Product Vision

Group:

PL5 - Totally Awesome Genome Coders

Members:

Kasper Grabarz	4083245
Thomas Oomens	4422597
Jeffrey Helgers	4318749
Youp Mickers	4246713
Matthijs Rijlaarsdam	4308417

Deliverable Date:

28-4-2016

Introduction

Over the last decades, antibiotic resistance under bacteria has risen alarmingly, while fewer and fewer new antibiotics have been found. This combination might result in widespread resistance against treatment, effectively setting us back to the dark ages in which common illnesses were a major death cause. In fact, this might happen within several decades.

The current solution is treating the patient with various antibiotics at the same time, hoping more or less that the bacteria will not grow resistant to all of them at the same time. Luckily, the cost of DNA sequencing has dropped significantly over the past years, making it feasible to sequence thousands upon thousands of unique bacteria strains, which can be analyzed for clues regarding antibiotics resistance.

As a single genome contains millions of nucleobases, any unaided human analysis is impossible. As such, the different genomes are put in a de Bruijn graph, creating a large graph of shared regions between the genomes. However, this graph is still too big for humans to comprehend. As such, we seek shelter in techniques from data visualization, in order to make human analysis of the data possible, hopefully finding the key to antibiotics resistance.

Client

The Programming Life context project has been tasked by the Broad Institute and the Kwazulu-Natal Research Institute For Tuberculosis and HIV (KRITH) with developing a visualization tool for comparative genomics data.

The Broad Institute aims to “Transform the process of therapeutic discovery and development” (“Areas of Focus | Broad Institute of MIT and Harvard,” n.d.). KRITH studies all aspects of TB, HIV and HIV/TB co-infection. One of the methods used by KRITH is the sequencing of multidrug-resistant Tuberculosis genomes. (“Scientists and Research,” n.d.). The context project is provided with a dataset from these sequenced Tuberculosis genomes.

Some of the challenges encountered by both clients when analyzing the genome data are:

- How do these genomes look in comparison to each other?
- Which mutations are there?
- Are any of these in ‘interesting’ genes?
- Do they associate with an ‘interesting’ phenotype (such as drug resistance)?
- Do mutations share a common ancestor?

To speed up the process of scientific discovery and to answer the above questions, being able to *look* at these large datasets can be an important tool. Implementing such a tool is the ultimate goal of this project.

As formulated in the project description:
The goal of this project is to develop a tool for the interactive visualization of 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. (*T. Abeel, 2016*)

Client needs

The client has formulated their expectations in nine feature requests:

1. The product should give the user the ability to interactively explore a sequence graph representing the genome architecture of multiple strains.
2. The product should provide semantic zooming to enable useful visual interpretation at various zoom levels (from whole-genome to individual mutations).
3. The sequence graph should be put in context of the evolutionary relationship between bacteria.
4. Visual encodings for different classes of mutations and the ability to filter on mutation class.
5. Identify mutations and determine the type of variant (insertion, deletion, SNP) uniformly across the samples.
6. Put bubbles in the graph in the context of well-known references genomes with their gene annotations and integrate with other reference databases.
7. Provide visual representation and encoding of meta-data associated with samples, such as drug resistance, location of isolation, isolation date etc.
8. Provide indications for convergent evolution of variants (and perform queries on the graph).
9. Integrate with other resources, such as literature databases, mutation databases, to identify graph features that are interesting for further investigation.

The client has furthermore stated multiple features they liked in previous versions of this project, being most importantly: responsive navigation, phylogenetic color-coding, edge thickness for graph prevalence, Quality of Life items, a radial tree, and curved edges.

Product attributes

Methods and tools for visualizing biological data have improved considerably over the last decades, but they are still inadequate for some high-throughput data sets. For most users, a key challenge is to benefit from the deluge of data without being overwhelmed by it. This challenge is still largely unfulfilled and will require the development of truly integrated and highly useable tools. (O'Donoghue et al., 2010)

The focus of TagC lies in the usability of the application. In order to satisfy the needs of the client (and avoid being "overwhelmed by the data") multiple product attributes are crucial:

Data Scaling

The datasets used by the application are of such a size that loading the entire dataset into the application at once is not feasible. Furthermore, for semantic zooming, it is necessary to filter and summarize that data in such a way that only the relevant information for a certain level of zooming is shown.
These two goals overlap, and result in the most critical feature of the product: data scaling. For summarizing the data, a weight algorithm is used; the more genomes are contained in a node or edge, the higher its weight.

