Product Vision

Tycho van Heems - tvheems - 4464567
Dex van Leeuwen - dqvanleeuwen - 4475461
Martijn Straatman - martijnstraatm - 4442504
Jochem Tiessen - jtiessen - 4478339
Michael Tran - michaeltran - 4499638

May 11, 2017

Contents

1 Introduction

Recently, research in DNA is becoming a popular field in science. DNA contains information about traits of a person and can be used to predict those. DNA analysis is used by the police to find criminals, but most research of DNA right now goes to predicting traits of living entities. These traits can be things from length of someone to whether someone might get a certain disease, like Leber's Hereditary Optic Neuropathy according to a research paper by Wallace, D.C., Singh, G., Lott, M.T., Hodge, J.A. & Schurr, T.G.(1988). When a disease has been predicted using DNA analysis, in some cases, its severeness can be predicted. "The polymerase chain reaction was used to detect four mutations in the DNA of 47 unrelated patients with type I Gaucher's disease (94 Gaucher's disease alleles)."(Zimran, A., et al., 1989), Zimran and his colleagues found a pattern in the DNA to see if someone who has Gaucher's disease has Type 1 of the disease.
If diseases can be predicted using this analysis, health care will improve a lot, since it will be known on time whether a person might get a certain disease or not and precautions can be taken. However, research in predicting these traits is not far yet, which means there is a lot to be discovered.
Most of these discoveries are by finding mutations in genomes. Genomes contain all the information of DNA. They consist of bases, where the genome is a sequence of any number of one of the following four bases: Adenine, Cytosine, Guanine and Thymine. These sequences determine all personal traits, where mutations cause traits to be different in living entities. "First of all, many human diseases are influenced by, if not caused by mutations in genes" (Daniel Nathans).
However, there are not enough tools yet to efficiently work with genomes. Genomes contain a lot of information, which makes it easy to miss important details when researching a genome. Especially when comparing multiple different genomes, it may be really difficult to keep track of those, let alone retrieve useful information from these genomes. The lack of tools to make it easier to analyze genomes halts the research on this topic, while it is important that research on DNA mutations keep going as lives can be saved by this research.
To help with research on genomes, we will make a genome browser, which can read in files containing the information of the genomes. This information should be visualized in a graph, so researchers have a visual representation of all the information. The researcher will be able to navigate through this representation and retrieve all the information they need in a clear, simple way.
The rest of this document will explain how we(the development team) envision the product by talking about our customer and what the customer wants. It will also be specified what the most crucial features of the product are and how this product should compare to other tools in the same industry. Lastly, the timeframe and budget in which the product has to be made will be discussed.

2 Target Customer

This chapter will introduce you to our target customer and why the product will be made for the customer.
The genome browser will be made for a company called GenomeViz Inc. GenomeViz Inc. is an aspiring company on DNA analysis. The company researcher all kinds of DNA, from the DNA of bacteria to the DNA of a tomato. The progress of the research of GenomeViz Inc. is slow however, as they cannot seem to find a way to organize and view the large amounts of information contained in the DNA they are researching.
For this reason, GenomeViz Inc. asked us to create a genome browser for the in which they have a clear overview of the genomes they are studying, so retrieving information of those genomes will be easier.
Communication with GenomeViz Inc. is done with four different employees working at the company, namely:

- T. Abeel is the CEO of GenomeViz Inc.
- T. Mokveld is a data scientist at GenomeViz Inc.
- J. Linthorst is the CTO of GenomeViz Inc.
- L. Krombeen is a software analyst at GenomeViz Inc.

Each of these contacts is specialized in a specific task around the company, so the information and demands we receive for the application are well rounded to all the fields the company specializes in.

3 Customer Needs

In this chapter, it will be explained what the customer expects from the product and what they do not want to see in the product.
GenomeViz Inc. has tasked us with several requirements for the genome browser they have requested. The following list contains the requirements set by GenomeViz Inc.:

- In a meeting, T. Abeel[1] stated: "Data scaling is #1 priority", so the data needs to be read as efficiently as possible and the visualization should not take long, so working with the data can be done efficiently.
- The application needs to be capable of reading genome data from .gfa files, these files can contain more data than can fit in memory. The application should then visualize the data from this file as a graph.
- The application should also allow the user to easily and intuitively find information about the genome in this graph.
- The user should be able to extract information out the graph and be able to use it elsewhere on the same computer.
- There should be a possibility of highlighting parts of the graph to keep track of the genome.
- The application should be able to run on the following operating systems: Windows, Apple and Linux.
- The shape of the nodes in the graph should not be circles.

All these requirements should be met to satisfy the needs of GenomeViz Inc. to make sure the customer is satisfied. If one of these requirements is not met, the product may not be suitable enough to use for GenomeViz Inc. which will not only be a letdown for GenomeViz Inc. but also for the development team.

[1] T. Abeel is the CEO of GenomeViz Inc.

