a dictionary of  $d$  patterns  $P_1, P_2, \dots, P_d$  of total length  $N$  for finding, given a query string  $Q$  of length  $m$ , whether  $Q$  is at Hamming distance at most  $k$  from some  $P_i$ . We provide a brief overview of their approach, following the exposition in [16] that provides some details not present in the original description.

For  $k = 0$ , this can be of course easily solved with a structure of size  $\mathcal{O}(N)$  and query time  $\mathcal{O}(m)$  by arranging the patterns in a trie. For larger values of  $k$ , the  $k$ -errata trie is defined recursively. In every step of the recursion, the input is a collection of  $x \leq d$  strings, each of them being a suffix of some pattern  $P_i$  decorated with its mismatch budget initially



set to  $k$ . We arrange the strings in a compact trie, and then recurse guided by the heavy-path decomposition [24] of the trie. The depth of the recursion is  $k$ , and on each level the overall number of strings increases by a factor of  $\log d$ , starting from  $d$ . Answering a query requires the following primitive: given a node of one of the compact tries and the remaining suffix of the query string  $Q[i..]$ , we need to navigate down starting from the given node while reading off the subsequent characters of  $Q[i..]$ . This needs to be done while avoiding explicitly scanning  $Q[i..]$ , as such a primitive is invoked multiple times. For a compact trie storing  $x$  suffixes of the patterns, such a primitive can be implemented by a structure of size  $\mathcal{O}(x \log x)$  and query time  $\mathcal{O}(\log \log N)$ , assuming that we know the position of every suffix of the query string in the suffix tree of  $P_1\$P_2\$P_3 \dots P_d\$P_d$  (also known as the generalized suffix tree of  $P_1, P_2, \dots, P_d$ ).

In our application, the query string will always be a fragment of the text  $T[i..j]$ . Thus, we can guarantee that the position of every suffix of the query string in the generalized suffix tree of  $P_1, P_2, \dots, P_d$  is known by building the generalized suffix tree of  $P_1, P_2, \dots, P_d, T$ . This gives us the position of every suffix  $T[i..]$  in the generalized suffix tree of  $P_1, P_2, \dots, P_d$ , from which we can infer the position of  $T[i..j]$ . We summarize the properties of such an implementation below.

► **Lemma 6** ([11]). *For any constant  $k$ , a dictionary of  $d$  patterns  $P_1, P_2, \dots, P_d$  of total length  $N$  and a text  $T[0..m-1]$  can be preprocessed in  $\mathcal{O}(m + N + d \log^{k+1} d)$  time to obtain a structure of size  $\mathcal{O}(m + N + d \log^k d)$ , such that for any fragment  $T[i..j]$  we can check in  $\mathcal{O}(\log^k d \log \log N)$  time whether  $\delta_H(T[i..j], P_i) \leq k$ , for some  $i$ .*

► **Theorem 7.** *For any constant  $k$ , an instance of APE WITH  $k$  MISMATCHES where the length of each pattern is at least  $B'$  and at most  $B$  can be solved in  $\mathcal{O}(mB \log^k m \log \log m + N + N/B' \cdot \log^{k+1} m)$  time.*

**Proof.** We start with applying Lemma 6 on the patterns  $P_1, P_2, \dots$  and the text  $T[0..m-1]$ . Then, iterate over every position  $j = 0, 1, \dots, m-1$ , length  $\ell = 1, 2, \dots, \min\{m-j, B\}$ , and  $d = 0, 1, \dots, k$  such that  $U_d[j] = 1$ . Next, for every  $d' = 0, 1, \dots, k-d$  we check if  $\delta_H(T[i..i+\ell], P_i) \leq d'$  for some  $i$ . If so, we set  $V_{d+d'}[j+\ell] = 1$ .

We analyze the overall time complexity. First, we need to construct the  $k$ -errata trie for  $P_1, P_2, \dots, P_d$  and  $T[0..m-1]$ . This takes  $\mathcal{O}(m + N + d \log^{k+1} d)$  time. Then, we consider  $\mathcal{O}(mB)$  possibilities for iterating over the position and the length, and for each of them spend  $\mathcal{O}(\log^k d \log \log N)$  time. As each  $P_i$  is of length at least  $B'$ , from the preprocessing we have  $\log d = \mathcal{O}(k \log m)$  and  $N \leq dm$ , the overall complexity is:

$$\mathcal{O}(m + N + d \log^{k+1} d + mB \log^k d \log \log N) = \mathcal{O}(N + N/B' \log^{k+1} m + mB \log^k m \log \log m)$$

as claimed. ◀

### 4.3 Long Case

In the most technical case, we assume that the length of each pattern  $P_i$  is at least  $B$ . We start with providing an overview, and then move to filling in the technical details.

The very high-level idea is to explicitly or implicitly process all occurrences of every pattern  $P_i$ . If a given pattern  $P_i$  occurs sufficiently few times in the text then we can afford to list and process each of its occurrences explicitly. Otherwise, we invoke the structural characterization of [9], which, roughly speaking, says that if there are many approximate occurrences of the same string sufficiently close to each other in the text, then the string and the relevant fragment of the text have a certain regular structure. Thus, we can certainly

hope to process all occurrences of such a pattern  $P_i$  together faster than by considering each of those occurrences one-by-one. However, this would not result in a speed-up, and in fact, we need to consider multiple such patterns  $P_i$  together. To this end, we need to further refine the characterization of [9]. Before we proceed with a description of our refinement, we start with a summary of the necessary tools from [9]. Then, we introduce some notation, introduce some simplifying assumptions, and then describe our refinement.

**Tools.** The authors of [9] phrase their algorithmic results using the framework of **PILLAR** operations. In this framework, we operate on strings, each of them specified by a handle. For two strings  $S$  and  $T$ , the following operations are possible (among others):

1. **Extract**( $S, \ell, r$ ): retrieve the string  $S[\ell \dots r]$ ,
2. **LCP**( $S, T$ ): compute the length of the longest common prefix of  $S$  and  $T$ ,
3. **IPM**( $S, T$ ): assuming that  $|T| \leq 2|S|$ , return the starting positions of all exact occurrences of  $S$  in  $T$  (at most two starting positions or an arithmetic progression of starting positions).

► **Lemma 8** ([9, Theorem 7.2]). *After an  $\mathcal{O}(N)$ -time preprocessing of a collection of strings of total length  $N$ , each **PILLAR** operation can be performed in  $\mathcal{O}(1)$  time.*

We apply the above lemma on the text and all the patterns in  $\mathcal{O}(m + N)$  time.

The first main result of [9] is the following structural characterization.

► **Lemma 9** ([9, Theorem 3.1]). *For each pattern of length  $|P| \leq 1.5|T|$ , at least one of the following holds:*

- $|\text{Occ}_k^H(P, T)| \leq 864k$ ,
- *There is a primitive string  $Q$  of length  $|Q| \leq |P|/128k$  such that  $\delta_H(P, Q^\infty[0 \dots |P|]) < 2k$ .*

Then, they convert the structural characterization (Lemma 9) into an efficient algorithm.

