PRODUCT VISION

GENOME VISUALIZATION TOOL

by

Programming Life #3

Student names and numbers:	Tom Brouws,	4278569
	Boris Mattijssen,	4233190
	Mathieu Post,	4286146
	Sam Smulders,	4225007
	Kasper Wendel,	4331362

TU Delft Delft University of Technology

CONTENTS

1

Product Vision

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

The problem that our customer gave us is that he would like to compare the genomes of a lot of bacteria that they extracted from the field. There are no good visualisation tools around right now and this is where our product comes in. Our product can be used for genome research and finding mutations that change bacteria which, for example, make them immune for antibiotics. This is done by visualizing the genomes of these bacteria and making one big sequence graph of it. Mutations of any kind are displayed by different paths in the nodes. Therefore this product is very useful for genome researchers and medical scientists.

1.2. Which customer needs will the product address?

1.2.1. Sequence graph interaction

Since there is a lot of data, if we get larger samples and we don't know what the customer exactly wants to inspect beforehand, we are giving the customer full control. He can choose want he would like to see and explore in the sequence graph. This will be possible in our product because we want to implement filters. These filters can be turned on and off by the customers. We want to implement a filter on genomes, so only the sequence graph of one genome will be displayed. This one genome can then be compared to the average genome of the whole sample or to the reference genome. And of course the customer can navigate to the whole sequence graph and zoom in on every part of it.

1.2.2. Semantic zooming

The customer asked for semantic zooming so only relevant information is displayed. We are going to implement this in our product since we think this is an important aspect. When looking at the sequence graph of the whole sample, mutations are grouped and then displayed depending on the percentage of the sample that is mutated in this spot and which kind of mutation it is. In this way it's easily visible where a lot of mutation occurs and which parts of the genome are interesting for genome research. When zooming in on a specific part of the genome, gradually more information is shown about which genomes are mutated in this part. When zoomed on the lowest level of the genome the sequence graph changes and all of the mutations are displayed individually. Information about the type of the mutation and which genome it is is displayed. The type of the base pair is now also visible.

1.2.3. Identify mutations / Visual encodings for mutations

Another important part the customer demanded in our product was the identification of mutation. We are going to implement this so we can cluster nodes which have the same type of mutation on a specific part of the sequence graph. This is very useful for semantic zooming and this will make our graph a lot less messy. We are also going to give each kind of mutation another color which will make them stand out visually.

1.2.4. Effects of mutation

Another key feature the customer asked for was the effect of a specific mutation. This is very useful to a genome researcher since a small mutation can have a huge impact. We would like to determine the effects of

a mutation by searching for it in a mutation database.

1.2.5. COMMON ANCESTORS

For every group of samples a common ancestor tree is provided. A feature that will determine the genome of these ancestors will be implemented. This way the customer can see were phenotypes like drug resistance were developed.

1.2.6. METADATA FOR EACH GENOME

There is a lot of data available for each genome, such as drug resistance, location of isolation, isolation date, etc. If the customer chooses a genome in our filter, all of this information will be displayed in the user interface.

1.3. WHICH PRODUCT ATTRIBUTES ARE CRITICAL TO SATISFY THE NEEDS SELECTED, AND THEREFORE FOR THE SUCCESS OF THE PRODUCT?

1.3.1. ONLY SHOWING RELEVANT INFORMATION

By implementing semantic zooming only information that is relevant is shown. Since we have a lot of data this is a very important feature. We want to have a very clear and obvious sequence graph that is understandable even for someone with little background knowledge. When zooming in, more information is shown for the customer so the specific parts of the genome can be researched and what the effects of these parts are.

1.3.2. FILTERS

The customer is interested in comparing genomes to see where phenotypes come from and which mutations are responsible for this. Filters are going to make this possible in our product since genomes can be compared with each other. It will also be possible to filter on each kind of mutation and on common ancestors. Every combination of genomes, ancestors and mutation that the customer can think of can be filtered on and will be made into a sequence graph. Our product will stand out due to this flexibility and will be very useful for research purposes.

1.3.3. CLUSTERING OF MUTATIONS

Our product will display a very clear sequence graph because it is possible to cluster mutations. When looking at the graph of the whole sample it will be very easy to determine where mutations are occurring, which kind of mutations these are and the percentage of mutations on the same spot. The customer is now able to see which part of the genome is interesting for his research and this will make our product a success.

1.3.4. ANCESTOR TREE FILTER

There will also be an ancestor filter in our product. The customer will be able to select an ancestor and then the sequence graph of all the genomes that have this ancestor will be shown. This feature is important because this way we can see what the effects are of different mutations on the same ancestor. The customer now is able to research if some phenotype is the effect of an ancestor or of some kind of mutation after this ancestor.

1.4. HOW DOES THE PRODUCT COMPARE AGAINST EXISTING PRODUCTS, BOTH FROM COMPETITORS AND THE SAME COMPANY? WHAT ARE THE PRODUCT'S UNIQUE SELLING POINTS?

Several visualization tools are already on the market, such as ..... A lot of these programs incorporate roughly the same feature set. Our product's unique selling point is that it consciously shows a selection of data about the genomes, instead of all possible data. This means that clutter can be avoided, and the data that is shown is easy to read and comprehend. This way we can cater to a select kind of research related to the data we provide. This is achieved by incorporating a clear and concise user interface, a visual representation of the data and filters to trim the amount of information that is shown.

1.4.1. INTEGRATIVE GENOMICS VIEWER

"We have developed the Integrative Genomics Viewer (IGV), a lightweight visualization tool that enables intuitive real-time exploration of diverse, large-scale genomic data sets on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNA interference screens, gene expression, methylation and genomic annotations."[1]

Both IGV and our application visualize genome data, but the advantage of our application is the semantic zooming and grouping of data from multiple strains. IGV shows very little information on a zoomed-out level, while our application shows useful information on every level.

1.4.2. JALVIEW

"Jalview Version 2 is a system for interactive WYSIWYG editing, analysis and annotation of multiple sequence alignments. Core features include keyboard and mouse-based editing, multiple views and alignment overviews, and linked structure display with Jmol." [2] Both JalView and our application are able to compare different genomes in a visual way. The advantage of our application is the way we visualize this information. JalView just shows you the genome as a sequence of A, C, G and T's. This is very inconvenient when working with larger data sets. Our application focuses on the demand of working with larger datasets, by introducing semantic zooming. The semantic zooming makes sure just the right information is shown at every zooming level.

1.4.3. INTEGRATED GENOME BROWSER

"IGB is an open source, desktop graphical display tool implemented in Java that supports real-time zooming and panning through a genome; layout of genomic features and datasets in moveable, adjustable tiers; incremental or genome-scale data loading from remote web servers or local files; and dynamic manipulation of quantitative data via genome graphs." [3] Both IGB and our product aim to visualize the genome structures. A big advantage of our product over IGB is the ability to show differences between different genomes and thereby inspect the evolution of different genomes.

1.5. WHAT IS THE TARGET TIMEFRAME AND BUDGET TO DEVELOP AND LAUNCH THE PRODUCT?

The target timeframe to develop and launch the product is nine weeks. Since this product is being developed as a study assignment, we have no budget for it.

1.6. SOURCES

[1] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

[2] Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) 'Jalview Version 2 - a multiple sequence alignment editor and analysis workbench' Bioinformatics25 (9) 1189-1191 doi: 10.1093/bioinformatics/btp033