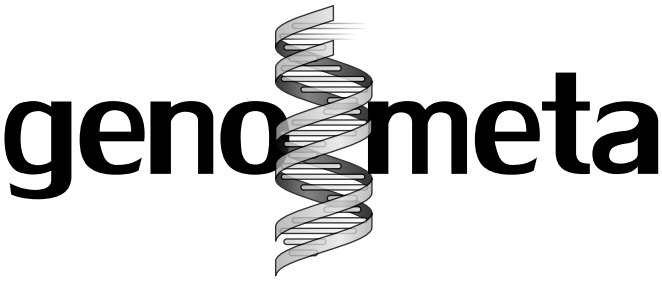
Fachhochschule Hannover
Fakultät IV - Wirtschaft und Informatik
Ricklinger Stadtweg 118
30459 Hannover

WS 2010/11

software requirements specification



FH Hannover

October 18, 2010

date	**version**	**comments**
2010-10-05	0.1	initial release
2010-10-06	0.2	add missing functional requirements
2010-10-07	0.3	first complete draft
2010-10-17	0.4	clarify some of the funtional reqs.

Contents

1 Introduction

This section gives a short introduction to the general scope of the project and ist primary goals.

1.1 Purpose

The SRS document is intended to represent a technical analysis of the product concept catalogue. It will try put the clients product specification into a technical perspective and to clarify main points for the programming staff.

1.2 Scope

In this project a software for visualizing DNA sequence reads is to be developed. The current project name "genometa" is subject for discussion and will be used as a preliminary name until a final name is decided upon.

The visualization will be useful in 2 distinct medical subcategories, namely in metagenomics - where bacteria are identified based upon their DNA - and transcriptomics - in which genes used by a specific bacterium are identified.

There are existing software products currently in use by the client, which only fulfill the requirements to a certain extent. The new software should incorporate the main advantages of these products and combine them into a single software solution.

1.3 Definitions, Acronyms, and Abbreviations.

The following definitions and abbreviations are used throughout this document

abbreviation	**explanation**
DNA	short for deoxyribonucleic acid
⋮	*vdots*

reference	document	explanation
PCC	lastenheft1.doc	product concept catalogue
SAMTools	http://samtools.sourceforge.net/SAM1.pdf	details about SAM format import/export
Picard	http://picard.sourceforge.net/index.shtml	alternative library for SAM/BAM import/-export
IGB	http://www.bioviz.org/igb	Integrated Genome Browser
Geneious	http://www.geneious.com/	Geneious visualization program
Tablet	http://boiinf.scri.ac.uk/tablet/	Tablet visualization program
NGSView	http://ngsview.sourceforge.net/	NGSView visualization program
MagicViewer		
SeqMonk		
Apollo		

1.4 References

The following documents and other resources are referenced throughout this document

1.5 Overview

The following chapters will give an a detailed description of the products functional and non-functional requirements and list the clients success criteria categorized to specific priority definitions.

2 Project Description

The genometa project will consist of a software implementation to visualize large datasets of DNA sequence reads in 2 different perspectives - metagenomics and transcriptomics. The current software packages used at MHH do not or only partially provide the tools needed for easy and efficient analyzation.

2.1 Product Perspective

Many software products have been listed as references by MHH and will be used to describe the specific advantages and disadvantages of each, hopefully defining an ideal software structure which will provide a complete package solution to the task that needs to be fulfilled.

The DNA sequence reads, which vary in sizes of around 10MB to 10GB will be provided in SAM format (Sequence Alignment/Map) and the corresponding binary format BAM. For importing and exporting these formats the SAMTools package as well as the Picard software library are provided.

The primary design goals will be an easy to use and visually appealing user interface as well as performance.

2.1.1 Interfaces

The system will be accesible to the user by means of a graphical user interface. A GUI design prototype will be presented in section 4.

2.1.2 Hardware Interfaces

The software is to be used on Windows, Linux and OS X platforms. Hardware is supposed to be on standard levels for office workspaces.

