

RNA-Seqlyze

Software Requirement Specification

Professor:

Prof. Dr. Georg Lipps

Student:

Patrick Pfeifer

Bachelor Thesis

Appendix 1

Version	Author	Comment	Date
0.1	Patrick Pfeifer	Document created	25. Mai 2012

Contents

1	Introduction	3
1.1	Purpose	3
1.2	Project Scope	3
1.3	Intended Audience and Structure of this Document	3
2	Overall Description	3
2.1	Product Environment	3
2.2	Product Features	3
2.3	User Characteristics	4
3	System Features	4
3.1	Priorities	4
3.2	RNA-seq Data Analysis	4
3.3	Custom Track Generation	4
4	Nonfunctional Requirements	5
4.1	Quality Requirements	5
4.1.1	Documentation	5
4.1.2	Usability	5
4.1.3	Reliability and Maintainability	5
4.1.4	Performance and Efficiency	5
4.1.5	Security	5
4.2	Interface Requirements	5
5	Appendices	6
5.1	Glossary	6

1 Introduction

1.1 Purpose

This document details the software requirements specification for the RNA-Seqlyze next-generation RNA-sequencing analysis software. It defines the intended use of the software and lists the required features.

1.2 Project Scope

The RNA-Seqlyze web application is the main product of the RNA-Seqlyze free software project. The project was started by Patrick Pfeifer as part of his bachelor thesis at the University of Applied Sciences and Arts Northwestern Switzerland FHNW at the School of Life Sciences.

RNA-seq data is of increasing importance to the analysis of prokaryotes. The intended use of this software is to aid researchers in analyzing data generated by next-generation RNA sequencing methods (RNA-seq). Specifically, the software is supposed to improve the annotation of existing genome data.

1.3 Intended Audience and Structure of this Document

All parties involved in the project may refer to this document as the definite source of information concerning it.

Software users will generally be most interested in chapters number two and three. They contain information regarding the available features, with detailed descriptions of the expected in- and outputs.

Developers planning to contribute code are strongly encouraged to read chapter four as well, as it contains important information about the reliability-, performance- and interface-requirements and -standards that must be fulfilled.

2 Overall Description

2.1 Product Environment

RNA-seq - the application of next-generation deep sequencing methods to DNA transcripts - carries the potential to deliver new insights into the organization and functionality of an organisms genome.

To achieve this, the large amounts of data produced by one of the various next generation sequencing (NGS) platforms available on the market, have to be analyzed in detail. Short reads of only a few dozen base pairs (bp) have to be mapped to a reference genome. The total number of reads mapped at each position in the genome is counted to produce a transcript coverage "signal". The deeper the coverage, the stronger, roughly, the gene is expressed in the organism.

2.2 Product Features

The envisioned web application is supposed to automate the process of analyzing the data generated by NGS platforms deployed for RNA-seq experiments with prokaryotes. The data will be made available to the application in FASTQ or SRA

format. In case SAM/BAM or coverage data files have already been produced, they can be made available to the application as well.

The RNA-Seqlyze software processes this data using a novel algorithm. It generates a list of transcribed sequences and assigns a confidence score to each of those regions. The score will be computed based on the provided input data as well as various other relevant information relating to the studied organism. This auxiliary data incorporated into the processing algorithms has been collected beforehand and is readily available to the application. Specifically, the application will include predictions for shine-dalgarno sequences, rho-independent terminators, operons (polycistronic transcripts) and, optionally, promoter sites.

2.3 User Characteristics

It is assumed that a researcher using the application has basic IT-skills and is used to interacting with popular computer interfaces. The handling of the the RNA-Seqlyze web application shall be easy for him.

He has a good understanding of the technology applied to generate the data that he wishes to analyze. A fair amount of specialized knowledge and fluency with the most popular technical terms *will be required* to use the application in an efficient manner.

3 System Features

3.1 Priorities

The priorities are assigned by the client and have the following meanings:

- | | |
|--------|---|
| high | This feature is indispensable and absolutely necessary for the correct functioning of the product. It must be implemented. |
| medium | This feature is not indispensable but makes a substantial contribution to the usability of the product. It should be implemented. |

3.2 RNA-seq Data Analysis

Description and Priority

TBD

Priority: **high**

Functional Requirements

TBD

3.3 Custom Track Generation

Description and Priority

TBD

Priority: **medium**

Functional Requirements

TBD

4 Nonfunctional Requirements

4.1 Quality Requirements

4.1.1 Documentation

The source code shall be documented, such that modifications or additional modules could be integrated by third-party developers. Furthermore, an extensive user manual shall be composed.

4.1.2 Usability

The application's user interface shall be easy to use and follow current conventions. When using the software, the user shall constantly be informed what the current state of the application is, which steps have been already carried out and which ones are to be carried out next.

4.1.3 Reliability and Maintainability

An error-free execution of all functions of the software shall be achieved by writing tests for all parts (unit tests) as well as for the whole system (integration tests). The coverage of the unit tests shall reach at least 80 Percent of the source code.

In case the server hosting the application is rebooted, the application shall be automatically restarted as well.

4.1.4 Performance and Efficiency

Load times and data transmission shall be within reasonable bounds. Except for the transmission of big data files from the users PC to the application server, there shall be no significant delays when using the application.

The time required to process a given dataset shall be estimated and the estimate shall be presented to the user before he triggers the processing by clicking the designated button.

4.1.5 Security

The data transmitted to the application for processing is not expected to be privacy-sensitive or confidential. Data submitted by individual users will generally not be presented to other users, but no efforts are made to protect the data from being retrieved by other authorized users of the server.

4.2 Interface Requirements

TBD

5 Appendices

5.1 Glossary

RNA-seq NGS Technology applied to whole genome transcriptome profiling