Product Vision

Programming Life

Team Dynamites & Butterflies

Table of Contents

This document covers the Product Vision of an interactive multiple genome browser. This document covers the description of the selected customer needs and the necessary product attributes meeting those needs.

DNA is the universal code of life and is represented by the four letters A, T, G and C. The genome is the blueprint and the source code of the operating system of an organism, whether it is a human, a plant or a bacterium. Mutations in the genome will result in changes in how genes work. Differences (small and big) between the genomes of people are responsible for many traits we use to differentiate people, such as eye color, but they are equally responsible for differences in our immune systems, allergies, the propensity to get cancer, etc.

Technological advances over the last few years have enabled scientists to determine the DNA sequence of thousands of organisms, including the human genome, but also many bacterial organisms. Bacteria are an essential part of who we are. The human body harbors more bacterial cells than 'real' human cells. While there are bacteria that perform various useful tasks for us, there are also plenty of bacterial species that make us sick. Bacterial infections are responsible for millions of deaths each year, and most people will be treated with antibiotics at some point in their lives to fight off an infection. Just like in human cells, spontaneous mutations also occur in bacterial cells. Some of these mutations will have negative effects, but others may make the bacteria immune to antibiotics. From the bacteria's point of view this is a very desirable feature, but from the medical perspective this is a serious problem as it limits our ability to treat patients.

1. Who is going to buy the product? Who is the target customer?

The interactive genome browser is made commissioned by GenomeViz Inc. From within this company, we have direct contact with the following stakeholders/employees:

1. Thomas Abeel, CEO of GenomeViz Inc. (also the Programming Life Coördinator). From his position, the team of developers keep him posted on their progress, and he tells the team of developers his wishes (which feature to add now).
2. Tom Mokveld, Data Scientist at GenomeViz Inc. and Jasper Linthorst, CTO of GenomeViz Inc., (also Context Teaching Assistants) are also giving us feedback from within the company, they are a bit closer to the subject and can provide more specific information about the product.
3. Lars Krombeen, Software Teaching Assistant, he is helping the developer team with everything the 3 people above can't help us with (The software side of the project).

Furthermore, our product will also be created for biologists interested in mutations between multiple genomes, specifically in visualizing them and determining the type of mutation between the genomes.

2. Which customer needs will the product address?

Abeel, T. (2017), has requested a product that can: "visualize DNA sequence graphs to represent the genome architecture of organisms of interest, such as drug-resistant human pathogens. The interactive visualization of large scale pan-genome graphs enables exploratory data analysis to formulate novel hypotheses, check existing ones, and to identify outliers, trends and patterns in the data.

Even though sequencing technology has enabled scientists to determine the DNA sequence of thousands of bacteria, there is a distinct lack in computational tools to interpret these big genomic data sets."

3. **Which product attributes are crucial to satisfy the selected needs, and therefore to the success of the product?**

To be able to browse interactively through genomes, the developer team was asked to build a program that will give the user a visualization of multiple genomes represented in a graph. There are several functions that need to be implemented to satisfy the customers needs.

Most important are the reading of large files (a lot of genomes / DNA sequences) in real time, the scalable layout of the graph, the semantic zooming to enable useful visual interpretations at various zoom levels, from whole-genome to individual mutations and the program should be able to recognize different kinds of mutations between multiple genomes and visualize those. Further it is important that we can identify the type of the mutations (insertion, deletion, SNP) and we have visual encodings for different classes of mutation and the ability to filter on mutation class.

According to Sugiyama, K. (2002), the following steps are required to uniformly distribute nodes en minimize edge crossings in a graph:
"Step I: Making the general directed graph acyclic.
Step II: The assignment of vertices to layers in the acyclic directed graph.
Step III: The determination of the order of vertices on each layer.
Step IV: The determination of the position of vertices on each layer." (p. 61)

For Step II we will use the procedure described by Tamassia, R. (2010), where he describes how to layerize graphs using the longest-path algorithm and the use of dummy nodes in order to reduce edge crossings.

For Step III we will use one of two methods described by Matuszewski C., Schönfeld R. and Molitor P. (1999) . "The most popular Heuristics for one sided crossing minimization are the barycenter and the median heuristic. There are other heuristics known from literature, …, are mostly outperformed by barycenter and median heuristics." (p. 218). The idea of barycenter and median heuristics is further explained by Patarasuk, P (2004): "The barycenter heuristic only needs to calculate barycenter values for each vertex, and then sorts the vertices according to these values. Hence, no comparison between numbers of crossings is made." (p. 18).

Finally we will pay attention to some quality-of-life features which should make the usage of the program much simpler, easier and more fun.

4. **How does the product compare against existing products, both from competitors and the same company? What are the product's unique selling points?**

While there are some existing tools that can visualize multiple genome graphs, none of them completely satisfy the customers needs (so far).

For instance Cytoscape is a pretty nice program with a kit of features, but misses the option to read .gfa files (which is wanted by our product owner).

In Bandage, the option to view the sequence specifics (ACTG) is absent, being able to see this is one of our main features.

Gephi is a really nice tool for displaying graphs, but misses some specific DNA properties like specific linear directed acyclic graphs which are topologically ordered.

But most important is the fact that all of these tools aren't able to load very big datasets dynamically and the lack of having a nice and fast coordinate system.

5. **What is the target timeframe and budget to develop and launch the product?**

The project will be completed over a timeframe of a total of 10 weeks (5 days a week, 140 hours per week in total (5 person team)). The budget for developing and launching the product is €0,-.

References

Abeel, T. (2017). Product Description of Programming Life. Retrieved from BlackBoard.

Matuszewski C., Schönfeld R., Molitor P. (1999) Using Sifting for k-Layer Straightline Crossing Minimization. In: Kratochvíyl J. (eds) Graph Drawing. GD 1999. Lecture Notes in Computer Science, vol 1731. Springer, Berlin, Heidelberg.

Patarasuk, P. (2004) Crossing Reduction for Layered Hierarchical Graph Drawing (master thesis). Retrieved from https://fsu.digital.flvc.org/islandora/object/fsu:180382/datastream/PDF/view

Sugiyama, K. (2002). Graph drawing and applications for software and knowledge engineers. River Edge, NJ: World Scientific.

Tamassia, R. (2010) Hierarchical drawing algorithms. London: Chapman & Hall/CRC.