4 Crucial Features

This chapter will explain the crucial features of the product. There will be some overlap with the customer needs specified in chapter 3. These features will be split in functional features and non-functional features. Functional features are all features that give the application some form of functionality and non-functional features are all other features.

4.1 Functional Features

- The application must have a parser that can parse data from .gfa files, so the data in those files can be visualized.
- When the data is parsed, the application must make a graph of it containing all the information contained in the file.
- It should be possible to view the information contained in the nodes and edges.
- The application should have a function, which makes it possible to create a sub-graph of a center node and close surrounding nodes.
- It should be possible to copy the genome information to the clipboard and view the information in some console.

4.2 Non-Functional Features

- Considering the potential size of the input files, the application must use a data structure that can efficiently store and retrieve data.
- The application must allow for fast navigation through the graph, so information can be handled efficiently.
- The graph should be visualized using visual encodings, so some information can be deduced without focusing on the specific part of the graph. These encoding can be made by using different colors/shapes for example.
- The application should work in the following operating systems: Windows, Apple and Linux.

These features should be met no matter what, because if one of these features lacks, the product becomes a lot less effective and maybe even unusable.

5 Comparison to other products

A potential issue is that the product might not be interesting because it doesn't provide users with any kind of functionality that other products don't already provide them with. In order to avoid this issue, we have researched what kind of alternative applications can be found that provide services similar to those our application provides.
There are not many products that are designed for this purpose, however there are two notable alternatives. The first is Cytoscape[2], and the second is Bandage[3].

5.1 Cytoscape

Cytoscape is used to visualize networks as graphs, and was originally designed to visualize molecules. However, when using Cytoscape to visualize genomes, it becomes very apparent that its focus is not assembly graphs, as its user interface is unnecessarily complex when used for linear graphs. Because our application is specifically designed for alignment graphs the user interface will be much clearer and more intuitive.

5.2 Bandage

Bandage is used specifically for De Novo assembly graphs, which contain cycles causing the graphs to look cluttered, while our application is only made for alignment graphs, which are linear and clear. The way the graphs of bandage handle the large amounts of data is efficient, which is their strongest point. This is also the reason that Bandage is one of the more popular applications in the field of DNA analysis.

5.3 Our application

The reason our application should be used in comparison to other applications is mainly that our application is focused towards linear alignment graphs of DNA. Cytoscape can represent large graphs really well, but because it is a tool with many uses, its user interface can be confusing. Our application is only focused on DNA alignment, which is why there will be less functions in the user interface, which will be less confusing because of that. Bandage is also focused on DNA analysis, but on a different field in DNA analysis, which is why our application will be the preferred choice for the task it is made for, creating linear alignment graphs.

[2] Cytoscape. *http://www.cytoscape.org/*
[3] Bandage. *https://rrwick.github.io/Bandage/*

6 Timeframe and Budget

This chapter discusses the timeframe and budget the development team received to make this application.

6.1 Timeframe

GenomeViz Inc. wants to use the multi-genome browser in ten weeks, which means that the development team has a total of ten weeks to finish the application. It is assumed that the development team puts in a total of 140 hours per week in the product. There will be several meetings throughout the timeframe with the customer, so the development team and the customer can talk about the application and the features is should contain. During those meetings, it is expected that the development team brings demo with new features to show the progress.

6.2 Budget

GenomeViz Inc. assigned a budget of €0,-. This means that the development team has to find all resources needed to make the product itself and cannot rely on outside sources.

7 Glossary

Genome - A genome is a piece of genetic information that consists of DNA.

Mutation - A mutation in a genome is a difference in a genome when compared to another genome.

Genome browser - A genome browser is an application that can display the information of genomes and navigate efficiently through that information.

Graph - A visual representation of relations of objects.

Data scaling - Data scaling is the ability of an application to work with amounts of data that are normally to large to handle efficiently.

Timeframe - The timeframe of a project is the amount of time the project needs to be finished in.

Budget - The budget is the amount of money that is made available for the project.

References

Nathans, D. (n.d.). BrainyQuote.com. Retrieved May 10, 2017, from BrainyQuote.com Web site: *https://www.brainyquote.com/quotes/quotes/d/danielnath320033.html*

Shannon, P., Markiel, A. & Ozier, O., et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". Genome Res. 13 (11): 2498–504.

Wallace, D.C., Singh, G., Lott, M.T., Hodge, J.A. & Schurr, T.G.(1988). Mitochondrial DNA Mutation Associated with Leber's Hereditary Optic Neuropathy. 242(4884), 1427-1430. doi: 10.1126/science.3201231

Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualisation of de novo genome assemblies. Bioinformatics, 31(20), 3350-3352

Zimran, A., Gross, E., West, C., Sorge, J., Kubitz, M., & Beutler, E. (1989). Prediction of severity of Gaucher's disease by identification of mutations at DNA level. The Lancet, 334(8659), 349-352.