Product Vision

Wouter Smit, 4401409
Justin van der Krieken, 4357116
Casper Athmer, 4329066
Faris Elghlan, 4341538
Cas Bilstra, 4381084
Delft University of Technology

Abstract

This document describes the features of our product. It does so by analyzing the target customer and defining his problems. Key attributes that define the product are presented, which are based directly on the customer needs. The document also includes a literature study on existing alternative of the product and defines the differences between our product and these alternatives.

Contents

Introduction

This document describes the analysis that has been made regarding the problem description for the Programming Life context. This work is part of the Contextproject course (course code: TI2806) at the EEMCS faculty of the Delft University of Technology.

Firstly, we define the user of the product and specify its characteristics. We will then discuss the needs of the user that the product will address. Based on these needs, we define the elements of the product that form the key to addressing these needs.

Next, we analyze existing products and compare the product against these. We identify unique differences that set this product apart from the rest and conclude with a short discussion on the lessons learnt from existing projects. Finally we provide information regarding organization of the project, such as the budget and timeframe.

User Definition

The Client. The project is executed at the request and under supervision of Dr. Thomas Abeel[1], who acts as representative of Broad Institute of MIT and Harvard[2]. He communicates to the team the needs of users, who the users are and feedback on iterations of the product.

The Users. The product will be developed specifically for Broad Institute and its researchers. The target customers are therefore researchers working at Broad institute, however, the product targets any researcher interested in exploring large genomic data sets.

These researchers work on analysis of DNA genomes of bacteria. They specifically analyze the differences between multiple genomes of a bacterium. The goal of this analysis is to identify patterns, outliers and trends in the data, which can be used to formulate novel hypotheses and check existing ones.

User Background. Researchers in this field of work are assumed to have vast backgrounds in biology of at least a Masters degree. They are specialized in the nature of genetics. While these users know the ins and outs of DNA, they are no computer scientists. They therefore cannot be expected to possess deep knowledge of a computer system or data structures.

Targeted Customer Needs

We will now discuss the customers needs by analyzing their problem description and extracting the needs that the product will address.

Problem Description

The customer wants to do analytical research on vast data sets of DNA genomes. Collecting Analytical data is a cumbersome process, as seen in a recent study of Comparative Genomic Analyses (Earl et al., 2016). The effectiveness of this research can be increased by utilizing Exploratory Data Analysis (Behrens & Yu, 2003). To enable this type of research, tools are needed that aide researchers

[1] http://www.abeel.be/

[2] https://www.broadinstitute.org/

in the exploration process. Tools that succeed in doing so do not currently exist. Such a tool should be able to load and process large data sets, with the current largest set consisting of 6000 genomes, as well as visualizing this data in an interactive graphical interface, as specified by the client.

Problem Analysis

We now analyze the problem as presented by the customer. The customer needs a tool (application) that enables them to perform Exploratory Data Analysis. This tool must address two main problems: scalability and semantic visualization

Scalability. "Over the past three years, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by over two orders of magnitude" (Oehmen & Nieplocha, 2006; Shendure & Ji, 2008, p. 1). For this reason, the application must be developed in such a way that it can handle dramatic increases in input size.

Semantic Visualization. In order for the application to successfully aide researchers, it should generate a visualization that interprets the semantics of the data. Such a visualization is assumed to be successful if it presents the user with the desired (interesting) information and temporarily hides not interesting information, to prevent the user from getting lost. This interpretation should be communicated to the user through colors, shapes or other encodings.

Localization. Due to the large size of the data, exploration through it can easily get the user lost. An important aspect in the semantic visualization is that the user does not get lost.

Semantic Interaction and Zoom. The semantic visualization must be interactive. When the user interacts with the visualized data, the visualization is adapted to reflect the current interest to the user. An example of this is semantic zoom, where detailed information of individual mutations is hidden at the top-level and gradually presented to the user while zooming in (Muthukumarasamy & Stasko, 1995).

Semantic Identification. Another aspect of semantic visualization is selectively presenting potentially interesting structures, trends or patterns in the data. Examples of these identifications are different types of mutations, convergent evolution and drug resistance.

Key Product Attributes

In order to successfully address all user needs, we define the attributes that directly satisfy these needs. These attributes will be the cornerstone of the product and define when the product is successful.

Data Scalability. The application will be able to load potentially unlimited data sets. This is done by selectively loading the data in memory when necessary. Additionally, the application will support a responsive interface that allows the user interface to work seamlessly.

Semantic Identification. The application will process the input data and and identify patterns and trends and presents these to the user when they are of interest for his research. Examples are mutations, genes, and cumulative information such as the amount of mutations in a specific area.

