МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №5

по дисциплине «Построение и анализ алгоритмов»

Тема: Ахо-Корасик

Студент гр. 3343 _____ Атоян М. А.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2025

Цель работы.

Изучить принцип работы алгоритма Кнута-Морриса_Пратта. Написать функцию, вычисляющую для каждого элемента строки максимальное значение длины префикса и с помощью данной функции решить поставленные задачи. А именно написать программу, осуществляющую поиск вхождений подстроки в строку, а также программу, определяющую, являются ли строки циклическим сдвигом друг друга, найти индекс начала вхождения второй строки в первую.

Задание №1.

Разработайте программу, решающую задачу точного поиска набора образцов.

Вход:

Первая строка содержит текст (T,1≤|T|≤100000*T*,1≤|*T*|≤100000).

Вторая - число n*n* (1≤n≤30001≤*n*≤3000), каждая следующая из n*n* строк содержит шаблон из набора P={p1,…,pn}1≤|pi|≤75*P*={*p*1,…,*pn*}1≤|*pi*|≤75

Все строки содержат символы из алфавита {A,C,G,T,N}{*A*,*C*,*G*,*T*,*N*}

Выход:

Все вхождения образцов из P*P* в T*T*.

Каждое вхождение образца в текст представить в виде двух чисел - i *i* p*p*

Где i*i* - позиция в тексте (нумерация начинается с 1), с которой начинается вхождение образца с номером p*p*

(нумерация образцов начинается с 1).

Строки выхода должны быть отсортированы по возрастанию, сначала номера позиции, затем номера шаблона.

Sample Input:

NTAG

3

TAGT

TAG

T

Sample Output:

2 2

2 3

Задание №2.

Используя реализацию точного множественного поиска, решите задачу точного поиска для одного образца с *джокером*.

В шаблоне встречается специальный символ, именуемый джокером (wild card), который "совпадает" с любым символом. По заданному содержащему шаблоны образцу P*P* необходимо найти все вхождения PP в текст TT.

Например, образец ab??c?a*b*??c? с джокером ?? встречается дважды в тексте *xabvccbababcax*.

Символ джокер не входит в алфавит, символы которого используются в T*T*. Каждый джокер соответствует одному символу, а не подстроке неопределённой длины. В шаблон входит хотя бы один символ не джокер, т.е. шаблоны вида ??? недопустимы.
Все строки содержат символы из алфавита {A,C,G,T,N}{*A*,*C*,*G*,*T*,*N*}

Вход:

Текст (T,1≤|T|≤100000*T*,1≤|*T*|≤100000)

Шаблон (P,1≤|P|≤40*P*,1≤|*P*|≤40)

Символ джокера

Выход:

Строки с номерами позиций вхождений шаблона (каждая строка содержит

только один номер).

Номера должны выводиться в порядке возрастания.

Sample Input:

ACTANCA

A$$A$

$

Sample Output:

1

Вариант 1. На месте джокера может быть любой символ, за исключением заданного.

Описание алгоритмов.

Описание алгоритма Ахо-Корасик.

Алгоритм создает префиксное дерево из букв искомых подстрок. Затем в полученном дереве ищутся суффиксные ссылки. Суффиксная ссылка вершины u – это вершина v, такая что строка v является максимальным суффиксом строки u. Для корня и вершин, исходящих из корня, суффиксной ссылкой является корень. Для остальных вершин осуществляется переход по суффиксной ссылке родителя и, если оттуда есть ребро с заданным символом, суффиксная ссылка назначается в вершину, куда это ребро ведет. Далее создаются терминальные ссылки – такие суффиксные ссылки, которые ведут в вершину, которая является терминальной.

Текст, в котором нужно найти подстроки побуквенно передается в автомат. Начиная из корня, автомат переходит по ребру, соответствующему переданному символу. Если нужного ребра нет, переходит по ссылке. Если встреченная вершина является терминальной, значит была встречена подстрока. Если

найдено совпадение нужно пройти по терминальным ссылкам, если они не None, чтобы вывести все шаблоны заканчивающиеся на этом месте. Номер подстроки (подстрок) хранится в поле *terminate* вершины. В ответ сохранятся индекс, на котором началась эта подстрока в тексте и сам номер подстроки.