Rapid data retrieval

The data will be stored in a RDBMS, which are built and optimized for random access to large data sets. Using these techniques, we are able to efficiently store and retrieve the data necessary. The querying will be done based on a coordinate system containing multiple complexity levels, in order to minimize the data that is needed for the visualization, resulting in a much more fluent user experience.

Simple visualization

Our visualization is intuitive and based on both human nature (e.g. brighter colours draw attention to more interesting part of the genome) and visual conventions (the genome is mapped from left to right). This maximizes both the amount of data we can show at once and the user friendliness of the program, while simultaneously reducing the learning curve. One of the key points are the visual clues indicating the current zoom level, giving the user a feeling of what he is looking at and how to navigate within the program.

Phylogeny

The phylogeny of the genomes will be shown as well, indicating when crucial mutations have taken place and showing how much certain genomes are related.

Highlighting

Selecting two or more genomes will allow the user to see the differences and similarities between the selected genomes, again using crucial visual clues helping the user analyze the data.

Complexity of nodes

Bubble collapsing will be used to visualize the de Bruijn graph, as it improves both the rapid data retrieval and the clarity of the graph. Bubbles will also contain visual clues, for example regarding their complexity, length or other metrics yet to be defined.

Existing products

http://www.ncbi.nlm.nih.gov/pubmed/27072794

https://tudelft.on.worldcat.org/atoztitles/link?rft.stitle=Nat%20Methods&rft.aulast=O%27Donoghue&rft.auinit1=S.%20I.&rft.volume=7&rft.issue=3%20Suppl&rft.spage=S2&rft.epage=S4&rft.atitle=Visualizing%20biological%20data-now%20and%20in%20the%20future.&rft_id=info:doi/10.1038/nmeth.f.301&rft_id=info:pmid/20195254&rft.genre=article&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&ctx_ver=Z39.88-2004&url_ver=Z39.88-2004&url_ctx_fmt=info:ofi/fmt:kev:mtx:ctx&rfr_id=info:sid/nar.oxfordjournals.org&ctx_tim=2016-04-28T01:40:29.139-07:00

Timeframe:

Since there is a hard deadline at the 17th of June, the timeframe exists of 8 weeks between the 18th of April and the 17th of June. Within these 8 weeks, the entire product is designed, developed and tested, multiple times. Between each design/develop/test-round, a meeting is held with the customer, checking whether the progress so far is that what the client wishes.

Costs:

Each week 5 developers will be working for 35 hours. This comes down to a total of 8 * 5 * 35 = 1120 hours. Each developer costs a total of 50 euro, including Vat per hour, coming down to a total of 56.000 euro, including Vat. This includes all meetings with the customer and the time the developers require to acquire information about the subject.

Launch:

After the system has been finished and been approved by the customer, it can be launched so that the customer can start using it him/herself. This can be done by setting up two servers, one web server containing the frontend of the system. Another server that will be used to process all data, protected from any outsiders except for the frontend server. This server needs to be able to run Java. After setting up these servers, an executable of the backend will be created and uploaded to the server. With that a Postgres database will be set up on the same backend server. These will then be started after which the backend is operational. The frontend then only needs to be given the correct ip-configuration in order to be able to communicate with the backend server, after which the system will be operational and ready for use. All of this will take approximately 15 hours to do and 8 hours to test. These hours are calculated within the earlier mentioned costs.

Bibliography

Areas of Focus | Broad Institute of MIT and Harvard. (n.d.). Retrieved April 28, 2016, from https://www.broadinstitute.org/what-broad/areas-focus/areas-focus

O'Donoghue, S. I., Anne-Claude, G., Nils, G., Goodsell, D. S., Jean-Karim, H., Nielsen, C. B., … Bang, W. (2010). Visualizing biological data—now and in the future. *Nature Methods*, *7*(3s), S2–S4.

Scientists and Research. (n.d.). Retrieved April 28, 2016, from http://www.k-rith.org/scientists-research