DEVELOPMENT OF A WEB APPLICATION TO SEARCH FOR GENE CLUSTERS IN GENOMIC DATABASES

Bc. Marek Koubek

Master’s thesis
Faculty of Information Technology
Czech Technical University in Prague
Department of Software Engineering
Study program: Informatics
Specialisation: Software Engineering
Supervisor: Ing. Jiří Novák, Ph.D.
January 5, 2026

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of master's thesis

Title:	Development of a Web Application to Search for Gene Clusters in Genomic Databases
Student:	Bc. Marek Koubek
Supervisor:	Ing. Jiří Novák, Ph.D.
Study program:	Informatics
Branch / specialization:	Software Engineering
Department:	Department of Software Engineering
Validity:	until the end of summer semester 2026/2027

Instructions

The aim of this master's thesis is to develop a web application for searching gene clusters in genomic data, based on the existing open-source software CluSeek. CluSeek is a cross-platform desktop application written in Python, designed for identifying colocalized genes in genomic databases available online. The thesis will focus on adapting and extending the existing data workflow for a new web-based application that can also perform local (offline) genomic data analysis.

Within the scope of the thesis, the student will:
1. Study the biological background of gene clusters in microorganisms and the structure of genomic data in the NCBI databases (Genbank/nuccore and IPG).
2. Analyze the data workflow of CluSeek application (available at http://cluseek.com) and the functionality of the web application CAGECAT (https://cagecat.bioinformatics.nl/).
3. Design a solution for refactoring the CluSeek desktop application into a web-based version, deployable also as a standalone desktop application with support for local (offline) genomic data analysis.
4. Implement the proposed solution using appropriate technologies for modern web application development.
5. Verify the functionality of the application and subject it to user testing.

Electronically approved by Ing. Michal Valenta, Ph.D. on 13 November 2025 in Prague.

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Koubek Marek. *Development of a Web Application to Search for Gene Clusters in Genomic Databases.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2026.

Chtěl bych poděkovat především sit amet, consectetuer adipiscing elit. Curabitur sagittis hendrerit ante. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Cras pede libero, dapibus nec, pretium sit amet, tempor quis. Sed vel lectus. Donec odio tempus molestie, porttitor ut, iaculis quis, sem. Suspendisse sagittis ultrices augue.

Declaration

FILL IN ACCORDING TO THE INSTRUCTIONS. VYPLŇTE V SOULADU S POKYNY. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Curabitur sagittis hendrerit ante. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Cras pede libero, dapibus nec, pretium sit amet, tempor quis. Sed vel lectus. Donec odio tempus molestie, porttitor ut, iaculis quis, sem. Suspendisse sagittis ultrices augue. Donec ipsum massa, ullamcorper in, auctor et, scelerisque sed, est. In sem justo, commodo ut, suscipit at, pharetra vitae, orci. Pellentesque pretium lectus id turpis.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Curabitur sagittis hendrerit ante. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Cras pede libero, dapibus nec, pretium sit amet, tempor quis. Sed vel lectus. Donec odio tempus molestie, porttitor ut, iaculis quis, sem. Suspendisse sagittis ultrices augue. Donec ipsum massa, ullamcorper in, auctor et, scelerisque sed, est. In sem justo, commodo ut, suscipit at, pharetra vitae, orci. Pellentesque pretium lectus id turpis.

In Prague on January 5, 2026

Abstract

Fill in the abstract of this thesis in English. Lorem ipsum dolor sit amet. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Cras pede libero, dapibus nec, pretium sit amet, tempor quis. Sed vel lectus. Donec odio tempus molestie, porttitor ut, iaculis quis, sem. Suspendisse sagittis ultrices augue.

Keywords enter, comma, separated, list, of, keywords, in, ENGLISH

Abstrakt

Fill in the abstract of this thesis in Czech. Lorem ipsum dolor sit amet. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Cras pede libero, dapibus nec, pretium sit amet, tempor quis. Sed vel lectus. Donec odio tempus molestie, porttitor ut, iaculis quis, sem. Suspendisse sagittis ultrices augue.

Klíčová slova enter, comma, separated, list, of, keywords, in, CZECH

Contents

List of Figures

List of Tables

List of Code listings

List of abbreviations

ABR	Abbreviation meaning
SMTH	Something

Chapter 1
Analysis

In this chapter the current state of the CluSeek application is described including the explanation of the main microbiology background. The chapter also includes the description of CAGECAT web application, an alternative tool to CluSeek. The second half of this chapter focuses on indetifying the application goals and specifications which are described in form of functional and nonfunctional requirements, use case diagram, activity diagram and domain model.

1.1 Problem statement

The main goal of CluSeek is to provide a simple and easy to use tool for identifying gene clusters in GenBank data. The main flaw of the current solution is the accessibility of the app, since the user needs to download and install the application to use it. Porting the application to the web solves this problem and it also enables the user to use the application from any device regardless of the operating system. This is also beneficial from the point of maintainability of the application.

NOTE: Shortened version of the problem statement, more will be added in the future.

1.2 Microbiology in context of CluSeek

This section focuses on description of the main microbiology concepts needed for the understanding of the application. One of the subsections focuses on the general terms used in the field of microbiology, while the other focuses on the way how such information and related data are stored in public online databases such as NCBI or IPG..

1.2.1 Genes, gene clusters and more

Deoxyribonucleic acid (DNA) is a sequence of nucleotides. A nucleotide is a monomeric unit composed of a phosphate group, a deoxyribose sugar, and one of four nitrogenous bases: adenine, cytosine, guanine, or thymine.

The transfer of genetic information involves two key processes: transcription and translation. Transcription is the process of duplicating a segment of DNA into ribonucleic acid (RNA). RNA is similar to DNA but consists of guanine, cytosine, adenine, and uracil (instead of thymine). There are different types of RNA, such as messenger RNA (mRNA), which codes for protein sequences, and non-coding RNA, which performs functions itself.