Сложность по времени:

Т.к при построении префиксного дерева запускается цикл по длине каждой подстроки (суммарная длина подстрок - n), и из каждой вершины может исходить максимум k ребер (где k – размер алфавита), то построение префиксного дерева происходит за O(n*k)

Алгоритм в цикле проходит по тексту длины s: O(s)

Итого: O(n*k + s)

Сложность по памяти:

Алгоритм создает префиксное дерево с n вершинами, каждая вершина хранит массив вершин, инцидентных ей, размером k (k – размер алфавита).

Итого: O(n*k) Описание алгоритма для нахождения шаблонов с маской.

Описание модифицированного алгоритма.

Алгоритм тот же, но в качестве подстрок берутся кусочки шаблона, разделенные джокером, запоминаются позиции полученных подстрок в исходном шаблоне. Создается массив C длины s, где s – длина текста, где ищется шаблон. При нахождении подстроки, в массиве C увеличивается на единицу число по индексу, соответствующему возможному началу шаблона. Индекс высчитывается по формуле: текущий индекс - (длина найденной подстроки - 1) - (позиция подстроки в шаблоне -1). Затем проходим по полученному массиву, каждый i для которого C[i] = количеству подстрок, является вероятным началом шаблона. В соответствии с индивидуализацией, для каждого найденного шаблона проверяются буквы, стоящие на месте джокера. Если не было встречено запрещенного символа, найденный шаблон добавляется в ответ.

Сложность по времени для модифицированного алгоритма:

Затраты по времени такие же как в обычном алгоритме, но дополнительно проход по массиву C длины s: Итого: O(n*k + s + s) = O(n*k + s + t)

Сложность по памяти для модифицированного алгоритма:

Затраты по памяти такие же как в обычном алгоритме, но дополнительно создается массив C длины s. Затраты по памяти O(n*k + s)

Описание функций.

1. **Структура Node**: Представляет узел в префиксном дереве (trie), содержащий ссылки на родителя, детей, суффиксные и терминальные ссылки, а также информацию о терминальности узла и его имени.
2. **func buildTrie(patterns []string) (root *Node)**: Создает префиксное дерево на основе переданных шаблонов, добавляя узлы для каждого символа шаблона и отмечая терминальные узлы.
3. **func buildAutomaton(root *Node)**: Обходит дерево в ширину и создает суффиксные и терминальные ссылки для всех узлов, начиная с корня.
4. **func findMatches(text string, root *Node, patternLengths map[int]int) (results [][2]int)** Реализует алгоритм Ахо-Корасик для поиска всех вхождений шаблонов в тексте, используя суффиксные и терминальные ссылки.

Тестирование.

Входные данные	Ответ	Комментарий
NTAG 3 TAGT TAG T	2 2 2 3	Верно
ACCGTACA 2 AC GT	1 1 4 2 6 1	Верно
ACGT 3 ACGT CG GT	1 1 2 2 3 3	Верно

Таблица 1 – Тестирование алгоритма Ахо-Корасик

Входные данные	Ответ	Комментарий
ACTANCA A$$A$ $ G	1	Верно
ACACAA ACXA X Y	3	Верно
ACGANGAAAT A$G $ C	4	Верно

Таблица 1 – Тестирование алгоритма поиска с джокером

Результат работы программы с отладочным выводом для первого задания (см. рис 1, 2, 3).

Исследование.