2.1.3 Software Interfaces

The source reads will be provided in SAM and BAM formatted files. The interfaces to read these files will be provided by either the SAMTools or Picard packages. A new import mechanism will be implemented if performance problems arise.

2.1.4 Memory Constraints

A total of 2GB to 4GB of RAM usage should not be exceeded.

2.2 Product Functions

This section will list the functional and non-functional requirements as specified in the product catalogue

2.2.1 Functional requirements

Priority 1

Graphs	
Graphs must be optically pleasing and some settings have to be editable.	• text options like color and size • background color changeable • no overlapping reads • visible read direction

Summary of all species	
The number of reads to each bacterial taxon are displayed as a summary graph.	• show bacteria taxon names • show summered read hits • summed reads in left hand text field

Reads aligned to one species	
When a bacterial name is clicked upon the reads aligned to that genome are displayed as a bar graph.	• show read positions

Zoomable Graphs	
The size of the histograms should be adjustable.	• show complete histogram • multiple zoom levels

Priority 2

Feature annotation track	
Features read in from a GFF file should be displayed as horizontal bars together with their orientation, positive or negative.	• the software should display genes at their corresponding start and end positions • the software should stack overlapping genes to allow distinction

Map RefSeq code to species name	
A name will be mapped to RefSeq code.	• software should display RefSeq and name

Map genera name to lineage name	
Genera name should be mapped to lineage.	• software should display lineage

Text export	
Export text analysis and summary statistics.	• software should be able to export information to CSV files

Analysis number of reads	
Count the number of reads.	• show numbers of reads attributed to each species/gene in text widget – Metagenomics: Normalise number of reads which hit each genome by genome length – Transcriptomics: Normalise number of reads which hit each gene by gene length to 1000bp and normalise total number of reads to one million (Reads per Kilobase per Million RPKM).

Priority 3

Run external alignment tools and reimport results	
Run alignment of specified reads file (in fastA, fastQ format) against a built reference database of bacterial species.	• the software should include funtionality to run an external aligment tool – the user should be able to choose program used – options for the external aligner should be settable via ui interaction – progress of external alignment should be visible to the user – as most external aligners are linux based this functionality should be linux only

Priority 4

Sequence export	
Export the sequence of a particular region.	• export sequence of DNA ACTG to a text widget

Graph export	
Export graph to vector and raster formats.	• possible formats: – SVG/PDF – PNG

Analysis for transcriptomes	
Find reads that group together into continuous segments and list them as new potential sRNAs or ORFs.	• export to a list with coordinates and sequence

Priority 5

SNP visualisation	
SNPs, i.e. mismatches between the letter in the reads and the letter in the genome string, should be flagged in the visualization.	• should be optional • only for one genome or transcriptome • maybe easier to find with external program, eg. Samtools, Varscan

2.2.2 Non-functional requirements

Priority 1

Usability	
Functions in the program should be easily accessible to non-technically oriented biologist users.	• the software should provide a clear and understandable menu system • an installation guide should be included

Memory efficiency	
The program should be able to read between 0.1M and 100M reads from a SAM/BAM file into RAM rapidly. The memory usage should not exceed 4GB.	• the software should include a fast SAM/BAM importer/parser • the software should provide a means of identifying the current memory usage (memory usage indicator)

Multi processor capable	
The program should use multiple CPU cores if available.	• the following features should be parallelized if possible – SAM/BAM import – statistics generation – graphics output (sliding window)

Documentation	

The entire project should be documented.	• documentation should be provided for
	  – source code (doxygen)
	  – how to
	  – tips and tricks
	• all documentation will be in english

Programming language	
The program code and any code for running external programs on the command line should be written in Java.	• the code will be written in SUN Java • all external libraries and other dependencies will be thoroughly documented

Priority 2

Platform independent	
The software should be usable on Windows, Linux and OS X systems.	• the software should be distributed my means of platform specific installer packages • the software should be well organised and packaged

