Bilkent University

Department of Computer Engineering

Senior Design Project

LIBRA: Genetic Filtering and Diagnosis Matching System

Project Specifications Report

Mahmud Sami Aydın, Berke Egeli, Naisila Puka, Halil Şahiner, Abdullah Talayhan

Supervisor: Can Alkan
Jury Members: Abdullah Ercüment Çiçek and Hamdi Dibeklioğlu

Project Specifications Report
Oct 14, 2019

This report is submitted to the Department of Computer Engineering of Bilkent University in partial fulfillment of the requirements of the Senior Design Project course CS491/2.

Table of Contents

1 Introduction

Many hospitals, laboratories and medical research centers want to collect and store genomic data, as well as explore and interpret that data based on specific needs. There are open-source programs developed and utilized for such purposes that have been accepted by the authorities [1]. Yet, they are not suitable for direct use and the installation requires specific expertise.

The collective data produced by these institutes possess valuable information related to genetic profiles. These genetic profiles can be useful for comparing and diagnosing rare diseases. For example, a child in San Francisco Bay Area had a rare disease caused by not being able to produce tears. The doctors were suspicious about a gene called NGLY1 after the results of genetic profiling. Yet, they were not sure about the cause because of the lack of genetic profiles of other patients for comparison. The issue was resolved after finding a patient with similar phenotypes studied by Duke University and NGLY1 was indeed the gene causing the disease [2]. Examples like this have created a high demand for genomic discovery through the comparison of genotypic/phenotypic profiles.

The aim of LIBRA is to provide a user-friendly genetic filtering and annotation system equipped with a genetic profile matching platform, that can be quickly integrated and easily used by medical institutions in order to explore genetic variation, detect and diagnose rare diseases, as well as safely collect and store their data.

1.1 Description

LIBRA is going to be a web application composed of two main modules: *Genetic Variation Query Interface* and *Patient Matching Platform*. The first module provides an interface for storing and annotating genomic data in order to query variants and explore the data, whereas the second one acts as a patient social network for doctors who seek similar genetic profiles related to a specific disease in order to understand and diagnose the disease further. These modules will be integrated into the same user interface. The potential users of this project are medical doctors in medium-sized hospitals, laboratories and research centers in Turkey.

1.1.1 Genetic Variation Query Interface

This module stems from an existing framework called GEMINI [3]. GEMINI is a framework for exploring genetic variation, in which genetic variants are loaded using VCF files. It automatically

annotates the variants by using existing databases.[1] This module of LIBRA aims to improve this framework. Doctors will be able to perform variation queries from a convenient user interface supplied by the web application. LIBRA will support annotation by comparing with all open genetic databases such as *1000 genomes*, *dbSNP*, *dbVar*, *ClinVar*, *OMIM*, *COSMIC*, *ENCODE* etc. in order to become a common access point for making genetic variation queries.

GEMINI uses a legacy portable database. This database is not compatible with our security and scalability requirements, as will be explained in more detail in the following sections. Therefore, LIBRA will reconstruct this infrastructure using a modern database structure, later extended to a distributed one. LIBRA will be an *SaaS* (Software as a Service) product, which implies that the users will not be responsible for setting up the databases and configuration files.

LIBRA will also have an integrated query editor that will enable users to request real time queries. Running basic SQL queries will be supported. The additional benefit of LIBRA's query editor will be the support of *Genotype Query Tools* [4] which provides faster queries computed over genomes represented in a specific format (compressed bitmap).

1.1.2 Patient Matching Platform

In this module, doctors can create accounts for patients as in social networks. The main purpose of the module is to help doctors understand rare diseases by comparing different patients with similar phenotypes. This module will make use of MatchMaker API [5] protocols in order to make queries for genotype/phenotype comparison and matching. Additional data for phenotypes will be supplied by using Human Phenotype Ontology (HPO), which provides a standardized vocabulary of phenotypic abnormalities encountered in human disease [6].

