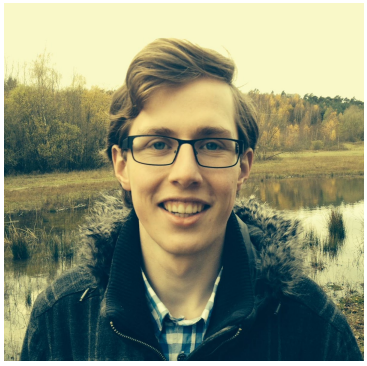
Product vision Genome Explorer

Programming Life: Context Project TI2806

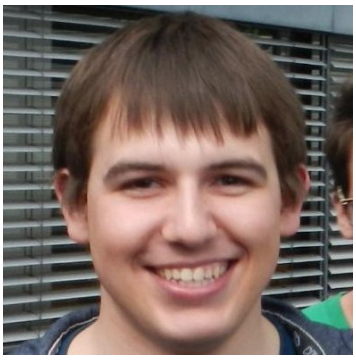
8-5-2015
Group 2
Supervisors: Dr. A. Bacchelli & Dr. T. Abeel

Gerben Oolbekkink (gjwoolbekkink, 4223896)



t

Jasper Boot (jboot, 1516272)



Jasper Nieuwdorp (jnieuwdorp, 4215796)

Jim Hommes (jhommes, 4306090)

René Vennik (rvennik, 4102959)

Table of contents

1. Abstract ..... p. 2
2. Introduction ..... p. 2
3. Customer ..... p. 2
4. Customer needs ..... p. 2
    4.1 Visualization ..... p. 2
    4.2 Identification ..... p. 3
    4.3 Integration ..... p. 3
    4.4 Requirements ..... p. 3
5. Product attributes ..... p. 3
6. Target ..... p. 3
    6.1 Timeframe ..... p. 4
    6.2 Project cost ..... p. 4
7. References ..... p. 4

1. Abstract

For research biologists of Broad Institute of MIT and Harvard and KwaZulu Natal Research Institute for Tuberculosis & HIV (KRITH), who need to explore large genomic datasets, Genome Explorer is a computer application that easily shows the differences of large genome sets and draws attention to potentially interesting mutations of those datasets. Unlike primary alternatives, our product interactively visualizes DNA sequence graphs and does exploratory data analysis on the datasets.

2. Introduction

This document will give an insight in the methodology used and process made during our project in the context programming life. In this updating and evolving document we will reflect our vision for the final product and in the end it can be read as a roadmap to our final result.

3. Customer

The customer is represented by research biologists of Broad Institute of MIT and Harvard and KwaZulu Natal Research Institute for Tuberculosis & HIV (KRITH)(Abeel, 2015). We assume that customer has a master's degree or higher and that he has basic knowledge about computers, but has no prior knowledge about any particular programming language or specific specialised applications. However, we assume that the customer can learn to use specialized programs, such as Genome Explorer, by trying out intuitive features that the program provides.

4. Customer needs

Bacterias, such as Tuberculosis mutations, are getting drug resistant (Mahr, 2013) and researchers are trying to find out how this resistance is developed. This plays a big part in the customer’s need to be able “to interactively explore a sequence graph representing the genome architecture of multiple strains” (Abeel, 2015).

4.1 Visualization

The customer will also need “semantic zooming to enable useful visual interpretation at various zoom levels from whole-genome to individual mutations” (Abeel, 2015). This graph should be put in the context of the evolutionary relationship between bacteria (Abeel, 2015) using the provided Phylogenetic tree (Newick file). There should be visual encodings for different types of mutations (Abeel, 2015) and there should also come a filter to show subsets of those types. There will be need of clear graph and one of the algorithms from ‘Algorithms For Drawing Graphs’ can be used (Eades and Tamassia, 1989).

4.2 Identification

Mutations should be identified and be labeled by their variant (Abeel, 2015). Possible variants are: insertion, deletion and Single-nucleotide polymorphism (Abeel, 2015), that means a single different base (“Single-nucleotide polymorphism - Wikipedia”, 2015). Convergent evolution should also be detected, because “drug resistant or fitness improving mutations are likely to evolve convergently” (Abeel, 2015).

4.3 Integration

On the subject of integration (Abeel, 2015):
The application should be annotated with mutations in the graph. These mutations should come from well-known reference genomes and retrieved from reference database . Sources for the annotations could be: tbdb.org, TuberQ, tuberculist and PolyTB. Also the different coordinate systems of those database should be integrated. Also the by the customer delivered metadata in week 3 should be integrated in the application.