► **Lemma 10** ([9, Main Theorem 8]). *For any pattern of length  $|P| \leq 1.5|T|$ , we can compute (a representation of)  $\text{Occ}_k^H(P, T)$  in time  $\mathcal{O}(k^2 \log \log k)$  plus  $\mathcal{O}(k^2)$  **PILLAR** operations.*

The representation is a set of  $\mathcal{O}(k^2)$  arithmetic progressions. Further, as the algorithm follows the proof of Lemma 9, in fact it either outputs a set of  $\mathcal{O}(k)$  occurrences or finds a primitive string  $Q$  of length  $|Q| \leq |P|/128k$  such that  $\delta_H(P, Q^\infty[0 \dots |P|]) < 2k$ .

**Notation.** We rephrase the APE WITH  $k$  MISMATCHES problem as follows. For a pattern  $P$  and a set of positions  $A$  in the text  $T$  we define:

$$\text{Ext}_k^H(P, T, A) := \left( \text{Occ}_k^H(P, T) \cap A \right) + |P|,$$

which is also a set of positions in  $T$ . Then, APE WITH  $k$  MISMATCHES for a given set of patterns  $P_1, P_2, \dots$  and a text  $T$  can be solved as follows. For every  $d = 0, 1, \dots, k$  we set  $A$  to be  $U_d$ . Then, for every  $d' = d, d + 1, \dots, k$  we create an instance of computing:

$$\bigcup_i \text{Ext}_{k'}^H(P_i, T, A),$$

where  $k' = d' - d$ . The obtained bitvector contributes to the bitvector  $V_{d'}$ . From now on, we focus on designing an algorithm that computes  $\bigcup_i \text{Ext}_k^H(P_i, T, A)$ , and identify the underlying sets with their characteristic vectors. For two sets of integers  $X$  and  $Y$ , we define their sumset  $X \oplus Y := \{x + y : x \in X, y \in Y\}$ . For a set of integers  $X$  and a shift  $s$ , we define  $X + s := \{x + s : x \in X\}$ . The following result is well-known.

► **Lemma 11** (e.g. [14]). *Given  $X, Y \subseteq [0, \ell]$ , we can compute  $X \oplus Y$  in  $\mathcal{O}(\ell \log \ell)$  time.*

**Simplifying Assumptions.** It is convenient to assume that each pattern has roughly the same length, similar to the length of the text. More formally, our algorithm will assume that:

1.  $|P_i| \in [\ell, 1.1\ell]$ , for every  $i$ ,
2.  $|T| \in [\ell, 1.5\ell]$ ,

for some  $\ell$ . Any instance can be reduced to  $\mathcal{O}(\log m)$  instances in which  $|P_i| \in [\ell, 1.1\ell]$ , for every  $i$ , by considering  $\ell = 1.1^0, 1.1^1, 1.1^2, \dots \leq m$ . For each such  $\ell \geq B$ , we create a separate instance containing only patterns of length from  $[\ell, 1.1\ell]$ . As each pattern  $P_i$  falls within exactly one such instance, designing an algorithm running in  $\mathcal{O}(f(m) + N)$  time for every such instance, implies an algorithm running in  $\mathcal{O}(f(m) \log m + N)$  time for a general instance. To additionally guarantee that  $|T| \leq 1.5\ell$  (so that Lemma 9 can be directly applied), we choose  $|T|/0.4\ell$  fragments  $T_j$  such that each potential occurrence of a pattern  $P_i$  in  $T$  falls within some fragments  $T_j$ . Formally, if  $T_j$  is the  $1.5\ell$ -length fragment (possibly shorter for the very last fragment) starting at position  $0.4j\ell$  then:

$$\text{Occ}_k^H(P_i, T) = \bigcup_j \text{Occ}_k^H(P_i, T_j) + 0.4j\ell,$$

where we disregard fragments shorter than  $\ell$  as they cannot contain an occurrence of any  $P_i$ . From now on, always assume that we deal with a single text  $T$ , with  $|T| \in [\ell, 1.5\ell]$ , and a set of patterns  $P_i$  with lengths in  $[\ell, 1.1\ell]$  (we will sometimes omit the index of a pattern and simply write  $P$ ). The preprocessing from Lemma 8 is performed only once, and then in each instance we assume that any PILLAR operation can be performed in  $\mathcal{O}(1)$  time. The input bitvectors  $U_0, U_1, \dots, U_k$  in such an instance are fragments of the original input bitvectors, and after computing the output bitvectors  $V_0, V_1, \dots, V_k$  we update appropriate fragments of the original output bitvectors by computing bitwise-OR. The final number of restricted instances is  $\mathcal{O}(m/\ell \cdot \log m)$ , and each original pattern appears in  $\mathcal{O}(m/\ell)$  instances.

Consider a restricted instance containing  $d$  patterns. Our goal will be to solve it in  $\mathcal{O}((d + \ell\sqrt{d \log \ell}) \text{poly}(k))$  time. Before we proceed to describe such an algorithm, we analyze what this time implies for an algorithm solving the original instance.

► **Theorem 12.** *For any  $k$ , an instance of APE WITH  $k$  MISMATCHES where the length of each pattern is at least  $B$  can be solved in  $\mathcal{O}(m + (Nm/B^2 + m^2 \log^2 m/B) \text{poly}(k))$  time.*

**Proof.** We assume that a restricted instance with  $d$  patterns can be solved in  $\mathcal{O}((d + \ell\sqrt{d \log \ell}) \text{poly}(k))$  time, and describe an algorithm for solving a general instance of APE WITH  $k$  MISMATCHES.

Let  $d_i$  denote the number of patterns of length from  $[\ell_i, \ell_{i+1})$ , where  $\ell_i = 1.1^i$ , in the original instance. Recall that we only consider  $i$  such that  $\ell_i \geq B$ . After the initial  $\mathcal{O}(m + N)$ -time preprocessing, ignoring factors polynomial in  $k$ , the total time to solve the restricted instances is:

$$\begin{aligned} \mathcal{O}\left(\sum_i m/\ell_i (d_i + \ell_i \sqrt{d_i \log \ell_i})\right) &= \mathcal{O}\left(\sum_i m/\ell_i^2 \cdot d_i \ell_i + \sum_i m \sqrt{d_i \log m}\right) \\ &= \mathcal{O}(Nm/B^2 + \sum_i m \sqrt{d_i \log m}), \end{aligned}$$

where we have used  $\sum_i d_i \ell_i = N$  and  $\ell_i \geq B$ . We split the sum by separately considering  $i$  such that  $m\sqrt{d_i \log m} \leq d_i \ell_i$ , i.e.,  $\sqrt{d_i} \geq m\sqrt{\log m}/\ell_i$ . This gives us:

$$\mathcal{O}(Nm/B^2 + \sum_i d_i \ell_i + \sum_i m^2 \log m/\ell_i) = \mathcal{O}(Nm/B^2 + N + m^2 \log^2 m/B).$$

Thus, as long as we indeed manage to solve a restricted instance in the promised complexity, we obtain the theorem. ◀

In what follows, we describe an algorithm for solving a restricted instance of APE WITH  $k$  MISMATCHES containing  $d$  patterns in  $\mathcal{O}((d + \ell\sqrt{d\log\ell})\text{poly}(k))$  time.