In this module, the privacy of the patients is a main concern. Further elaboration will be done regarding privacy and security in the following sections.

1.2 Constraints

1.2.1 Implementation Constraints

LIBRA will be implemented using Python Django as backend supplied with React JS as frontend. The underlying database will be PostgreSQL, later extended to a distributed version.

[1] **Genome annotation** is the process of identifying the locations of genes so that it serves as an explanation about the functionality of genes.

1.2.2 Economic Constraints

The primary economic constraints on this project will be imposed by the cost associated with server hosting and maintenance of the web application. The servers will be hosted on Google Cloud Platform [7].

1.2.3 Privacy and Legal Constraints

The application should comply with the instated laws and regulations of the host country/international community, such as KVKK (Kişisel Verilerin Korunması Kanunu) in Turkey [8].

1.2.4 Health and Safety Constraints

The system will handle any unexpected hardware/network failure while annotating genetic variants and make sure that no wrong result is displayed to the doctors. In this way, the system will never cause the doctors to diagnose the patient with an incorrect disease since that may impose a health and safety risk to the patient.

1.2.5 Sustainability Constraints

In order to increase the capacity of our system to endure through time, we are mainly focused in the underlying database. The more scalable the database is, the more sustainable it will be when compared with the exponential data growth (in particular genomic data) within the years. Also, updates according newly released versions will contribute in sustainability of the system.

Moreover, the system supports concerns in the above-mentioned constraints (legal concerns, economic concerns, etc), which makes the software available to continue operating in the future.

1.3 Professional and Ethical Issues

There are not many commercial products like this project. Similar products exist to meet research demands. The existing products are either provided as additional services or are sold at high prices. Our project aims to provide a ubiquitous and uniform service to hospitals around Turkey to allow diagnosis, research and treatment of diseases, especially rare ones.

Privacy of the patients is important. Certain genetic variations and diseases can lead to the discrimination against the patients in their social and professional lives. Therefore, it is imperative that the privacy of the patients is protected. Data will not be utilized other than the intended

purposes of LIBRA. The doctors will filter what they want to share regarding their patients. The information shared between doctors in the platform will be restricted to ensure only the permitted data will be displayed.

2 Requirements

2.1 Functional Requirements

2.1.1 User Accounts

- Users will be able to create an account in LIBRA given that they work in a hospital that LIBRA accepts (a list of hospitals will be prepared).
- After the hospital is accepted by LIBRA, they will provide personal and work information.
- A verification email or SMS will be sent to the hospital the user claims to be working at, based on the hospital's contact data found in LIBRA.

2.1.2 Genetic Variant Upload and Automatic Annotation

- Users will be able to load genetic variants of their patients of interest to LIBRA using a specific format (VCF).
- After loading, they will be given the list of the databases supported for annotation and the users will choose the ones they prefer.
- The genetic variants will then be automatically annotated by comparing them to several online genome annotation sources such as dbSNP, KEGG, etc.

2.1.3 Annotation Based Variant Querying

- Users can query variants based on specific requirements related to variants' attributes since the underlying system will organize the genotypes and annotations.
- Users can choose to write their own query or use LIBRA's built-in query editor. This editor provides most common genome queries (e.g. filtering on genotypes, finding which samples have a specific variant, variants with specified allele frequency percentage, etc.) Users will be able to modify these query templates through the editor.

- Users can also customize their own queries and add it as a template for future queries.
- Users will be able to combine results from different queries through the editor.
- Users can save previous query results and use them on another query.

2.1.4 Custom Annotation and Variant Analysis

- Users will be able to annotate the variants with their own specific annotated file, which might describe genome regions particularly relevant to user's purpose.
- Users will be able to share this annotated file with another hospital based on their desire. In this way, the users belonging to the other hospital will also be able to annotate their variants with that custom file.
- Users can run analytics tools in the system since the underlying system will support analytical queries. These analytics tools include identifying potential variants related to some specific disorder (e.g. compound heterozygotes cause many autosomal recessive disorders [9]).
- Users can choose to locally save a descriptive file of the run analytics.