Translation is the process in biological cells in which proteins are produced using RNA molecules as templates. The mRNA is decoded in a ribosome, where transfer RNA (tRNA) bonds to the mRNA and releases amino acids. These amino acids chain together to create polypeptides, which are long, continuous, unbranched peptide chains. Proteins are polypeptides that have a molecular mass of 10,000 Da or more.

The sequence of nucleotides (A, C, G, U) is mapped to amino acids via the genetic code. A sequence of three nucleotides, known as a codon, translates to a single amino acid. There are specific codons that start the translation process (start codons) and others that stop it (stop codons). While there are approximately 500 amino acids, only about 20 appear in the genetic code.

A gene is defined as a segment of DNA that is transcribed into RNA and eventually translated into a protein. In the context of this application, a gene cluster refers to a group of such genes that are physically close to each other in the genome and often function together in a specific metabolic pathway.

1.2.2 Genomic data and databases

Usefull links:

GenBank `https://www.ncbi.nlm.nih.gov/genbank/`

BLAST `https://blast.ncbi.nlm.nih.gov/Blast.cgi`

- **BLAST Guide** `https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf`
- **ElasticBLAST** `https://blast.ncbi.nlm.nih.gov/doc/blast-help/cloudblast.html`
- **ElasticBLAST Guide** `https://link.springer.com/article/10.1186/s12859-023-05245-9`
- **BLAST+ Docker Documentation** `https://github.com/ncbi/blast_plus_docs`

IPG

1.3 Current state of CluSeek desktop application

1.3.1 Technology and documentation

All of the code is written in Python language. The documentation of code is basically nonexistent. The code itself contains a few comments, but without any structure or organization. The code also lacks type annotations.

1.3.2 Code

Currently the application structure does not follow any specific design pattern. The file structure is quite shallow and the code is not properly organized.

```
/
├── __init__.py
├── __main__.py
├── about.py ....................................... licensing information
├── asdf ................................................... mystery file
├── dbdl.py .................................... database manipulation
├── dframe.py ................................... data handling classes
├── monitoring.py ................................................ a
├── uiqt.py .................................. UI and application logic
├── diamond ........................... DIAMOND alignment tool binaries
│   ├── diamond-linux64-2
│   ├── diamond-osx-arm64
│   └── diamond-win64.exe
└── uilayouts .................................... UI layout definitions
    ├── *.ui .................................. xml UI layout definitions
    ├── *.py ............................... python UI layout definitions
    └── *.png ................................................ png icons
```

The `about.py` file contains information licencing information of the software and third party libraries used in the application. It serves mainly as a text file.

The `dframe.py` contains all the classes used for data handling and processing, including `Taxon`, `Scaffold`, `Feature`, `GeneticRegion`, `Protein`, `IdenticalProteinSet`, `ReferenceDummy`, `ProteinClusterHierarchy`, and `ProteinCluster`. Not all of these classes are used in the application.

The `dbdl.py` file contains implementations regarding the manipulation with database. The file contains few sections that are not relevant to the original purpose of this file and should be moved to a separate file.

The `uiqt.py` file contains most of the application logic, though it should contain only the logic of the user interface. This file needs a major refactorization and should be split into multiple files. Currently the business logic is not separated from the UI logic, which violates the single responsibility principle and the open/closed principles of software design.

1.3.3 Data workflow

1.3.4 UI/UX

There is no official guide or user manual for the application. However, the application contains descriptive tooltips for most of the elements. UI is implemented using Qt framework.

The UI of the CluSeek application is built using the Qt framework. The `uilayouts` directory contains `.ui` files, which are XML descriptions of widget hierarchies (buttons, tables, layouts), and corresponding Python scripts.

The main application workflow centers around a tabbed interface containing:

- **Info tab**: Displays metadata for selected elements.
- **Gene clusters tab**: The main visualization view, split into headers (taxa/sequence IDs) and contents (arrows representing genes).
- **Protein groups tab**: A summary view of protein groups.
- **Filtering tab**: Controls for including or excluding clusters based on protein composition.

1.4 CAGECAT web application

1.4.1 Data workflow

1.4.2 UI/UX

1.4.3 Technology and documentation

1.5 Application goals

1.6 Functional and nonfunctional requirements

1.6.1 Functional requirements

1.6.2 Nonfunctional requirements

1.7 Use case diagram

1.8 Activity diagram

1.9 Domain model

1.10 TODO

■ **Table 1.1** Software Architecture Considerations for Cloud/Web Migration

Category	**Questions to Answer**	**Cloud/Web Impact**
Data Input	How does it get data? (User-typed string, local file path, or database?)	Web apps must handle "Uploads" rather than local paths.
State Management	Does it use global variables or save things to self in a class?	Multiple users will use the same code; variables must be "session-aware."
CPU/RAM Usage	Is it a light calculation or a heavy processing task?	Heavy tasks need their own dedicated server/worker.
Execution Time	Does it run in <1 sec, or does it take minutes?	Tasks over 10s need a "Background Worker" and a loading bar.
Output Type	Does it print to console, show a GUI window, or save a file?	Needs to be converted to JSON for the frontend or a downloadable URL.

Appendix A
Nějaká příloha

Sem přijde to, co nepatří do hlavní části.

Obsah příloh

```
/
├── readme.txt .................. stručný popis obsahu média
├── exe ......... adresář se spustitelnou formou implementace
├── src
│   ├── impl .................... zdrojové kódy implementace
│   └── thesis ......... zdrojová forma práce ve formátu LaTeX
└── text ...................................... text práce
    └── thesis.pdf ................ text práce ve formátu PDF
```