4.4 Requirements

The customer indicated some constraints (Abeel, 2015):

- It should work on 32 bit systems of possible African standards
- 4.5 million bases per sequence
- Multiple of hundreds of bacterial strains
- Graph will contain thousands of edges

5. Product attributes

The application will specialize in comparative DNA sequence analysis on a large numbers of genomes. This can not only be used to find convergent evolution and parallel evolution (Zhang & Kumar, 1997) but also has shown potential for unraveling the function of noncoding sequence (Cliften et al., 2001). By making the genome browser interactive with semantic zooming we make it possible to quickly identify which parts of the data are relevant for research and allow omitting non-relevant information at certain levels. We can also use semantic zooming to annotate parts of the data (Loraine & Helt, 2002)

There already are quite a few genome browsers (“Genome browser - Wikipedia”, 2015). Most of these genome browsers focus on visualizing genomes or annotating them. For most operations with dna sequences computers are the only option due to the large nature of DNA. Algorithms such as BLAST have been used for a long time for aligning dna sequences (Altschul et al., 1997). One of these genome browsers is the Integrative Genomics Viewer (Robinson et al., 2011), which is used for exploring large genomic datasets. Other genome browsers focus, like our product, on mutations. For instance COSMIC (Forbes et al. 2014) focuses on mutations in the human genome which cause cancer.

Our product will differ from alternatives because of the scale. It can load a relatively large amount of genome strands (around 500) and compare them. This gives our application a wide area in which it can be used. By including data and information from external sources our program will serve a wide audience while being relevant for specific research goals. Because our program can handle larger strands it can also be used for other species with larger genomes. This could make our application usable for research in all the fields between bacterial evolution to genome mutations in cancerous cells.

6. Target

6.1 Timeframe

The product will be developed over a time period of 10 weeks ("Times and locations table", 2015). The target will to have a working product every week according to the scrum methodology (Schwaber and Sutherland, 2013). And at the end of the 10 weeks a product that will improve the search for drug resistant strains.

6.2 Project cost

For now this project will be developed by students free of charge. And will be released open source to "study, change, and distribute the software to anyone and for any purpose" ("Open-source software - Wikipedia", 2015).

7. References

Abeel, T. (2015, 13-4-2015). Contextproject Programming Life [slides]. Retrieved 8-5-2015, from https://blackboard.tudelft.nl/bbcswebdav/pid-2443556-dt-content-rid-8387922_2/courses/34575-141504/customer_meeting.pdf

Single-nucleotide polymorphism - Wikipedia. (2015, 28-4-2015). Retrieved 8-5-2015, from http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

Eades, P. and Tamassia, R. (1989). Algorithms For Drawing Graphs: An Annotated Bibliography. Retrieved 8-5-2015, from ftp://ftp.cs.brown.edu/pub/techreports/89/cs89-09.pdf

Times and locations table. (2015). Retrieved 8-5-2015, from https://www.dropbox.com/s/t0z65qrjg1z3qd4/TI2806_2015Plan.pdf?dl=0

Schwaber, K. and Sutherland, J. (2013, 7-2013). The Scrum Guide. Retrieved 8-5-2015, from http://www.scrumguides.org/docs/scrumguide/v1/Scrum-Guide-US.pdf

Zhang, J & Kumar, S (1997). Detection of convergent and parallel evolution at the amino acid sequence level.. *Molecular Biology and Evolution, 14*, 527-536. Retrieved 8-5-2015, from http://mbe.oxfordjournals.org/content/14/5/527.full.pdf+html

Cliften, P., Hillier, L., Lucinda, F. Graves, T., Miner, T., Gish, W., Waterston, R. and Johnston, M. (2001). Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. *Genome Research, 11*, 1175-1186. Retrieved 8-5-2015 from http://genome.cshlp.org/content/11/7/1175.full.pdf+html

Mahr, K (2013, March 4). Contagion, why drug-resistant tuberculosis threatens us all. *Time, 181*, 28-36.

Genome browser. (2015, March 10). In *Wikipedia, The Free Encyclopedia*. Retrieved 19:35, May 8, 2015, from http://en.wikipedia.org/w/index.php?title=Genome_browser&oldid=650821329

Loraine, A & Helt, G (2002). Visualizing the genome: techniques for presenting human genome data and annotations. *BMC Bioinformatics, 3-19*. Retrieved 8-5-2015, from http://www.biomedcentral.com/1471-2105/3/19

Open-source software - Wikipedia. (2015, 5-5-2015). Retrieved 8-5-2015, from http://en.wikipedia.org/wiki/Open-source_software

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389-3402.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. (2011, January 10) Integrative Genomics Viewer. *Nature Biotechnology 29*, 24–26 (2011)

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U. & Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research, 43*, D805-D811.