**Using a machine learning model to predict coding and non-coding sequence in DNA**

|  |  |
| --- | --- |
| Report Name | Project Outline |
| Author (User Id) | Tommy Boyle(tob31) |
| Supervisor | Wayne Aubrey (waa2) |
|  |  |
| Module | CS39440 |
| Degree Scheme | G40F (Computer Science with foundation year) |
|  |  |
| Date | February X, 2022 |
| Revision | 0.1 |
| Status | Draft |

# Project description

Analyse the presence and frequency frequency of tandem stop codons in a genetic sequence and their impact on gene expression. Tandem stop codons are two or more stop codons located in close proximity in a genetic sequence, and they can affect gene expression by altering the process of translation. This project will focus on understanding the impact of tandem stop codons on gene expression and how they can be used as markers for gene regulation.

# Proposed tasks

1. Data Collection: Collect a large dataset of genetic sequences, including both coding and non-coding regions.

2. Data Preprocessing: Preprocess the collected data to prepare it for analysis, including cleaning, normalizing, and filtering the data as necessary.

3. Tandem Stop Codon Identification: Develop an algorithm to identify tandem stop codons in the genetic sequences and quantify their frequency.

4. Gene Expression Analysis: Analyze the impact of tandem stop codons on gene expression, including the effect of codon density and the influence of tandem stop codons on the translation process.

5. Comparison with Existing Methods: Compare the results of this project with existing methods for detecting tandem stop codons and assessing their impact on gene expression.

# Project deliverables

1. Detailed report: A comprehensive report documenting the methodology, results, and conclusions of the project, including a discussion of the implications for gene regulation and future research directions.

2. Source code: A complete and well-documented source code for the algorithms developed during the project, including the tandem stop codon identification algorithm and the gene expression analysis algorithm.

3. Data sets: A collection of the data sets used in the project, including the raw data, preprocessed data, and final results.

4. Presentation: A presentation summarizing the results and conclusions of the project, suitable for a final project review.

5. Poster: A poster summarizing the results and conclusions of the project, suitable for display at a scientific conference or other event

# Initial annotated bibliography

The following is a simple list, i.e. not using EndNote or Microsoft Word’s Referencing tool. You could insert any citations as cross-references in Word [1][3][3][4].

1. Sylvia Duckworth. A picture of a kitten at Hellifield Peel. <http://www.geograph.org.uk/photo/640959>, 2007. Copyright Sylvia Duckworth and licensed for reuse under a Creative Commons Attribution-Share Alike 2.0 Generic Licence. Accessed August 2011.  
     
   *This is my annotation. I should add a description here.*
2. Mark Neal, Jan Feyereisl, Rosario Rascunà, and Xiaolei Wang. Don’t touch me, I’m fine: Robot autonomy using an artificial innate immune system. In *Proceedings of the 5th International Conference on Artificial Immune Systems*, pages 349–361. Springer, 2006.   
     
   *This is my annotation. I should add a description here.*
3. W.H. Press et al. *Numerical recipes in C*. Cambridge University Press Cambridge, 1992.  
     
   *This document…*
4. Various. Fail blog. <http://www.failblog.org/>, August 2011. Accessed August 2011.  
     
   *This is my annotation. I can add comments that are in* ***bold*** *as well as italics. It isn’t just the formatting – do mention what is useful about the resource.*