**Additional Assumptions.** We start with applying Lemma 10 on every pattern  $P_i$  to obtain a representation of its occurrences in  $T$  in  $\mathcal{O}(d\text{poly}(k))$  time. As mentioned earlier, the algorithm underlying Lemma 10 either outputs a set of  $\mathcal{O}(k)$  occurrences of  $P_i$  in  $T$  or finds a primitive  $2k$ -period  $Q_i$  of  $P_i$  such that  $|Q_i| \leq |P|/128k < \ell/100k$  (note that the second inequality holds because  $|P| < 1.1\ell$ ). In the latter case, we also obtain a representation of the whole set of occurrences as  $\mathcal{O}(k^2)$  arithmetic progressions.

If there are  $\mathcal{O}(k)$  occurrences of  $P_i$  in  $T$ , then we process each of them naively in  $\mathcal{O}(dk)$  time. From now on we can thus assume otherwise for every pattern  $P_i$ . Then, we consider the text  $T$ , and ensure that it is fully covered by approximate occurrences of the patterns:

- some pattern  $P$  is a  $k$ -mismatch prefix of  $T$ , formally  $\delta_H(P, T[0..|P|]) \leq k$ ; and
- some pattern  $P'$  is a  $k$ -mismatch suffix of  $T$ , formally  $\delta_H(P', T[|T| - |P'|..|T|]) \leq k$ .

This is guaranteed by removing some prefix and some suffix of the text; it can be implemented in  $\mathcal{O}(d\text{poly}(k))$  time by extracting the first and the last occurrence from each arithmetic progression in the representation. Then the following claim can be inferred from the characterization of the period case in [9]. We provide a proof in the appendix for completeness.

► **Lemma 13.** *All  $Q_i$ s are cyclically equivalent, and every  $Q_i$  is a  $6k$ -period of the text  $T$ .*

We choose  $Q$  to be a cyclic shift of the period  $Q_1$  that we got for the pattern  $P_1$ , so  $Q$  is a  $2k$ -period of every pattern, and a  $6k$ -period with offset 0 of the text. This can be implemented in  $\mathcal{O}(d\text{poly}(k) + \ell)$  time as follows. For every  $P_i$ , because  $|Q_i| < \ell/100k$  and  $\delta_H(P_i, Q_i^\infty[0..|P_i|]) \leq 2k$ , we have  $P_i[j \cdot |Q_i|..(j+2) \cdot |Q_i|] = Q_i Q_i$  for some  $j$ . Further, such a  $j$  can be computed in  $\mathcal{O}(k)$  time by just trying  $j = 0, 1, 2, \dots$  and verifying each  $j$  by computing the longest common prefix twice. Overall, this takes  $\mathcal{O}(d\text{poly}(k))$  time. We start with setting  $Q' := Q_1$ . Then, we search for a cyclic shift  $Q$  of  $Q'$  such that  $\delta_H(T, Q^\infty[0..|T|]) \leq 6k$ . To this end, we check all possible  $\mathcal{O}(\ell/k)$  cyclic shifts. To verify whether  $Q$  is a good cyclic shift, we extract the mismatches between  $T$  and  $Q^\infty[0..|T|]$ , terminating when there are more than  $6k$ . The next mismatch can be found in constant time by first computing the longest common prefix of the remaining suffix of  $T$  with an appropriate cyclic shift of  $Q$ , and if there is none by computing the longest common prefix of the remaining suffix of  $T$  with the suffix shortened by  $|Q|$  characters. Overall, this takes  $\mathcal{O}(\ell)$  time. After having found  $Q$ , we also compute, for every pattern  $P_i$ , an integer  $r$  such that  $\delta_H(P_i, Q^\infty[r..r + |P_i|]) \leq 2k$ , which can be done with a single internal pattern matching query to find an occurrence of  $Q$  in  $Q_i Q_i$ .

**The Algorithm.** To obtain an efficient algorithm, we will partition the set of all positions  $[0..|T|]$  into  $\mathcal{O}(k)$  consecutive regions  $R_0, R_1, \dots, R_b$  with the property that if we restrict the text to any region  $R_i$ , the corresponding fragment is *almost* periodic with respect to  $Q$ ; more specifically, it may have a single periodic mismatch at the rightmost position. Then, for each pair of regions  $R_s$  and  $R_t$ , with  $s \leq t$ , we separately calculate the set of extensions induced by occurrences of the pattern starting at positions  $x \in R_s$  such that  $(x + |P|) \in R_t$ :

$$\bigcup_i \text{Ext}_k^H(P_i, T, A) = \bigcup_s \bigcup_t \bigcup_i \text{Ext}_k^H(P_i, T, A \cap R_s) \cap R_t.$$

Since  $b = \mathcal{O}(k)$ , this allows us to reduce the problem to  $\mathcal{O}(k^2)$  separate instances of calculating:

$$\bigcup_i \text{Ext}_k^H(P_i, T, A \cap R_s) \cap R_t.$$

Consider a single pattern  $P$ , and recall that by the additional assumptions, we have a primitive string  $Q$  of length  $|Q| < \ell/100k$  such that:

$$\begin{aligned} \delta_H(P, \bar{P}) &\leq 2k & \text{for } \bar{P} &:= Q^\infty[r \dots r + |P|], \\ \delta_H(T, \bar{T}) &\leq 6k & \text{for } \bar{T} &:= Q^\infty[0 \dots |T|]. \end{aligned}$$

Further, let  $C_r$  be the positions congruent to  $r$  modulo  $|Q|$  in the text. We start with recalling from [9] that the positions of all  $k$ -mismatch occurrences of  $P$  in  $T$  are congruent modulo  $|Q|$ . We provide a proof in the appendix for completeness.

► **Lemma 14.**  $\text{Occ}_k^H(P, T) \subseteq \{r + i|Q| : i \in \mathbb{Z}\}$ .

Following Lemma 14, choose  $r$  such that  $\text{Occ}_k^H(P, T) \subseteq C_r$ . In order to characterize  $\text{Occ}_k^H(P, T)$ , let us now analyze the values  $\delta_H(P, T[x \dots x + |P|])$  for  $x \in C_r$ . From triangle inequality, we have

$$\delta_H(P, T[x \dots x + |P|]) \leq \delta_H(P, \bar{P}) + \delta_H(\bar{P}, \bar{T}[x \dots x + |P|]) + \delta_H(\bar{T}[x \dots x + |P|], T[x \dots x + |P|]). \quad (1)$$

Observe that

$$\bar{P} = \bar{T}[x \dots x + |P|], \quad (2)$$

since both strings have a period  $Q$  with offsets congruent modulo  $|Q|$ , which gives us

$$\delta_H(P, T[x \dots x + |P|]) \leq \delta_H(P, \bar{P}) + \delta_H(\bar{T}[x \dots x + |P|], T[x \dots x + |P|]). \quad (3)$$

We will later show that the above inequality is in fact an equality for all  $x \in C_r$  except for  $\mathcal{O}(k^2)$  exceptions  $E$ . Specifically, define

