Dissertation Title

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
vorgelegt von

Enrico Siragusa

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Berlin 201X

Datum des Disputation: ***XX.XX.201X***

Gutachter:
Prof. Dr. Knut Reinert, *Freie Universität Berlin, Deutschland*
Prof. Dr. XXX XXX, *XXX, XXX*

Abstract

Bla bla bla.

CONTENTS

Part I

APPROXIMATE STRING MATCHING

Chapter 1 Background

1.1 Introduction

1.1.1 Problem definition

1.1.2 Bioinformatics applications

1.2 Overview of existing methods

1.2.1 Online methods

1.2.2 Indexed methods

1.2.3 Filtering methods

1.3 Related problems

1.3.1 Local similarity search

1.3.2 Dictionary search

1.3.3 Overlaps computation

Chapter

2

Online Methods

CHAPTER 3

Indexed Methods

3.1 Classic Full-Text Indices

3.1.1 Trie

3.1.2 Suffix trie and suffix tree

3.1.3 Suffix array

3.1.4 q-Gram index

3.2 Compressed Full-Text Indices

3.2.1 Burrows-Wheeler transform

3.2.2 FM-index

3.2.3 Rank dictionaries

3.3 Backtracking

3.3.1 Pruning methods

3.3.2 Multiple backtracking

Chapter 4 Filtering Methods

4.1 q-Gram filters

4.1.1 Exact seeds

4.1.2 Gapped seeds

4.2 Factor filters

4.2.1 Exact seeds

4.2.2 Approximate seeds

4.3 Suffix filters

Part II

APPLICATIONS

CHAPTER 5

Read Mapping

5.1 Related work

5.1.1 Best mappers

5.1.2 All mappers

5.2 The Masai mapper

5.2.1 Single-end mapping

5.2.2 Paired-end mapping

5.3 The Masai 2 mapper

5.3.1 Single-end mapping

5.3.2 Paired-end mapping

5.3.3 Parallelization

5.3.4 Hardware acceleration

5.4 Experimental results

5.4.1 Comparison of filtration strategies

5.4.2 Rabema benchmark results

5.4.3 Variant detection results

5.4.4 Runtime results

5.5 Discussion

CHAPTER

6 String Similarity Search / Join

6.1 The competition

6.2 Our method

6.2.1 Implementation

6.2.2 Parallelization

6.3 Related methods

6.4 Experimental results

6.4.1 Search

6.4.2 Join

6.5 Discussion

Appendix A Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Enrico Siragusa
October 6, 2013

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF NOTATIONS

bp	base pair, character of the alphabet {A,C,G,T}	??
Mb	megabase, 1 million base pairs or characters	??
MB	megabyte, 1 MB = 1024 kB = 1,048,576 byte	??
GB	gigabyte, 1 GB = 1024 MB	??
Σ, Ψ, Φ	finite alphabets	??
Σ^*	set of all possible strings over the alphabet Σ	??
Σ^n	set of all possible strings over the alphabet Σ with length n	??
ϵ	empty string	??
$\lvert s \rvert$	length of string s	??
$s[i]$	character of s at position i (counting from 0)	??
$s_i, \mathrm{suf}(s, i)$	suffix of s beginning at position i	??
$[i..j]$	set of integers $i, i + 1, \ldots, j$	??
$[i..j)$	set of integers $i, i + 1, \ldots, j - 1$	??
$\mathbb{N}_0$	set of non-negative integers	??
$\prec, \preceq$	strict and non-strict substring relation	??
$\langle \ldots \mid \ldots \rangle$	definition of a string analogous to the set notation	??
lcp $\mathcal{S}$	longest common prefix of a set $\mathcal{S}$ of strings	??
$<_{\text{lex}}$	lexicographical order	??
$<_q$	lexicographical prefix order, compares only prefixes of length q	??
$\$, \$^j$	(virtual) sentinel characters to well-define the suffix tree	??
$concat(v)$	edge label concatenation on the path from root to suffix tree node v	??
$\overline{s}$	suffix tree node whose edge label concatenation is s	??
$\mathrm{rank}(a)$	rank of character a in the underlying alphabet	??
$\mathtt{R}, \mathtt{D}, \mathtt{I}$	edit operations that replace, delete, or insert a character	??
$\lVert T \rVert_{\mathrm{E}}$	number of edit operations in transcript T	??
$\lVert T \rVert_{\mathrm{R}}$	number of matches, replacements, and deletions in transcript T	??
$\mathcal{R}$	sequenced reads, set of strings	??
G	reference sequence, string	??
$\mathcal{D}$	database, set of strings	??
$\mathrm{freq}(\phi, \mathcal{D})$	absolute number of strings in $\mathcal{D}$ that contain ϕ at least once	??
$\mathrm{supp}(\phi, \mathcal{D})$	relative number of strings in $\mathcal{D}$ that contain ϕ at least once	??