Suffix Array  
User Guide

Tal E. Cohen

**Table of contents**

[1. Overview 3](#_Toc485835692)

[2. Acronyms and glossary 3](#_Toc485835693)

[3. Versions 3](#_Toc485835694)

[4. Introduction 4](#_Toc485835695)

# Overview

In this project our aim was to develop an optimized approach for storing and analyzing data concerning sub-sequences within char sequences of given text files. The Suffix Array we implemented in this project allows handling large files, and is divided into two entities: a builder that takes most of the runtime and an analyzer that offers different types of analysis to be performed on the output of the builder in a relatively short runtime.

Both programs were tested on generic text files and specific DNA files alike. The developed software (builder, analyzer and classifier) allows classification of genetic diseases by comparing a subject’s DNA data to reference groups of people representing various conditions of genetic disease immunity.

# Acronyms and glossary

|  |  |
| --- | --- |
|  | Meaning |
| SA | Suffix Array |
| .saf | SA file extension of each of the SA’s splits, saved as compressed references to the input |
| .sai | SA input extension of the compressed saved input |
| Split | Sub-tree of a SA independent of other sub-trees |

# Versions

Document versions

|  |  |  |
| --- | --- | --- |
| # | Description | Date |
| 1.0 | CS B.Sc. final project submission | June 21st 2017 |

Software versions

|  |  |  |
| --- | --- | --- |
| # | Description | Date |
| 1.5 | SABuilder – builds and saves splits of Suffix Array | June 16th 2017 |
| 1.5 | SAAnalyser – generates Infixes out of splits and ranks them | June 16th 2017 |

# Introduction

The software package consists of three programs:

* SuffixArrayBuilder

A program that receive as input text files in which each line is a sequence of characters and the analysis is done on sub-sequences of these lines. The output of this program is a representation of the SA structure in the form of a series of binary files (.saf) along with a compressed input file (.sai).

* SuffixArrayAnalyzer

A program that receives as input the output folder of the builder (containing one .sai file and several .saf files) and analyzes each sub-sequence in order to generate a “top” list that meets selected criteria (currently Z-score or count).

* Classifier

This part is documented in a separate guide.

In the current version the input should be raw text files in which all characters are significant (an additional utility is available for extracting the relevant fields from standard DNA files).

The tested capacity of the current software is up to 1.5 million lines of up to 1,000 characters and the tested performance for a 0.5G input file ranges between 30m to 60m runtime (depending on hardware).

The output of the SuffixArrayAnalyser includes:

1. Top list of Infixes, using user-selected criteria (Z-score, count, etc).
2. Optional lists of user-selected length of:
   1. Infixes with a count over user-defined threshold.
   2. Infixes with a Z-score over user-defined threshold.