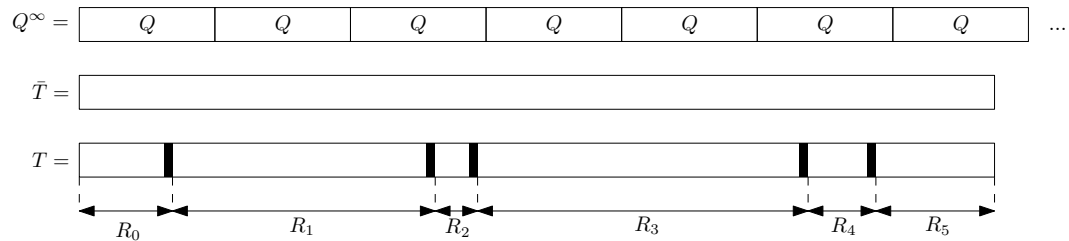
$$E := \{\tau - \pi : \pi \in \text{Mis}(P, \bar{P}), \tau \in \text{Mis}(T, \bar{T})\} \cap [0, |T|],$$

which is the set of all starting positions  $x$  in  $T$  such that when comparing  $P$  and  $T[x \dots x + |P|]$ , at least one pair of mismatches aligns with each other. Note that  $|E| \leq \delta_H(P, \bar{P}) \cdot \delta_H(T, \bar{T}) \leq 2k \cdot 6k = 12k^2$ . Finally, we define  $R_0, R_1, \dots, R_b$ . Let:

- $\tau_0 < \tau_1 < \dots < \tau_{b-1}$  denote the sorted elements of  $\text{Mis}(T, \bar{T})$ ,
- $R_i := (\tau_{i-1} \dots \tau_i]$  for all  $0 \leq i \leq b$ , where we set  $\tau_{-1} := -1$  and  $\tau_b := |T| - 1$ .

This is illustrated in Figure 1.

The elements of  $\text{Mis}(T, \bar{T})$  can be computed in  $\mathcal{O}(k)$  time by extracting the mismatches with longest common prefix queries, which allows us to find the regions in  $\mathcal{O}(\text{poly}(k))$  time. Similarly, we can compute the elements of  $\text{Mis}(P, \bar{P})$  in  $\mathcal{O}(k)$  time, so we can also compute the set  $E$  in  $\mathcal{O}(\text{poly}(k))$  time. We stress that the regions are the same for every considered pattern  $P$  (but the set  $E$  does depend on the pattern  $P$ ).



■ **Figure 1** Regions  $R_0, R_1, \dots, R_b$ . Black rectangles denote the elements of  $\text{Mis}(T, \bar{T})$ .

Now we can state our extension of the structural characterization of [9].

► **Lemma 15.** *There exist an integer  $r$  and a set  $E$ , with  $|E| = \mathcal{O}(k^2)$ , such that, for each pair  $s, t$ , with  $0 \leq s \leq t \leq b$ :*

$$\text{Ext}_k^H(P, T, A \cap R_s) \cap R_t = \begin{cases} ((C_r \cap A \cap R_s) + |P|) \cap R_t & \text{if } \delta_H(P, \bar{P}) + t - s \leq k, \\ \text{some subset of } E + |P| & \text{otherwise.} \end{cases}$$

**Proof.** For any  $s > t$  the resulting set is trivially empty. Now let us fix any  $s \leq t$  and define

$$I_{st}^P := C_r \cap A \cap R_s \cap (R_t - |P|).$$

We will show that either  $I_{st}^P \subseteq \text{Occ}_k^H(P, T)$  or  $I_{st}^P \cap \text{Occ}_k^H(P, T) \subseteq E$ , depending on whether  $\delta_H(P, \bar{P}) + t - s \leq k$  by using the following property:

► **Proposition 16.** *For all  $x \in I_{st}^P$ , we have  $\delta_H(T[x \dots x + |P|], \bar{T}[x \dots x + |P|]) = t - s$ .*

**Proof.** We know that

- $x \in R_s$ , so  $\tau_{s-1} < x \leq \tau_s$ ,
  - $x + |P| \in R_t$ , so  $\tau_{t-1} < x + |P| \leq \tau_t$ ,
- therefore

$$[x, x + |P|) \cap \{\tau_0, \tau_1, \dots, \tau_{b-1}\} = \{\tau_s, \tau_{s+1}, \dots, \tau_{t-1}\},$$

and finally

$$\begin{aligned} \delta_H(T[x \dots x + |P|], \bar{T}[x \dots x + |P|]) &= |[x, x + |P|) \cap \text{Mis}(T, \bar{T})| \\ &= |[x, x + |P|) \cap \{\tau_0, \tau_1, \dots, \tau_{b-1}\}| \\ &= |\{\tau_s, \tau_{s+1}, \dots, \tau_{t-1}\}| \\ &= t - s. \end{aligned} \quad \blacktriangleleft$$

Now assume that  $\delta_H(P, \bar{P}) + t - s \leq k$ . In that case, recall that by (3) combined with Proposition 16, for every  $x \in I_{st}^P$ , we have

$$\delta_H(P, T[x \dots x + |P|]) \leq \delta_H(P, \bar{P}) + \delta_H(\bar{T}[x \dots x + |P|], T[x \dots x + |P|]) = \delta_H(P, \bar{P}) + t - s \leq k.$$

Consequently  $x \in \text{Occ}_k^H(P, T)$ , and therefore  $I_{st}^P \subseteq \text{Occ}_k^H(P, T)$ . In that case observe that

$$\begin{aligned} \text{Ext}_k^H(P, T, A \cap R_s) \cap R_t &= \left( (\text{Occ}_k^H(P, T) \cap A \cap R_s) + |P| \right) \cap R_t \\ &= \left( \text{Occ}_k^H(P, T) \cap A \cap R_s \cap (R_t - |P|) \right) + |P| \\ &= \left( \text{Occ}_k^H(P, T) \cap C_r \cap A \cap R_s \cap (R_t - |P|) \right) + |P| \\ &= \left( \text{Occ}_k^H(P, T) \cap I_{st}^P \right) + |P| \\ &= I_{st}^P + |P|, \end{aligned}$$

where the third equality follows from  $\text{Occ}_k^H(P, T) \subseteq C_r$  and the fifth from  $I_{st}^P \subseteq \text{Occ}_k^H(P, T)$ . Since  $I_{st}^P + |P| = ((C_r \cap A \cap R_s) + |P|) \cap R_t$  as desired, the proof for this case is complete.

For the second case, when  $\delta_H(P, \bar{P}) + t - s > k$ , we need to make use of the following property:

► **Proposition 17.** *For all  $x \in C_r \setminus E$ , we have*

$$\delta_H(P, T[x \dots x + |P|]) = \delta_H(P, \bar{P}) + \delta_H(\bar{T}[x \dots x + |P|], T[x \dots x + |P|]).$$



**Proof.** Consider the triangle inequality (1) stated explicitly for each position  $i \in [0 \dots |T|]$ :

$$\delta_H(P[i], T[i+x]) \leq \delta_H(P[i], \bar{P}[i]) + \delta_H(\bar{P}[i], \bar{T}[i+x]) + \delta_H(\bar{T}[i+x], T[i+x]).$$

From (2) we already know that  $\delta_H(\bar{P}[i], \bar{T}[i+x]) = 0$ , thus

$$\delta_H(P[i], T[i+x]) \leq \delta_H(P[i], \bar{P}[i]) + \delta_H(\bar{T}[i+x], T[i+x]).$$

We will now show that the above inequality holds with equality by considering two cases. The proof is completed by summing the equations for all  $i \in [0 \dots |T|]$ .