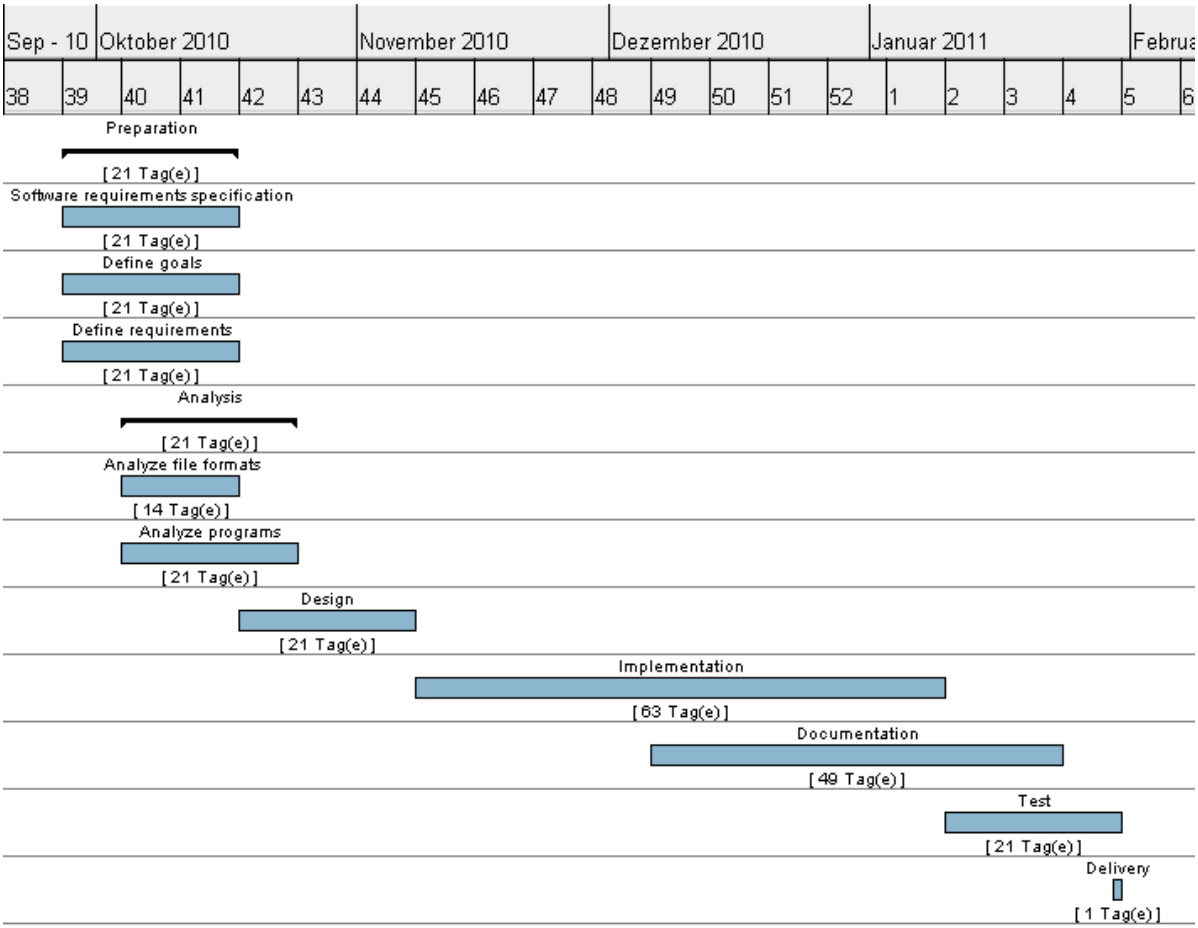
Website	
A simple html/php website should be implemented for documentation and distribution of the program.	• the files and website will be hosted on MHH servers • should include documentation and download pages

Open source	
Software source code should be made available sometime in or after the main development process.	• the source code should be made available by means of an online codebase system (e.g. sourceforge) or as source file downloads

3 Time management

The project will start in October 2010, and end on January 2011. On that date a stable product version should be made available included the defined documents and connected requirements (website, installer programs).

A preliminary time management plan is shown here:



The plan will be updated according to project revisions and definition of the work breakdown structure.

4 GUI prototype

The following image will show a GUI prototype.