CHAPTER 1

Background

1.1 Introduction

1.1.1 Motivation

This work has been motivated by recent advances of molecular genetics. The human genome has been sequenced in 2001. Also mouse, drosophila, etc. Nowadays # reference model genomes are available in genbank.

Next-generation sequencing has been the second revolution. NGS produces billions of reads for 1000$ dollars. Why should one re-sequence a known genome? Resequencing applications include variant calling, etc. So NGS impacts biomedicine.

Given a set of reads, two approaches are possible: assembly and mapping.

Assembly methods are based on overlaps, de brujin graphs, or...

Read mapping methods work on a previously assembled reference genome.

The typical SNPs analysis pipeline 1.1 consists of...

In this work we focus on read mapping, although many core algorithms considered are also applicable to assembly, as well as to later pipeline stages.

Figure 1.1: *NGS pipeline.*

1.1.2 Fundamental stringology

We now introduce fundamental definitions and problems of stringology, in order to keep the manuscript self-contained. The reader familiar with basic stringology can skip this section and proceed to section **??**.

Definitions

Let us start by defining primitive objects of stringology: alphabets and strings. An alphabet is a finite set of symbols (or characters); a string (or word) over an alphabet is a finite sequence of symbols from that alphabet. We denote the length of a string s by $|s|$, and by ϵ the empty string s.t. $|\epsilon| = 0$. Given an alphabet Σ, we define $\Sigma^0 = \{\epsilon\}$ as the set containing the empty string, Σ^n as the set of all strings over Σ of length n, and $\Sigma^* = \cup_{n=0}^{\infty}\Sigma^n$ as the set of all strings over Σ. Finally, we call any subset of Σ^* a language over Σ.

We now define concatenation, the most fundamental operation on strings. The concatenation operator of two strings is denoted with $\cdot$ and defined as $\cdot : \Sigma^* \times \Sigma^* \to \Sigma^*$. Given two strings, $x \in \Sigma^n$ with $x = x_1 x_2 \dots x_n$, and $y \in \Sigma^m$ with $y = y_1 y_2 \dots y_m$, their concatenation $x \cdot y$ (or simply denoted xy) is the string $z \in \Sigma^{n+m}$ consisting of the symbols $x_1 x_2 \dots x_n y_1 y_2 \dots y_m$.

From concatenation we can derive the notion of prefix, suffix, and substring. A string x is a prefix of y iff there is some string z s.t. $y = x \cdot z$. Analogously, x is a suffix of y iff there is some string z s.t. $y = z \cdot x$. Moreover, x is a substring of y iff there is some string w, z s.t. $y = w \cdot x \cdot z$, and then we say that x occurs within y at position $|w|$.

Example 1.1. These definitions allow us to model basic biological sequences. Let us consider the alphabet consisting of DNA bases: $\Sigma = \{\text{A}, \text{C}, \text{G}, \text{T}\}$. Examples of strings over Σ are $x =$A, $y =$AGGTAC, $z =$TA. For instance, $y \in \Sigma^6$ and $|y| = 6$. Moreover, the concatenation $x \cdot z$ produces ATA. The string x is a prefix of y, and the string z is a substring of y occurring at position 4 in y.

Transformations

The next step is to define the minimal set of edit operations to transform one string into another: substitutions, insertions and deletions. Given two strings x, y of equal length n, the string x can be transformed into the string y by substituting (or replacing) all symbols x_i s.t. $x_i \neq y_i$ into y_i, for $1 \leq i \leq n$. If the given strings have different lengths, insertion and deletion of symbols from x become necessary to transform it into y. Therefore, given any two strings x, y, we define as edit transcript for x, y any finite sequence of substitutions, insertions and deletions transforming x into y.

Example 1.2. TODO: example of edit transcript.

Edit transcripts lead us to the definition of distance functions between strings. The Hamming distance between two strings $x, y \in \Sigma^n$ is defined as the function $d_H : \Sigma^n \times \Sigma^n \to \mathcal{N}$ counting the number of substitutions necessary to transform x into y. More generally, the edit (or Levenshtein) distance between two strings $x, y \in \Sigma^*$ is defined as the function $d_E : \Sigma^* \times \Sigma^* \to \mathcal{N}$ counting the minimum number of edit operation necessary to transform x into y.