1°  $i+x \notin \text{Mis}(T, \bar{T})$ . In that case, observe that by triangle inequality

$$\delta_H(P[i], \bar{P}[i]) \leq \delta_H(P[i], T[i+x]) + \delta_H(T[i+x], \bar{T}[i+x]) + \delta_H(\bar{T}[i+x], \bar{P}[i])$$

and since  $\delta_H(T[i+x], \bar{T}[i+x]) = 0$  by the assumption and  $\delta_H(\bar{T}[i+x], \bar{P}[i]) = 0$  by (2), we get

$$\delta_H(P[i], \bar{P}[i]) + \delta_H(\bar{T}[i+x], T[i+x]) \leq \delta_H(P[i], T[i+x]).$$

2°  $i \notin \text{Mis}(P, \bar{P})$ . In that case, observe that by triangle inequality

$$\delta_H(\bar{T}[i+x], T[i+x]) \leq \delta_H(\bar{T}[i+x], \bar{P}[i]) + \delta_H(\bar{P}[i], P[i]) + \delta_H(P[i], T[i+x])$$

and similarly, since  $\delta_H(\bar{P}[i], P[i]) = 0$  and  $\delta_H(\bar{T}[i+x], \bar{P}[i]) = 0$ , we again get

$$\delta_H(P[i], \bar{P}[i]) + \delta_H(\bar{T}[i+x], T[i+x]) \leq \delta_H(P[i], T[i+x]).$$

It remains to show that every  $i \in [0 \dots |T|]$  falls into at least one of these two cases. Indeed, if for some  $i$  we would have  $i \in \text{Mis}(P, \bar{P})$  and  $i+x \in \text{Mis}(T, \bar{T})$ , then by the definition of  $E$ , we would get  $x \in E$ , which is a contradiction.  $\blacktriangleleft$

By Proposition 17, combined with Proposition 16 for every  $x \in I_{st}^P \setminus E$ , we have

$$\delta_H(P, T[x \dots x + |P|]) = \delta_H(P, \bar{P}) + \delta_H(\bar{T}[x \dots x + |P|], T[x \dots x + |P|]) = \delta_H(P, \bar{P}) + t - s > k,$$

therefore  $x \notin \text{Occ}_k^H(P, T)$ , which implies  $I_{st}^P \cap \text{Occ}_k^H(P, T) \subseteq E$ . Following the same reasoning as before, up to the fourth equality, we get

$$\text{Ext}_k^H(P, T, A \cap R_s) \cap R_t = (\text{Occ}_k^H(P, T) \cap I_{st}^P) + |P| \subseteq E + |P|$$

as required.  $\blacktriangleleft$

We apply Lemma 15 on every pattern  $P_i$ . Whenever the second case applies, we process all occurrences of  $P_i$  naively. We observe that by definition we have

$$\text{Ext}_k^H(P, T, A \cap R_s) \cap R_t = ((\text{Occ}_k^H(P, T) \cap E \cap A \cap R_s) + |P|) \cap R_t.$$

Since we already have access to the previously calculated representation of  $\text{Occ}_k^H(P, T)$ , we can simply check for each element in  $E$  whether it is in  $\text{Occ}_k^H(P, T) \cap A \cap R_s$  in  $\mathcal{O}(k^2)$  time, as the representation of  $\text{Occ}_k^H(P, T)$  is of size  $\mathcal{O}(k^2)$ , so  $\mathcal{O}(|E| \cdot k^2) = \mathcal{O}(k^4)$  time in total.

For the remaining patterns that fall into the first case, we still use the naive approach if  $|Q| > z$  for some threshold  $z$  to be chosen later. Since  $|C_r| = \mathcal{O}(\ell/|Q|)$ , this takes  $\mathcal{O}(\ell/z)$  time per pattern. Otherwise,  $|Q| \leq z$ . We partition the remaining patterns into  $|Q|$  groups with the same  $r$ . Formally, let  $\mathcal{P}_r$  denote the set of patterns  $P$  with a specific value of  $r$ :

$$\bigcup_i \text{Ext}_k^H(P_i, T, A \cap R_s) \cap R_t = \bigcup_{r=0}^{|Q|-1} \bigcup_{P \in \mathcal{P}_r} ((C_r \cap A \cap R_s) + |P|) \cap R_t.$$

We calculate the result for each  $r$  separately by phrasing it as a sumset of some common set of positions with the set of pattern lengths, where the result is then truncated to  $R_t$ :

$$\bigcup_{P \in \mathcal{P}_r} ((C_r \cap A \cap R_s) + |P|) \cap R_t = ((C_r \cap A \cap R_s) \oplus \{|P| : P \in \mathcal{P}_r\}) \cap R_t.$$

This takes  $\mathcal{O}(z\ell \log \ell)$  total time by Lemma 11. Overall, the time complexity is  $\mathcal{O}(d \text{poly}(k) + \ell + d\ell/z + z\ell \log \ell)$ , which by choosing  $z = \sqrt{d/\log \ell}$  becomes  $\mathcal{O}(d \text{poly}(k) + \ell\sqrt{d \log \ell})$  as promised.

#### 4.4 Combining the Cases

After designing an algorithm for every case, we show how to combine them to obtain the claimed bounds.

► **Theorem 1.** *Given a pattern  $P$  of length  $m$  and an ED-string  $\tilde{T}$  of length  $n$  and size  $N$ , EDSM WITH  $k$  MISMATCHES, for  $k = 1$ , can be solved in  $\mathcal{O}(nm^{1.5} \text{polylog } m + N)$  time.*

**Proof.** By Lemma 4, to prove the theorem it is enough to show how to solve APE WITH  $k$  MISMATCHES, where  $k = 1$ , in  $\mathcal{O}(m^{1.5} \text{polylog } m + N)$  time. We choose  $B' = \log^2 m$  and  $B = \sqrt{m}$ . For patterns of length at most  $B'$ , we use Theorem 5. For patterns of length at least  $B'$  but at most  $B$ , we use Theorem 7. Finally, for patterns of length at least  $B$ , we use Theorem 12. Summing up the time complexities, we obtain

$$\mathcal{O}(m \log^4 m + N) + \mathcal{O}(m \log^3 m \log \log m + N) + \mathcal{O}(m + N + m^{1.5} \log^2 m) = \mathcal{O}(m^{1.5} \text{polylog } m + N)$$

as required. ◀

► **Theorem 2.** *Given a pattern  $P$  of length  $m$  and an ED-string  $\tilde{T}$  of length  $n$  and size  $N$ , EDSM WITH  $k$  MISMATCHES, for any constant  $k \geq 1$ , can be solved in  $\mathcal{O}((nm^{1.5} + N) \text{polylog } m)$  time.*