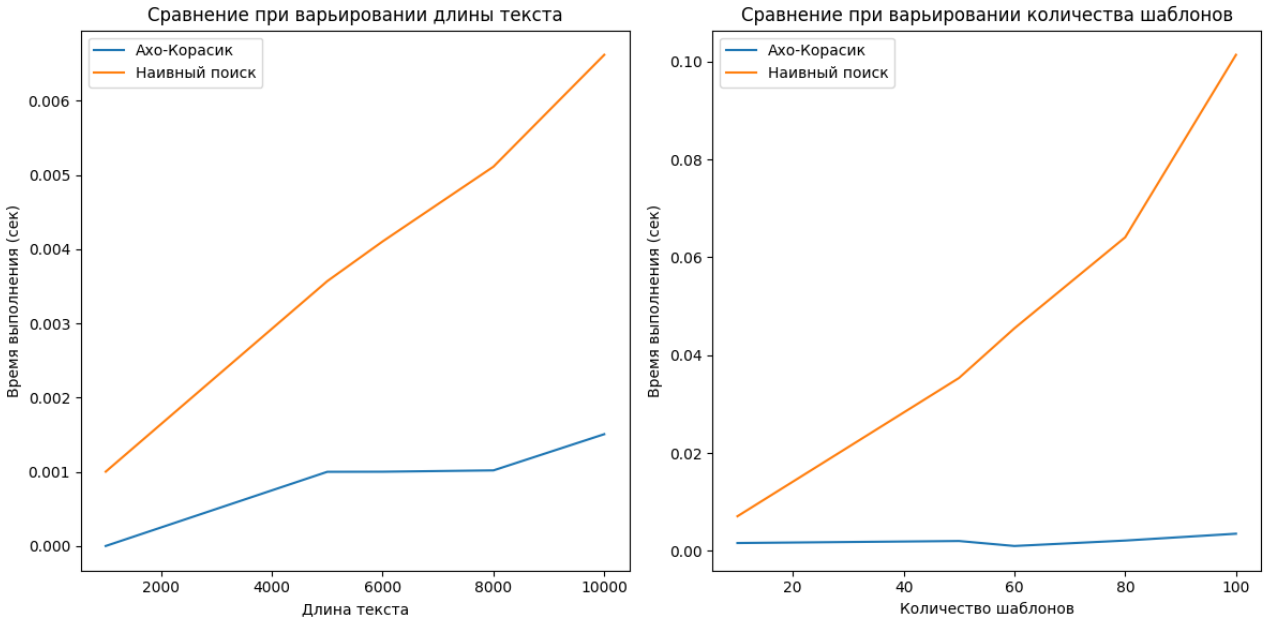

Рисунок 8 – Сравнение алгоритма Ахо-Корасик и наивного поиска

Можно сделать вывод, что Ахо-Корасик выполняется значительно быстрее, чем наивный алгоритм.

Выводы.

Изучен принцип работы алгоритма Ахо-Корасик. Написаны программы, корректно решающие задачу поиска набора подстрок в строке, в также программа поиска подстроки с джокером.