Example 1.3. TODO: example of edit and hamming distance.

Edit distance computation

The edit distance problem is to compute the edit distance between two given strings, along with an optimal edit transcript that describes the transformation **?**. The edit distance problem is a minimization problem and can be efficiently computed via dynamic programming (DP). Below we describe the three essential components of the DP approach: the recurrence relation, the DP table, and the traceback.

Given two strings x, y, for all $1 \leq i \leq |x|$ and $1 \leq j \leq |y|$ we define with $d_E(x_{1..i}, y_{1..j})$ the edit distance between their prefixes $x_{1..i}$ and $y_{1..j}$. The base conditions of the recurrence relation are:

$$d_E(\epsilon, \epsilon) = 0 \tag{1.1}$$

$$d_E(x_{1..i}, \epsilon) = i \text{ for } 1 \leq i \leq |x| \tag{1.2}$$

$$d_E(\epsilon, y_{1..j}) = j \text{ for } 1 \leq j \leq |y| \tag{1.3}$$

and the recursive case is defined as follows:

$$d_E(x_{1..i}, y_{1..j}) = min\{d_E(x_{1..i-1}, y_{1..j}) + 1, d_E(x_{1..i}, y_{1..j-1}) + 1, d_E(x_{1..i-1}, x_{1..j-1}) + \delta(x_i, y_j)\} \tag{1.4}$$

The recurrence relation can be computed in time $\mathcal{O}(|x| \cdot |y|)$ using a table of $(n+1) \times (m+1)$ cells. However only $\mathcal{O}(min\{n, m\})$ space is required.

The table can be filled in four different ways: column-wise, row-wise, diagonal-wise or antidiagonal-wise.

Figure 1.2: *DP table representing the computation of the edit distance $d_E(x_{1..5}, y_{1..4})$.*

An optimal alignment can be computed in time $\mathcal{O}(n + m)$.

Alignments

An alignment is a way of visualizing a transformation between two strings.

The problem of finding the optimal alignment between two strings is the dual of the edit distance problem.

Example 1.4. TODO: example of alignment.

1.2 Overview of string matching

1.2.1 Problem definition

We can now define exact string matching, perhaps the most fundamental problem in stringology. Given a string p called the pattern and a longer string t called the text, the exact string matching problem is to find all occurrences, if any, of pattern p into text t **?**. This problem has been extensively studied from the theoretical standpoint and is well solved in practice. The reader is referred to **?** for an extensive treatment of the subject.

The definition of distance functions between strings let us generalize exact string matching into a more challenging problem: approximate string matching. Given a text t, a pattern p, and a distance threshold $k \in \mathcal{N}$, the approximate string matching (a.s.m.) problem is to find all occurrences of p into t within distance k. The a.s.m. problem under

the Hamming distance is commonly referred as the k-mismatches problem and under the edit distance as the k-differences problem.

Existing methods to solve approximate string matching problems can be classified in three categories: online, indexed and filtering. An extensive survey on online methods is provided by **?**, while a more succint survey on indexed methods is given in **?**. In the following of this section we give only a brief and non-exhaustive overview of the fundamental techniques, and discuss their advantages and limitations. This overview serves as an introduction to more involved methods presented in chapter **??**.

1.2.2 Online methods

Introduce and motivate online methods. Online methods work by scanning the text from left to right (or right to left). Good: require in the best case memory proportional to the pattern length. Bad: worst case runtime at least linear in the text size.

Automata

Exact search of one pattern. Boyer-moore automaton.

Exact search of multiple patterns. Aho-corasick automaton.

Approximate search of one pattern. Ukkonen automaton.

Dynamic programming

The dynamic programming algorithm **??** to compute the edit distance of two strings can be easily turned into a pattern matching algorithm. Since an occurrence of the pattern can start and end anywhere in the text, a.s.m. consists of computing the edit distance between the pattern and all substrings of the text. The problem can be thus solved by computing the edit distance between the text and the pattern without penalizing leading and trailing deletions in the text.

