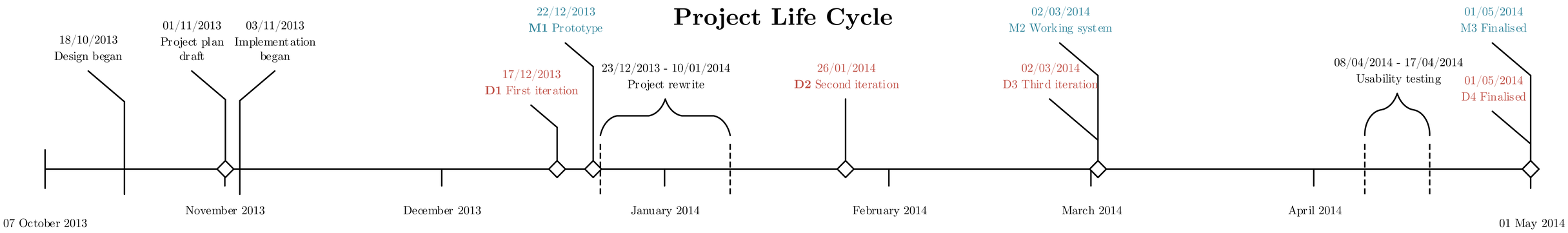
Protein Isoelectric Point Database

Chris Cummins | Supervisor: Ian Nabney | Moderator: Zuoyin Tang | MEng Electronic Engineering & Computer Science

Project Life Cycle



Introduction

The purpose of the project is to make an intuitive and powerful bioinformatics search engine which provides online access to a large dataset of protein isoelectric points which has been compiled by Aston University researchers and students over the course of several years.

Deliverables

1. An updatable relational database warehousing a set of protein isoelectric point data.
2. A web-accessible GUI with searching and downloading functionality.

Objectives

1. To build a free (as in freedom) web application for searching and viewing protein isoelectric points.
2. To produce a bioinformatics tool with real world value for future research.
3. The application should provide intuitive but powerful searching facilities.
4. The application should provide a convenient means for a certified user to edit and upload additional data.
5. The application should present information in a usable and efficient form.
6. Users should be allowed to download generated results for offline use.
7. Adequate security precautions should be taken to minimise the risk of data being sabotaged or stolen.
8. The implementation should use a clean model view controller architecture.
9. Comprehensive test coverage of the API and common use cases should be automated.
10. The application should be scalable for much larger datasets.

Overview

This is an interdisciplinary software project with potential value for molecular biology researchers. There are many similar existing projects, and so a user-centred approach to design was coupled with an incremental development pattern in order to ensure that the project offers real benefits compared to existing tools. Additionally, a significant amount of time was invested into developing the tools and infrastructure for large software projects. Work for this project fell within three categories:

Process

- Human-centred design.
- Rapid prototyping.
- Hi & Lo-fidelity prototyping.
- Usability for bioinformatics.
- Open source.
- Version control and accountable development.
- Issue tracker.
- Usability testing.

Infrastructure

- Homogeneous build and deploy system.
- Autotools and inotify events.
- Test dataset generation.
- Unit testing framework & coverage.
- Online hosting.
- Task automation and pipbot.
- YAPS encoding and toolset.

Product

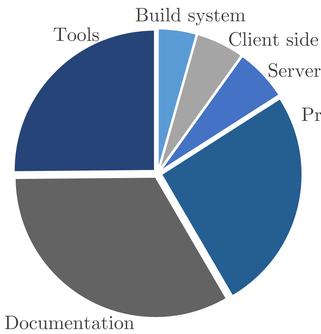
- Search engine design and implementation.
- SQL back-end.
- LISP server.
- Client-side JavaScript.
- BLAST+ searching
- Passive controller.
- JSON API.
- Literate programming style.

Summary

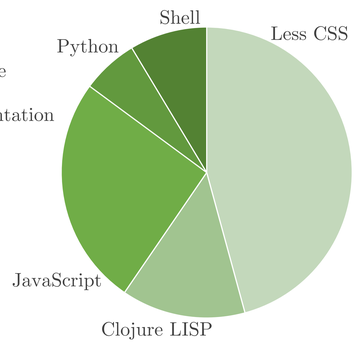
This is a polylingual project including source code written in Clojure LISP, JavaScript, Less CSS, M4sh, Make, Python, and shell programming languages. Key project deliverables include:

1. A schema and toolset for storing protein isoelectric point data, and an online instance of the supplied dataset.
2. A web-accessible search interface incorporating FASTA sequence searching and a public API.
3. A build system framework for web application back-ends.
4. A toolset for generating plausible test payloads using machine learning techniques with scientific datasets.

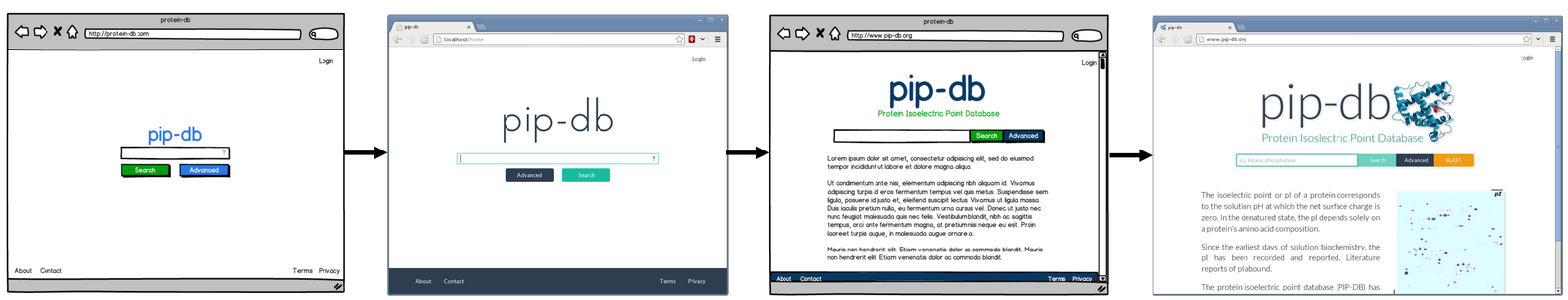
Project composition



Languages used



Product Life Cycle



Paper prototype | **First implementation** | **Refined design** | **Completed implementation**