Investigating Patterns in the LRMDA Gene Utilizing the 2-Jump Algorithm

Computer Science 450 Project Proposal

Derik Dreher
Department of Computer Science
University of Northern British Columbia
Prince George, B.C.
ddreher@unbc.ca

Duncan Render
Department of Computer Science
University of Northern British Columbia
Prince George, B.C.
drender@unbc.ca

***Abstract*—An approach to substring mining and analysis using the 2-Jump algorithm in order to find patterns that may help in the investigation of diseases such as autosomal recessive oculocutaneous albinism type-7. To model and extrapolate information in a user-accessible format using Python, Kivy, and an NCBI, Nucleus-provided dataset.**

***Index Terms*—substring, mining, albinism, 2-Jump, analysis, Python**

Introduction

Leucine-Rich Melanocyte Differentiations Associated (LRMDA) is a gene responsible for encoding proteins. In the past, mutations in the transcript variant 2 mRNA of this gene have been suspected to be involved with diseases including autosomal recessive oculocutaneous albinism type-7 [1].

This report outlines our procedure in building an application that allows for DNA sub-sequence pattern matching, frequent pattern mining using a proprietary algorithm – perhaps creating a tool that can be used to investigate protein chains that may be involved in these diseases. Furthermore, it describes extending our solution to any text-based DNA sequence.

Objective

The objective of the project was to create an efficient way to mine our dataset containing a DNA sequence such that we could perhaps find a notable substring that may indicate the origins of oculocutaneous albinism; doing this led us to implement the 2-jump algorithm.

Originally our objective was localized to the aforementioned, but later extended to including all frequent patterns for further information on possible interesting substrings. Furthermore, we wanted to display the information in an easy-to-use and user-friendly manner; this later extended to data visualization.

Requirements

- Understand data mining techniques
- Understand biological terminology
- Competency in the Python programming language
- Teamwork

Challenge Components

One of the major challenges with this project was understanding the pseudocode for the algorithm; furthermore, after implementing the algorithm to the best of our ability, it was very apparent that the algorithm did not work as intended. This led to us having to build upon the paper and develop our own algorithm based on the ideas of the published algorithm.

Another challenge was learning the Python programming language. As we both have more experience with more statically-typed languages (and low-level languages in the case of Duncan), learning the syntax and oddities of Python was challenging at first.

The final challenge relates to learning the Kivy framework as we both have more experience working with raw data and algorithms rather than front-end development. This required learning a new language - the Kivy markup language - as well as interacting with an unfamiliar application programming interface (API).

Methodology

For this project, we used the Agile programming methodology and incorporated the practice of pair programming. The timeline for completion of notable events is as follows:

1) **September 21st** Initial research for project ideas
2) **September 27th** Literature survey of selected topics
3) **October 2nd** Research on suitable pattern matching algorithms
4) **October 8th** Project proposal presentation
5) **October 19th** Formal project proposal
6) **November 2nd** Initial design plans and layout for project
7) **November 4th** Initial implementation of algorithm
8) **November 17th** Project progress presentation
9) **November 18th** Major changes to design plan
10) **November 23rd** Algorithm rewritten and validated
11) **November 25th** Prototype GUI created
12) **December 2rd** Project validation and debugging

13) **December 4th** Presentation preparation and frequent pattern miner added
14) **December 5th** Project presentation
15) **December 6th** GUI finalization and project final report

DELIVERABLE

An easy-to-use frequent pattern miner and substring sequence searcher for any string-based DNA sequence; a desktop GUI that can be run platform independent.

LEARNING OUTCOMES

Communication and Teamwork

In order to avoid conflicting on components of our application, we had to effectively communicate our changes and what we were working on to each other. This was an effective way of enforcing efficient communication and good teamwork.

Time Management

With both of us having extremely loaded semesters, with eleven courses between the two of us, managing time was extremely difficult. This project was a great experience for exposure to high-stress, high-workload situations.

Exposure to Biology

As computer science stays relatively localized to pure algorithms and mathematics, exposure to a separate area of science was beneficial in widening our breadth of knowledge. The exposure to biology also showed how computer science can be directly used for advancing other fields.

Application of Data Mining Techniques

Substring mining is already a subset of data mining, but what we had learned in data mining had a direct application in our project. This gave good context as to how important data mining is in current computer science avenues.

REFERENCES

[1] Grønskov, K., Dooley, C., Østergaard, E., Kelsh, R., Hansen, L., Levesque, M., Vilhelmsen, K., Møllgård, K., Stemple, D. and Rosenberg, T., 2020. Mutations In C10orf11, A Melanocyte-Differentiation Gene, Cause Autosomal-Recessive Albinism. [Online]. Available at: https://pubmed.ncbi.nlm.nih.gov/23395477/ [Accessed: 20-Oct-2020].

[2] Bhukya, Raju & Somayajulu, Dvln. (2011). 2-Jump DNA Search Multiple Pattern Matching Algorithm. International Journal of Computer Science Issues. 8. [Online]. Available: https://www.researchgate.net/publication/268325804_2-Jump_DNA_Search_Multiple_Pattern_Matching_Algorithm [Accessed: 20-Oct-2020]

[3] "Homo sapiens leucine rich melanocyte differentiation associated (LRMDA - Nucleotide - NCBI," National Center for Biotechnology Information, Sep-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NM_032024.5?report=genbank [Accessed: 20-Oct-2020].