Let pose $x = t$ and $y = p$ and consider equations **??**. Since an occurrence of the pattern can start anywhere in the text, we change the initialization of the top row as:

$$d_E(\epsilon, y_{1..j}) = 0 \text{ for } 1 \leq j \leq |y| \tag{1.5}$$

and since an occurrence of the pattern can end anywhere in the text, we check every cell of the bottom row for the condition:

$$d_E(x_{1..m}, y_{1..j}) \leq k \text{ for } 1 \leq j \leq |y|. \tag{1.6}$$

Figure 1.3: *DP table representing the match of $p = ...$ in $t = ...$*

1.2.3 Indexed methods

Motivate indexed methods. Good: runtime independent of the text size. Bad: require memory linear in the text size; exponential in k.

No matter how fast online search can be, these approaches quickly become impractical. Since the text is static and searched frequently, we decide to preprocess it. We build an index of the text beforehand and use it to speed up subsequent searches. To this intent, we first introduce *suffix trees*, optimal data structures to index strings. Later on, we consider algorithms solving string matching problems on suffix trees.

Suffix tree and suffix trie

The suffix tree [**?**] is a lexicographically ordered tree data structure representing all suffixes of a string. Assume w.l.o.g. a string s of length n, padded with a *terminator symbol* \$ not being part of the string alphabet Σ[1]. The suffix tree $\mathbb{S}$ of the string s has one node designated as the root and n leaves, one per suffix, labeled with numbers from 1 to n. Each internal node has more than one child, and each edge is labeled with a non-empty substring of s. Each path from the root to a leaf i spells the suffix $s_{i..n}$. Figure **??** illustrates.

Figure 1.4: *Suffix tree.*

In the following of this manuscript we consider w.l.o.g. *suffix tries* instead of suffix trees. On suffix tries, internal nodes can have only one child and each edge is labeled by one single character. This fact simplifies the exposition of all given algorithms without affecting their runtime complexity nor their result. However, we remark that all given algorithms can be generalized to work on trees. Therefore, from now on we assume the text t to be indexed using a suffix trie $\mathbb{T}$.

Figure 1.5: *Suffix trie.*

Given a node x, we denote with $label(x)$ the label of the edge entering into x, with $\mathcal{C}(x)$ the set of children of x being internal nodes[2], with $\mathcal{E}(x)$ the set of children of x being leaves[3].

Exact search

Using the suffix trie $\mathbb{T}$ of a text t, we can find all occurrences of a pattern p into t in optimal time $\mathcal{O}(|p|)$ and independently of $|t|$. We use the property of the suffix trie that each path from the root to an internal node spells a different unique substring of t and consequently all equal substrings of t are compressed in a single path. Algorithm **??** locates a pattern p by starting in the root node of $\mathbb{T}$ and following the path spelling the pattern. If we end up in a node x, each leaf $l_x \in \mathcal{E}(x)$ points to a distinct substring of t that is equal to p.

[1] The terminator symbol is necessary to ensure that no suffix $s_{i..n}$ is a prefix of another suffix $s_{j..n}$.

[2] Entering edges of internal nodes are labeled with symbols in Σ.

[3] Entering edges of leaves are labeled with terminator symbols.

Algorithm 1.1 Exact search on a suffix trie.

```
1: procedure SEARCH(s, q)
2:     report E(s) × E(q)
3:     for all c_q ∈ C(q) do
4:         if ∃ c_s ∈ C(s) : label(c_s) = label(c_q) then
5:             SEARCH(c_s, c_q)
6:         end if
7: end procedure
```

Figure 1.6: *Exact pattern matching on a suffix tree.*

Backtracking k-mismatches

By backtracking [**??**] on a suffix tree we can find all occurrences in t within distance k from a pattern p, in average time sublinear in $|t|$ [**?**]. A top-down traversal on the suffix trie $\mathbb{T}$ spells incrementally all distinct substrings of t. While traversing each branch of the trie, we incrementally compute the distance between the query and the spelled string. If the computed distance exceeds k, we stop the traversal and proceed on the next branch. Conversely, if we completely spelled the pattern p, and we ended up in a node x, each leaf $l_x \in \mathcal{E}(x)$ points to a distinct string $S_x \in \mathcal{S}$ that is within distance k of p.