**Proof.** By Lemma 4, to prove the theorem it is enough to show how to solve APE WITH  $k$  MISMATCHES for any constant  $k \geq 1$  in  $\mathcal{O}((m^{1.5} + N) \text{polylog } m)$  time. We choose  $B = \sqrt{m}$ . For patterns of length at most  $B$ , we use Theorem 7. Finally, for patterns of length at least  $B$ , we use Theorem 12. Summing up the time complexities, we obtain

$$\mathcal{O}(m^{1.5} \log^k m \log \log m + N \log^{k+1} m) + \mathcal{O}(m + N + m^{1.5} \log^2 m) = \mathcal{O}((m^{1.5} + N) \text{polylog } m)$$

as required. ◀

---

#### References

- 1 Josh Alman, Ran Duan, Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou. More asymmetry yields faster matrix multiplication. In Yossi Azar and Debmalya Panigrahi, editors, *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025, New Orleans, LA, USA, January 12–15, 2025*, pages 2005–2039. SIAM, 2025. doi:10.1137/1.9781611978322.63.
- 2 Kotaro Aoyama, Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Faster online elastic degenerate string matching. In Gonzalo Navarro, David Sankoff, and Binhai Zhu, editors, *Annual Symposium on Combinatorial Pattern Matching, CPM 2018, July 2–4, 2018 - Qingdao, China*, volume 105 of *LIPICs*, pages 9:1–9:10. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPICs.CPM.2018.9.

- 3 Rocco Ascone, Giulia Bernardini, Alessio Conte, Massimo Equi, Estéban Gabory, Roberto Grossi, and Nadia Pisanti. A unifying taxonomy of pattern matching in degenerate strings and founder graphs. In Solon P. Pissis and Wing-Kin Sung, editors, *24th International Workshop on Algorithms in Bioinformatics, WABI 2024, September 2-4, 2024, Royal Holloway, London, United Kingdom*, volume 312 of *LIPICs*, pages 14:1–14:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPICs.WABI.2024.14.
- 4 Giulia Bernardini, Estéban Gabory, Solon P. Pissis, Leen Stougie, Michelle Sweering, and Wiktor Zuba. Elastic-degenerate string matching with 1 error or mismatch. *Theory Comput. Syst.*, 68(5):1442–1467, 2024. doi:10.1007/S00224-024-10194-8.
- 5 Giulia Bernardini, Pawel Gawrychowski, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Elastic-degenerate string matching via fast matrix multiplication. *SIAM J. Comput.*, 51(3):549–576, 2022. doi:10.1137/20M1368033.
- 6 Giulia Bernardini, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Approximate pattern matching on elastic-degenerate text. *Theor. Comput. Sci.*, 812:109–122, 2020. doi:10.1016/J.TCS.2019.08.012.
- 7 Karl Bringmann, Marvin Künnemann, and Philip Wellnitz. Few matches or almost periodicity: Faster pattern matching with mismatches in compressed texts. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1126–1145. SIAM, 2019. doi:10.1137/1.9781611975482.69.
- 8 Thomas Büchler, Jannik Olbrich, and Enno Ohlebusch. Efficient short read mapping to a pangenome that is represented by a graph of ED strings. *Bioinform.*, 39(5), 2023. doi:10.1093/BIOINFORMATICS/BTAD320.
- 9 Panagiotis Charalampopoulos, Tomasz Kociumaka, and Philip Wellnitz. Faster approximate pattern matching: A unified approach. In Sandy Irani, editor, *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 978–989. IEEE, 2020. doi:10.1109/FOCS46700.2020.00095.
- 10 Aleksander Cislak, Szymon Grabowski, and Jan Holub. Sopang: online text searching over a pan-genome. *Bioinform.*, 34(24):4290–4292, 2018. doi:10.1093/BIOINFORMATICS/BTY506.
- 11 Richard Cole, Lee-Ad Gottlieb, and Moshe Lewenstein. Dictionary matching and indexing with errors and don't cares. In László Babai, editor, *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 91–100. ACM, 2004. doi:10.1145/1007352.1007374.
- 12 Martin Farach. Optimal suffix tree construction with large alphabets. In *38th Annual Symposium on Foundations of Computer Science, FOCS '97, Miami Beach, Florida, USA, October 19-22, 1997*, pages 137–143. IEEE Computer Society, 1997. doi:10.1109/SFCS.1997.646102.
- 13 Nathan J. Fine and Herbert S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16:109–114, 1965. doi:10.1090/S0002-9939-1965-0174934-9.
- 14 Martin Fürer. How fast can we multiply large integers on an actual computer? In Alberto Pardo and Alfredo Viola, editors, *LATIN 2014: Theoretical Informatics - 11th Latin American Symposium, Montevideo, Uruguay, March 31 - April 4, 2014. Proceedings*, volume 8392 of *Lecture Notes in Computer Science*, pages 660–670. Springer, 2014. doi:10.1007/978-3-642-54423-1\_57.
- 15 Esteban Gabory, Moses Njagi Mwaniki, Nadia Pisanti, Solon P. Pissis, Jakub Radoszewski, Michelle Sweering, and Wiktor Zuba. Pangenome comparison via ED strings. *Frontiers in Bioinformatics*, 4, 2024. doi:10.3389/fbinf.2024.1397036.
- 16 Pawel Gawrychowski, Gad M. Landau, and Tatiana Starikovskaya. Fast entropy-bounded string dictionary look-up with mismatches. In Igor Potapov, Paul G. Spirakis, and James Worrell, editors, *43rd International Symposium on Mathematical Foundations of Computer*

- Science, MFCS 2018, August 27-31, 2018, Liverpool, UK*, volume 117 of *LIPICs*, pages 66:1–66:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPICS.MFCS.2018.66.
- 17 Roberto Grossi, Costas S. Iliopoulos, Chang Liu, Nadia Pisanti, Solon P. Pissis, Ahmad Retha, Giovanna Rosone, Fatima Vayani, and Luca Versari. On-line pattern matching on similar texts. In Juha Kärkkäinen, Jakub Radoszewski, and Wojciech Rytter, editors, *28th Annual Symposium on Combinatorial Pattern Matching, CPM 2017, July 4-6, 2017, Warsaw, Poland*, volume 78 of *LIPICs*, pages 9:1–9:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPICS.CPM.2017.9.
  - 18 Dov Harel and Robert Endre Tarjan. Fast algorithms for finding nearest common ancestors. *SIAM J. Comput.*, 13(2):338–355, 1984. doi:10.1137/0213024.
  - 19 Costas S. Iliopoulos, Ritu Kundu, and Solon P. Pissis. Efficient pattern matching in elastic-degenerate strings. *Inf. Comput.*, 279:104616, 2021. doi:10.1016/J.IC.2020.104616.
  - 20 Gad M. Landau and Uzi Vishkin. Efficient string matching with  $k$  mismatches. *Theor. Comput. Sci.*, 43:239–249, 1986. doi:10.1016/0304-3975(86)90178-7.
  - 21 Veli Mäkinen, Bastien Cazaux, Massimo Equi, Tuukka Norri, and Alexandru I. Tomescu. Linear time construction of indexable founder block graphs. In Carl Kingsford and Nadia Pisanti, editors, *20th International Workshop on Algorithms in Bioinformatics, WABI 2020, September 7-9, 2020, Pisa, Italy (Virtual Conference)*, volume 172 of *LIPICs*, pages 7:1–7:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPICS.WABI.2020.7.
  - 22 Solon P. Pissis, Jakub Radoszewski, and Wiktor Zuba. Faster approximate elastic-degenerate string matching – Part A. In Paola Bonizzoni and Veli Mäkinen, editors, *36th Annual Symposium on Combinatorial Pattern Matching, CPM 2025, June 17-19, 2025, Milan, Italy*, volume 331 of *LIPICs*, pages 28:1–28:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPICS.CPM.2025.28.
  - 23 Milan Ruzic. Constructing efficient dictionaries in close to sorting time. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfssdóttir, and Igor Walukiewicz, editors, *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part I: Track A: Algorithms, Automata, Complexity, and Games*, volume 5125 of *Lecture Notes in Computer Science*, pages 84–95. Springer, 2008. doi:10.1007/978-3-540-70575-8\_8.
  - 24 Daniel Dominic Sleator and Robert Endre Tarjan. A data structure for dynamic trees. *J. Comput. Syst. Sci.*, 26(3):362–391, 1983. doi:10.1016/0022-0000(83)90006-5.
  - 25 Peter Weiner. Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory, Iowa City, Iowa, USA, October 15-17, 1973*, pages 1–11. IEEE Computer Society, 1973. doi:10.1109/SWAT.1973.13.