ПРИЛОЖЕНИЕ А
ИСХОДНЫЙ КОД ПРОГРАММЫ

Название файла: var1/main.go

```
package main

import (
	"bufio"
	"fmt"
	"os"
	"sort"
)

const alphabetSize = 5

type Node struct {
	children [alphabetSize]*Node
	failure  *Node
	outputs  []int
}

type FixedChar struct {
	Position int
	Char     byte
}

func charToIndex(c byte) (int, error) {
	switch c {
	case 'A':
		return 0, nil
	case 'C':
		return 1, nil
	case 'G':
		return 2, nil
	case 'T':
		return 3, nil
	case 'N':
		return 4, nil
	default:
		return 0, fmt.Errorf("invalid character: %c", c)
	}
}
```

```
func buildTrie(fixed []FixedChar) (*Node, error) {
	root := &Node{}
	for _, fc := range fixed {
		current := root
		idx, err := charToIndex(fc.Char)
		if err != nil {
			return nil, err
		}

		if current.children[idx] == nil {
			current.children[idx] = &Node{}
		}
		current = current.children[idx]
		current.outputs = append(current.outputs, fc.Position)
		fmt.Printf("Добавлен фиксированный символ %c на позиции %d
в узел %p\n",
			fc.Char, fc.Position, current)
	}
	return root, nil
}

func buildFailureLinks(root *Node) {
	queue := make([]*Node, 0)

	for i := 0; i < alphabetSize; i++ {
		if child := root.children[i]; child != nil {
			child.failure = root
			queue = append(queue, child)
			fmt.Printf("Инициализация failure для узла %p ->
корень\n", child)
		}
	}

	for len(queue) > 0 {
		u := queue[0]
		queue = queue[1:]

		for i, child := range u.children {
			if child == nil {
				continue
			}

			fail := u.failure
			for fail != nil && fail.children[i] == nil {
```

```
                        fail = fail.failure
                    }

                    if fail == nil {
                        child.failure = root
                    } else {
                        child.failure = fail.children[i]
                    }

                    child.outputs = append(child.outputs,
child.failure.outputs...)
                    queue = append(queue, child)

                    fmt.Printf("Установка failure для узла %p -> %p
(outputs: %v)\n",
                        child, child.failure, child.outputs)
                }
            }
        }

        func main() {
            scanner := bufio.NewScanner(os.Stdin)
            scanner.Buffer(make([]byte, 0, 1024*1024), 1024*1024*10)

            scanner.Scan()
            text := scanner.Text()

            scanner.Scan()
            pattern := scanner.Text()

            scanner.Scan()
            wildcardChar := scanner.Text()[0]

            scanner.Scan()
            excludedChar := scanner.Text()[0]

            var fixed []FixedChar
            for i := 0; i < len(pattern); i++ {
                if pattern[i] != wildcardChar {
                    fixed = append(fixed, FixedChar{i, pattern[i]})
                }
            }

            var wildcardPositions []int
```

```
	for j := 0; j < len(pattern); j++ {
		if pattern[j] == wildcardChar {
			wildcardPositions = append(wildcardPositions, j)
		}
	}

	if len(fixed) == 0 {
		fmt.Println("Ошибка: шаблон содержит только джокеры")
		return
	}

	fmt.Println("\n=== Этап 1: Построение префиксного дерева ===")
	root, err := buildTrie(fixed)
	if err != nil {
		fmt.Println("Ошибка построения дерева:", err)
		return
	}

	fmt.Println("\n=== Этап 2: Построение failure-ссылок ===")
	buildFailureLinks(root)

	fmt.Println("\n=== Этап 3: Поиск кандидатов ===")
	counters := make(map[int]int)
	current := root
	patternLen := len(pattern)

	for textPos, c := range text {
		idx, err := charToIndex(byte(c))
		if err != nil {
			continue
		}

		for current != nil && current.children[idx] == nil {
			current = current.failure
		}

		if current == nil {
			current = root
		} else {
			current = current.children[idx]
		}

		fmt.Printf("Обработан символ %c на позиции %d, текущий
узел: %p\n",
```

```
                    c, textPos+1, current)

                for _, p := range current.outputs {
                    i := textPos - p
                    if i >= 0 && i+patternLen <= len(text) {
                        counters[i]++
                        fmt.Printf("Найден кандидат на позиции %d
(счетчик: %d)\n", i+1, counters[i])
                    }
                }
            }

            fmt.Println("\n=== Этап 4: Проверка кандидатов ===")
            var results []int
            required := len(fixed)
            for i, cnt := range counters {
                if cnt == required {
                    valid := true
                    for _, f := range fixed {
                        if i+f.Position >= len(text) ||
text[i+f.Position] != f.Char {
                            valid = false
                            break
                        }
                    }
                    if valid {
                        for _, pos := range wildcardPositions {
                            if text[i+pos] == excludedChar {
                                valid = false
                                break
                            }
                        }

                        results = append(results, i+1)
                        fmt.Printf("Подтверждено вхождение на позиции
%d\n", i+1)

                    }
                }
            }

            fmt.Println("\n=== Результаты ===")
            sort.Ints(results)
            for _, pos := range results {
```

```
            fmt.Println(pos)
        }
    }
```