Algorithm 1.2 k-mismatches on a suffix trie.

```
1: procedure SEARCH(s, q)
2:     report E(s) × E(q)
3:     for all c_q ∈ C(q) do
4:         if ∃ c_s ∈ C(s) : label(c_s) = label(c_q) then
5:             SEARCH(c_s, c_q)
6:         end if
7: end procedure
```

Backtracking k-differences

We can compute k-differences on a suffix tree in two different ways: by explicitly enumerating errors with a five-fold recursion on the suffix trie, or by computing the DP matrix on the suffix trie.

1.2.4 Filtering methods

Why filtering

Motivate filtering methods.

Figure 1.7: *k-mismatches on a suffix tree.*

Algorithm 1.3 k-differences on a suffix trie.

Pigeonhole principle

Exact seeds

A simple solution to the problem is provided by a filtering algorithm proposed in **?** which reduces an approximate search into smaller exact searches. A pattern p is partitioned into $k + 1$ non-overlapping seeds which are searched in t with the help of $\mathbb{T}$. Since each edit operation can affect at most one seed, for the pigeonhole principle each approximate occurrence of p in t contains an exact occurrence of some seed. However the converse is not true, consequently we must verify whether any candidate location induced by an occurrence of some seed corresponds to an approximate occurrence of p in t.

Filtration specificity in terms of candidate locations to verify is strongly correlated to seed length. Since we want to maximize the length of the shortest seed, we let the minimum seed length be $\lfloor |p|/(k+1) \rfloor$. If we want to improve filtration specificity by increasing seed length, we can resort to approximate seeds.

Approximate seeds

A more involved filtering algorithm proposed in **?** reduces an approximate search into smaller approximate searches. We partition p into $s \leq k + 1$ non-overlapping seeds. According to the pigeonhole principle each approximate occurrence of p in t then contains an approximate occurrence of some seed within distance $\lfloor k/s \rfloor$.

Approximate seeds are searched via backtracking on $\mathbb{T}$. We search $(k \bmod s) + 1$ seeds within distance $\lfloor k/s \rfloor$ and the remaining seeds within distance $\lfloor k/s \rfloor - 1$. To prove full-sensitivity it suffices to see that, if none of the seeds occurs within its assigned distance, the total distance must be at least $s \cdot \lfloor k/s \rfloor + (k \bmod s) + 1 = k + 1$. Hence all approximate occurrences of p in t within distance k will be found.

q-Gram lemma

1.3 Related problems

1.3.1 Local similarity search

Define score and scoring scheme.

Define local similarity.

Online methods

Give dynamic programming solution.

Figure 1.8: *k-differences on a suffix tree.*

Figure 1.9: *Filtration with exact seeds.*

Indexed methods

Backtracking over substring index. BWT-SW.

Filtering methods

SWIFT/Stellar is based on the q-gram lemma.

1.3.2 Dictionary search

Dictionary search is a restriction of string matching. Given a set of database strings $\mathcal{S}$ and a query string q find all strings in $\mathcal{S}$ within distance k from q. Note that strings in $\mathcal{S}$ usually have length similar to $|q|$, as $||s| - |q|| \leq k$ is a necessary condition for $d_E(s, q) \leq k$.

Online methods

The problem can be solved by checking whether $d_E(s, q) \leq k$ for all $s \in \mathcal{S}$. Answering the question whether the distance $d_E(s, q) \leq k$ is an easier problem than computing the edit distance $d_E(s, q)$: a band of size $k + 1$ is sufficient.

Indexed methods

Using a radix tree $\mathbb{S}$ we can find all strings in $\mathcal{S}$ equal to a query string q, in optimal time $\mathcal{O}(|q|)$ and independently of $||\mathcal{S}||$.

Algorithm 1.4 Exact dictionary search on a radix trie.

Filtering methods

1.3.3 Overlaps computation

Define problem.

Online methods

DP solution.

Indexed methods

Indexed solution, exact and approximate.

Figure 1.10: *Filtration with approximate seeds.*

Algorithm 1.5 Approximate dictionary search on a radix trie.