2.1.5 Patient Profiles

- Users will be able to create accounts for their patients in LIBRA with the national ID of the patient if there is no account for that patient in the system.
- After creating the account, users will enter the information of the patient based on user's stored genomic data in LIBRA. While doing that, they can decide which information will be shared with matchmaker system.
- If patient account is already available, users can edit the current information of the patient, e.g. add new diseases for the patient.
- Users can alter/delete their patient accounts.

2.1.6 Patient Matching

- Users will be able to search for similar patients based on customized attributes of the genomic profile of their patient of interest.

- Users can also run customized filters on the matching results in terms of information points shared through the databases to focus on different groups of patients.

2.2 Non-Functional Requirements

2.2.1 Scalability

The Gemini tool uses SQLite database. In order to increase the scalability of our system we plan to upgrade the infrastructure by switching to PostgreSQL database, later extended to the distributed version.

2.2.2 Backup and Recovery

The hospitals should be able to do their backup on their local database. Also the server needs to do regular backups for the accounts of doctors and their patients in the genetic matching platform.

2.2.3 Availability

The application should be accessible to users at all times, with only the possible exceptions of server maintenance.

2.2.4 Accessibility

The application should provide language support for Turkish language since the application's focus group is medical doctors in Turkey.

2.2.5 Reliability

The system will ensure that the comparison of the patients' information to find matches for diseases must give reliable results and handle unexpected failures. That is, if the system fails while doing comparison (server/client side errors), the error will be reported and the comparison procedure will handle the process appropriately, making sure only reliable matches/mismatches will be displayed.

2.2.6 Portability

The system should be compatible with different browsers. This means that LIBRA will be developed as a platform independent web application.

References

[1] J. E. Stajich and H. Lapp, "Open source tools and toolkits for bioinformatics: significance, and where are we?" *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 287–296, Sep. 2006. [Online]. Available: https://doi.org/10.1093/bib/bbl026

[2] "Crying without tears unlocks the mystery of a new genetic disease - scope," Mar. 2014. [Online]. Available: https://scopeblog.stanford.edu/2014/03/20/crying-without-tears/

[3] U. Paila, B. A. Chapman, R. Kirchner, and A. R. Quinlan, "GEMINI: Integrative exploration of genetic variation and genome annotations," *PLoS Computational Biology*, vol. 9, no. 7, p. e1003153, Jul. 2013. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003153

[4] R. M. Layer, N. Kindlon, K. J. Karczewski, and A. R. Quinlan, "Efficient genotype compression and analysis of large genetic-variation data sets," *Nature Methods*, vol. 13, no. 1, pp. 63–65, Nov. 2015. [Online]. Available: https://doi.org/10.1038/nmeth.3654

[5] O. J. Buske, F. Schiettecatte, B. Hutton, S. Dumitriu, A. Misyura, L. Huang, T. Hartley, M. Girdea, N. Sobreira, C. Mungall, and M. Brudno, "The matchmaker exchange API: Automating patient matching through the exchange of structured phenotypic and genotypic profiles," *Human Mutation*, vol. 36, no. 10, pp. 922–927, Sep. 2015. [Online]. Available: https://doi.org/10.1002/humu.22850

[6] S. Köhler, L. Carmody, and N. V. et al., "Expansion of the human phenotype ontology (hpo) knowledge base and resources," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1018–D1027, Nov. 2018. [Online]. Available: https://doi.org/10.1093/nar/gky1105

[7] "Cloud computing services — google cloud." [Online]. Available: https://cloud.google.com

[8] "Kişisel verilerin korunması kanunu," Apr. 2016. [Online]. Available: https://www.mevzuat.gov.tr/MevzuatMetin/1.5.6698.pdf

[9] "comp_hets: Identifying potential compound heterozygotes." [Online]. Available: https://gemini.readthedocs.io/en/latest/content/tools.html#comp-hets-identifying-potential-compound-heterozygotes