Finding repetition in genomes

Report Name	Finding repetition in genomes
Author (User Id)	Raeesah Khan (rak12)
Supervisor (User Id)	Amanda Clare (afc)
Module	CS39440
Degree Scheme	G400 (Computer Science)
Date	February 10, 2023
Revision	2.0
Status	Release

---

1 Project description

The finding repetition in genomes project will develop a web-based application for bioinformaticians so that they are able to find microsatellites in genomes. Genomes are stored as strings in FASTA files, using string manipulation bioinformaticians can analyse the genomes and are then able to make conclusions about the genomes that they've collected. Microsatellites (also known as Simple Sequence Repeats (SSRs)) are repetitive sections of DNA that repeat throughout the genome. Microsatellites are important for bioinformatics because they can be used to advance our understanding of biological systems, genetic diversity, and evolutionary processes. Furthermore, microsatellites are also useful for identifying and tracking genetic changes in populations, including those that may be important for human health or that are associated with diseases. Most commonly, Microsatellites are considered to be between 2 and 7 bases long. Currently, most of the software used to locate microsatellites are relatively old.

This project will initially find microsatellites within this range. It is essential that this project is able to read a FASTA file and the user is able to upload their file to the website and receive a GFF file as an output with the results. The user may be limited on the file size of their upload. The user should also be able to add a penalty score for deviations as this project needs to take into account microsatellites that have slight deviations from each other ie. ATGCC and ATACC may be counted as the same microsatellite but the final match score may have a penalty to reflect the deviation. The user should also be able to decide the K-mer size of the microsatellite. Finally, the project should display the results graphically if the user wishes.

For this project, it would be nice to have an input where the user is able to upload a matrix for the penalty score, so that different variations have different penalties, for example, if there is a base A in place of base C in the microsatellite the program adds a greater penalty then if the A was replaced with T. Furthermore, it would be nice to be able to give users the option of displaying the results in multiple different graphical methods. Additionally, it would be nice to be able to find microsatellites of longer lengths if time permits the project will try and find microsatellites between 7 and 10 bases long.

The project will use a Kanban approach to manage the tasks. The project will have weekly Monday reflections and tasks such as backing up work and writing down the past week's work. The project will also be managed with weekly Tuesday individual meetings and Thursday group meetings with the project supervisor.

2 Proposed tasks

Initial set up: It is vital to set up the project so that the workflow is consistent. The initial set up includes the following tasks.

- **Set up Kanban board**: This task will first require investigation on physical Kanban boards versus virtual boards. If the board is virtual then the tasks requires some investigation on virtual Kanban boards. This task requires decisions that need to be made about what columns should be added to the Kanban board. The other tasks that are currently known should be added to the Kanban board with labels indicating whether the task is research, documentation, coding, general or urgent.

- **Set up weekly diary/review**: Setting up a diary for the weekly Monday review is simple. This document is going to be written on weekly and would include a section about work done the previous week, reflections on the work and what to do to improve for the next

week. This document is going to be written using a word processing software, most likely Word or LaTex.

- **Setting up local build environment and version control system**: To set up a version control system for this project Git could used, either GitLab, which the university provides, or Github. Either way, the project supervisor needs to be added to the Git repository.

Project meetings and diary: This project will have weekly individual and group meetings in addition to a weekly review which will be written down in the project diary so that the tasks completed are tracked and then the diary can be used to write up the final report.

Research and investigation: This task requires investigation for the following topics; microsatellites, legacy software that finds microsatellites, the advantages and disadvantages of using Python or C for this project and suitable visual formats to present the results. This task also requires some spike work to be completed, such as, reading a FASTA file, producing a GFF file and finding 2-mers in a FASTA file.

Design: This task will be split into two sections, designing the implementation of the website and designing the implementation of the program that will find microsatellites.

Implementation: This task requires the following work, creating code to read FASTA file, inputs data in a GFF file format, find micro satellites of sizes between 2 and 7, display results using a graph, set up a website for users to input data and download their results.

3 Project deliverables

Mid-Project Demonstration reflection: A written reflection on the mid-project demonstration will be produced and included in the final project report and will include what was discussed, feedback and plan for improvement.

Microsatellite program: A microsatellite program will be produced that is at least able to read a FASTA file and find microsatellites that are at least 2-7 bases long and produce a graphical representation of the results.

Microsatellite Website: A website will be produced at least to allow a user to upload a FASTA file, with limitations, and download a GFF file.

Project Report: The project report will include sections on design, implementation, testing and conclusion. This project will detail the motivations behind choices made during this project, how the project was implementation and what was used to implement the project.

Diary: A diary will be produced and included in the project report, the diary summarises each weeks work and will showcase the work done and the process followed.

Test files: Test files will be produced and included on the projects version control software. The test files will be specifically designed for this project to test the program.

Final Demonstration: A PowerPoint presentation and other additional visual tools will be produced for the final demonstration to showcase the work done.

Annotated Bibliography

[1] B. D. Pickett, J. B. Miller, and P. G. Ridge, "Kmer-SSR: a fast and exhaustive SSR search algorithm," *Bioinformatics*, vol. 33, no. 24, pp. 3922–3928, 08 2017. [Online]. Available: https://doi.org/10.1093/bioinformatics/btx538

Kmer-SSR is a current tool to find microsatellites.

[2] B. D. Pickett, S. M. Karlinsey, C. E. Penrod, M. J. Cormier, M. T. W. Ebbert, D. K. Shiozawa, C. J. Whipple, and P. G. Ridge, "SA-SSR: a suffix array-based algorithm for exhaustive and efficient SSR discovery in large genetic sequences," *Bioinformatics*, vol. 32, no. 17, pp. 2707–2709, 05 2016. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw298

Details an algorithm that uses suffix arrays to find microsatellites.

[3] X. Wang and L. Wang, "Gmata: an integrated software package for genome-scale ssr mining, marker development and viewing," *Frontiers in plant science*, vol. 7, p. 1350, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2016.01350/full

Current software used to find microsatellites.

[4] M. L. C. Vieira, L. Santini, A. L. Diniz, and C. d. F. Munhoz, "Microsatellite markers: what they mean and why they are so useful," *Genetics and molecular biology*, vol. 39, pp. 312–328, 2016. [Online]. Available: https://www.scielo.br/j/gmb/a/HQFhwk9mttszSG7th39ygmS/?format=html&lang=en

Overview of what microsatellite are and why they are important to bioinformaticians.