## **A**    **Very Short Case (for $k = 1$ )**

Before stating the algorithm, we need a few standard tools.

**Suffix Tree.** A *trie* is a (rooted) tree, where every edge is labeled with a single character. Each node of a trie represents the string obtained by concatenating the labels on its path from the root. We consider only deterministic tries, meaning that the labels of all edges outgoing from the same node are pairwise distinct. Then, a *compact trie* is obtained from a trie by collapsing maximal downward paths on which every inner node has exactly one child. The remaining nodes are called *explicit*, and the nodes that have been removed while collapsing the paths are called *implicit*. In a compact trie, every edge is labeled with a nonempty string, and the first characters of all edges outgoing from the same node are pairwise distinct.

The *suffix tree* of a string  $T[0..n-1]$  is the compact trie of all the suffixes of  $T\$$ , where  $\$$  is a special character not occurring anywhere in  $T$  [25]. Thus, there are  $n+1$  leaves in the suffix tree of  $T$ , and it contains  $\mathcal{O}(n)$  nodes and edges. The label of each edge is equal to some fragment  $T[i..j]$ , and we represent it by storing  $i$  and  $j$ ; thus the whole suffix tree needs only  $\mathcal{O}(n)$  space. For constructing the suffix tree we apply the following result.

► **Lemma 18** ([12]). *The suffix tree of a string  $T[0..n-1]$  over a polynomial alphabet can be constructed in  $\mathcal{O}(n)$  time.*

The suffix tree of string  $T$ , denoted by  $ST_T$ , allows us to easily check if any string  $X$  is a substring of  $T$  by starting at the root and simulating navigating down in the trie storing the suffixes of  $T$ . In every step, we need to choose the outgoing edge labeled by the next character  $X[i]$ . If the current node is implicit, this is trivial to implement in constant time. Otherwise, we might have multiple outgoing edges, and we need to store the first characters of their edges in an appropriate structure. To this end, we use deterministic dictionaries.

► **Lemma 19** ([23, Theorem 3]). *Given a set  $S$  of  $n$  integer keys, we can build in  $\mathcal{O}(n(\log \log n)^2)$  time a structure of size  $\mathcal{O}(n)$ , such that given any integer  $x$  we can check in  $\mathcal{O}(1)$  time if  $x \in S$ , and if so return its associated information.*

We apply Lemma 19 at each explicit node of the suffix tree. This takes  $\mathcal{O}(n(\log \log n)^2)$  total time, and then allows us to implement navigating down in  $\mathcal{O}(|X|)$  total time. This also gives us, for each prefix  $X[..i]$ , its unique identifier: if we are at an explicit node then it is simply its preorder number, and otherwise it is a pair consisting of the preorder number of the nearest explicit ancestor and the length of the current prefix. If in any step we cannot proceed further, we set the identifier to null denoting that the prefix does not occur in  $T$ . Such identifiers have the property that the identifier of  $X[..i]$  is null if and only if  $X[..i]$  does not occur in  $T$ , and the identifiers of  $X[..i]$  and  $Y[..j]$  that both occur in  $T$  are equal if and only if the strings themselves are equal. Further, we can think that each identifier is an integer from  $\{1, 2, \dots, n^2\}$ .

► **Theorem 5.** *An instance of APE WITH  $k$  MISMATCHES where  $k = 1$  and the length of each pattern is at most  $B'$  can be solved in  $\mathcal{O}(m(B')^2 + m(\log \log m)^2 + N)$  time.*

**Proof.** We assume that the length of each pattern  $P_i$  is at most  $B'$ . Recall that we are given bitvectors  $U_0, U_1$  and the goal is to compute the bitvectors  $V_0, V_1$ . This will be done by explicitly listing all fragments  $T[j..j']$  such that  $\delta_H(T[j..j'], P_i) = 0$ , for some  $i$ , and  $\delta_H(T[j..j'], P_i) = 1$ , for some  $i$ . For each such a fragment, we propagate the appropriate information from the input bitvectors to the output bitvectors.

We begin with constructing the suffix trees of  $T$  and  $T^r$ , including the dictionary structures at each explicit node. Then, we distinguish two cases as follows.

**Exact Occurrences.** For each pattern  $P_i$ , we find its identifier in  $ST_T$  in  $\mathcal{O}(|P_i|)$  time. If the identifier is non-null then we include it in a set  $S$ .

We iterate over every position  $j = 0, 1, \dots, m-1$  and length  $\ell = 1, 2, \dots, \min\{B', m-j\}$ . While we iterate over the lengths, we simultaneously navigate in  $ST_T$  to maintain the identifier of  $T[j..j+\ell]$ . To check if  $T[j..j+\ell] = P_i$  for some  $i$ , we thus need to check if a specific identifier belongs to  $S$ . Recall that the identifiers are integers from  $\{1, 2, \dots, m^2\}$ . To avoid the necessity of paying extra logarithmic factors or using randomization, we answer all such queries together. In more detail, we gather all the queries. Then, we sort the elements of  $S$  and the queries together with radix sort. Then, we scan the obtained sorted list and obtain the answer to each query in linear total time. Finally, if  $T[j..j+\ell] = P_i$  for some  $i$  and  $U_d[j] = 1$ , then we set  $V_d[j+\ell] = 1$ , for  $d = 0, 1$ .

**Occurrences with one Mismatch.** For each pattern  $P_i$ , we iterate over every position  $j = 0, 1, \dots, |P_i| - 1$ , assuming that the mismatch is at position  $j$ . We would like to have access to the identifiers of  $P_i[. . j - 1]$  and  $P_i[j + 1 . .]$ . This can be guaranteed by first navigating in  $ST_T$  to compute the identifier of every prefix  $P_i[. . j]$  in  $\mathcal{O}(|P_i|)$  time, and similarly navigating in  $ST_{T^r}$  to compute the identifier of the reversal of every suffix  $(P_i[j . .])^r$ . After such a preliminary step, for every position  $j = 0, 1, \dots, |P_i| - 1$ , if both identifiers are non-null then we form a pair consisting of the identifier of  $P_i[. . j - 1]$  and the identifier of  $(P_i[j + 1 . .])^r$ . Let  $S$  denote the obtained set of pairs.