Localization. To prevent the user from getting lost, there should always be a reference to the current location relative to the whole data set.

Semantic Zoom. The application will support semantic zooming. This is done in two parts. Firstly, 'ordinary' semantic zooming is used to show only structures that are relevant based on their size in the current level of zoom (Muthukumarasamy & Stasko, 1995). Secondly, 'stationary' semantic zooming (Smith, 2006) is used to show structures based on their level of importance. For instance, a certain mutation in a piece of gene is more important than the same mutation in a piece of noncoding DNA (DNA that does not encode any gene).

Comparison with Previous Work

There exist other applications and tools that attempt to achieve analytical processing of DNA sequences and these tools have existed for quite some time (Pearson & Lipman, 1988; Minton, Flanagan, & Ellard, 2011). These tools, however, do not provide realtime analytical data, instead requiring a lot of processing before an output is provided. Our product will be able to perform analytical calculations and present these to the user in realtime through an interactive interface. This heavily speeds up the process and intuitivity of obtaining the data.

In previous iterations of this project, applications were developed that did provide an interface in realtime (Abeel, 2015), but these applications were unable to provide a good semantic visualization. Particularly, they lacked semantic zooming and a means of keeping overview throughout the data.

Project Organization

Timeframe. The product will be developed over the course of 10 weeks. A working version of the product will be delivered and presented to the customer at the end of every week, according to the Scrum methodology (Schwaber & Sutherland, 2013). At the end of the timeframe, the product will be released.

Budget. The project will be executed free of charge by students of the Delft University of Technology. The product will be released under the Apache License v2.0 (Apache Software Foundation, 2004), which is considered a free software license (Free Software Foundation, 2015) by the Free Software Foundation (Free Software Foundation, 2016).

References

Abeel, T. (2015). *Abeellab.* Retrieved 4 May 2016, from `https://github.com/AbeelLab/`

Apache Software Foundation. (2004, January). *Apache license, version 2.0.* Retrieved 4 May 2016, from `http://www.apache.org/licenses/LICENSE-2.0`

Behrens, J. T., & Yu, C.-H. (2003, April 15). Exploratory data analysis. In *Handbook of psychology.* John Wiley & Sons, Inc. doi: 10.1002/0471264385.wei0202

Earl, J. P., de Vries, S. P., Ahmed, A., Powell, E., Schultz, M. P., Hermans, P. W., ... Ehrlich, G. D. (2016, February 24). Comparative genomic analyses of the moraxella catarrhalis serosensitive and seroresistant lineages demonstrate their independent evolution. *Genome Biology and Evolution*, *8*(4), 955-974. Retrieved 4 May 2016, from `http://gbe.oxfordjournals.org/content/8/4/955.abstract` doi: 10.1093/gbe/evw039

Free Software Foundation. (2015, September 1). *What is free software?* Retrieved 4 May 2016, from `https://www.gnu.org/philosophy/free-sw.en.html`

Free Software Foundation. (2016). *Various licenses and comments about them.* Retrieved 4 May 2016, from https://www.gnu.org/licenses/license-list.html

Minton, J. A. L., Flanagan, S. E., & Ellard, S. (2011). Pcr mutation detection protocols. In D. B. Theophilus & R. Rapley (Eds.), (pp. 143–153). Totowa, NJ: Humana Press. doi: 10.1007/978-1-60761-947-5_10

Muthukumarasamy, J., & Stasko, J. T. (1995). *Visualizing program executions on large data sets using semantic zooming.* Graphics, Visualization & Usability Center, Georgia Institute of Technology. doi: 10.1109/VL.1996.545283

Oehmen, C., & Nieplocha, J. (2006, August). Scalablast: A scalable implementation of blast for high-performance data-intensive bioinformatics analysis. *IEEE Transactions on Parallel & Distributed Systems*, *17*(8), 740-749. doi: 10.1109/TPDS.2006.112

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, *85*(8), 2444-2448. Retrieved 4 May 2016, from http://www.pnas.org/content/85/8/2444.abstract

Schwaber, K., & Sutherland, J. (2013). *The scrum guide*. Retrieved 4 May 2016, from http://www.scrumguides.org/docs/scrumguide/v1/scrum-guide-us.pdf

Shendure, J., & Ji, H. (2008, October 9). Next-generation dna sequencing. *Nature Biotechnology*, *26*, 1135 - 1145. doi: 10.1038/nbt1486

Smith, R. (2006, September 19). *Stationary semantic zooming.* Google Patents. Retrieved 4 May 2016, from https://www.google.com/patents/US7109998 (US Patent 7,109,998)