We iterate over every position  $j = 0, 1, \dots, m - 1$ , position  $j' = j, j + 1, \dots, m - 1$  and position  $j'' = j', j' + 1, \dots, m - 1$ , where  $T[j . . j'']$  is the considered fragment of  $T$  and  $j'$  is the position of the mismatch. We would like to have access to the identifier of  $(T[j . . j'])^r$  in  $ST_{T^r}$  and the identifier of  $T(j' . . j'')$  in  $ST_T$ . This can be assumed without increasing the time complexity by first iterating over  $j'$  (in any order), then over  $j''$  in the increasing order, and finally over  $j$  in decreasing order, all while simultaneously navigating in  $ST_T$  and  $ST_{T^r}$ , respectively. With the identifiers at hand, we need to check if the pair consisting of the identifier of  $T(j' . . j'')$  and the identifier of  $(T[j . . j'])^r$  belongs to  $S$ . Similarly as for exact occurrences, this is done by answering the queries together with radix sort. Then, if  $\delta_H(T[j . . j''], P_i) \leq 1$ , for some  $i$ , and  $U_0[j] = 1$ , we set  $V_1[j'' + 1] = 1$ .

**Summary.** The algorithm described above consists of the following steps. First, we need to construct the suffix trees of  $T$  and  $T^R$  in  $\mathcal{O}(m)$  time. Constructing the deterministic dictionaries storing the outgoing edges takes  $\mathcal{O}(m(\log \log m)^2)$  time. Second, listing and processing the exact occurrences takes  $\mathcal{O}(mB' + N)$  time. Third, listing and processing occurrences with one mismatch takes  $\mathcal{O}(m(B')^2 + N)$  time. ◀

## B Omitted Proofs

► **Lemma 13.** *All  $Q_i$ s are cyclically equivalent, and every  $Q_i$  is a  $6k$ -period of the text  $T$ .*

**Proof.** We first prove that the periods  $Q_i$  obtained for all the patterns  $P_i$  must be cyclically equivalent. Let  $T_{\text{mid}} := T[\ell/2 . . \ell)$  be the middle part of  $T$ . Since all patterns are of length  $|P| \geq \ell$  and the text is of length  $|T| \leq 1.5\ell$ , all pattern occurrences must cover the middle part of  $T$ . Recall that we assume that every pattern  $P_i$  has some  $2k$ -period  $Q_i$ . By triangle inequality, every  $Q_i$  must be a  $3k$ -period of  $T_{\text{mid}}$ . We will first show that if the strings  $Q_i$  are primitive and of length  $|Q_i| < \ell/100k$ , then they are all cyclically equivalent. Select any two such periods of  $T_{\text{mid}}$ , denoted by  $Q_1$  and  $Q_2$ , and assume (only to avoid clutter) that both of their offsets are equal to 0.

First, assume that  $Q_1$  and  $Q_2$  are not of the same length. Observe that since the size of the combined set of periodic mismatches  $\text{Mis}(T_{\text{mid}}, Q_1^\infty[0 . . \ell/2)) \cup \text{Mis}(T_{\text{mid}}, Q_2^\infty[0 . . \ell/2))$  is at most  $6k$ , there must exist a substring  $T_{\text{sub}}$  of  $T_{\text{mid}}$  that does not contain any such mismatch of length at least

$$|T_{\text{sub}}| \geq \left\lceil \frac{|T_{\text{mid}}| - 6k}{6k + 1} \right\rceil \geq \frac{\ell/2 + 1}{6k + 1} - 1 > \frac{\ell}{14k} - 1.$$

The strings  $Q_1$  and  $Q_2$  are thus exact periods of  $T_{\text{sub}}$ . In addition we have

$$|Q_1| + |Q_2| \leq \ell/100k + \ell/100k < \ell/14k < |T_{\text{sub}}| + 1$$

which by the periodicity lemma of Fine and Wilf [13] induces a period of length  $\gcd(|Q_1|, |Q_2|)$ , and contradicts the assumption that  $Q_1$  and  $Q_2$  are primitive.



In the other case, when  $|Q_1| = |Q_2|$ , assume that  $Q_1 \neq Q_2$ . We would then have

$$\delta_H(Q_1^\infty[0.. \ell/2], Q_2^\infty[0.. \ell/2]) \geq \left\lfloor \frac{\ell/2}{|Q_1|} \right\rfloor \geq \left\lfloor \frac{\ell/2}{\ell/100k} \right\rfloor = 50k.$$

On the other hand, by triangle inequality

$$\delta_H(Q_1^\infty[0.. \ell/2], Q_2^\infty[0.. \ell/2]) \leq \delta_H(Q_1^\infty[0.. \ell/2], T_{\text{mid}}) + \delta_H(T_{\text{mid}}, Q_2^\infty[0.. \ell/2]) \leq 3k + 3k,$$

which again gives us a contradiction and proves that  $Q_1$  must be equivalent to  $Q_2$ .

Since we have assumed that some  $P$  is a  $k$ -mismatch prefix of  $T$  and some  $P'$  is a  $k$ -mismatch suffix of  $T$ , both having  $2k$ -period  $Q$ , it can be proven with similar arguments that  $Q$  is a  $6k$ -period of  $T$ . ◀

► **Lemma 14.**  $\text{Occ}_k^H(P, T) \subseteq \{r + i|Q| : i \in \mathbb{Z}\}$ .

**Proof.** For any  $x \in \text{Occ}_k^H(P, T)$ , by triangle inequality, we have

$$\delta_H(\bar{P}, \bar{T}[x..x + |P|]) \leq \delta_H(\bar{P}, P) + \delta_H(P, T[x..x + |P|]) + \delta_H(T[x..x + |P|], \bar{T}[x..x + |P|])$$

and since

- $\delta_H(\bar{P}, P) \leq 2k$ ,
- $x \in \text{Occ}_k^H(P, T) \Rightarrow \delta_H(P, T[x..x + |P|]) \leq k$ ,
- $\delta_H(T[x..x + |P|], \bar{T}[x..x + |P|]) \leq \delta_H(T, \bar{T}) \leq 6k$

we get

$$\delta_H(\bar{P}, \bar{T}[x..x + |P|]) \leq 9k.$$

Recall that  $\bar{P} = Q^\infty[r..r + |P|]$  and  $\bar{T}[x..x + |P|] = Q^\infty[x..x + |P|]$  both have a primitive period  $Q$  (with offsets  $r$  and  $x$ , respectively). If their offsets are not congruent modulo  $|Q|$ , we can bound the number of mismatches by

$$\delta_H(\bar{P}, \bar{T}[x..x + |P|]) \geq \lfloor |P|/|Q| \rfloor > 9k,$$

which yields a contradiction (the second inequality follows from  $|Q| \leq |P|/128k$ ). Therefore  $x \in \{r + i|Q| : i \in \mathbb{Z}\}$ . ◀