AUT UNIVERSITY COMPUTER + MATHEMATICAL SCIENCES

Master of Computer and Information Sciences

INFS812

BIOINFORMATICS

ASSESSMENT ONE

Exercise on Bioinformatics Tools and Methods

Data Analysis of Bioinformatics Dataset

Date: Friday 5 September 2019, 11pm (40%)

Submission: through *Turnitin* in AUTOnline

Students are required to complete the assignment on an independent basis. Any references cited in the assignment must be given in full. APA 6th style should be used for referencing.

Task 1 (20 marks)

Search and discover the information about a human genetic disease of interest to you using bioinformatics tools and public databases

1. Choose a human genetic disease/disorder and give its short description (correct citation is required).
2. Find the information about the genes that are involved with your chosen disease. There might be several genes that satisfy this criterion. You only need to pick up one gene for this assignment using the tool - Online Mendelian Inheritance in Man (OMIM) or another bioinformatics knowledge database, such as ebi, NHGRI (http://www.genome.gov/). Record the related information of the gene in your assignment sheet.
3. Give the genomic context information of the selected gene, including the locus, its adjacent genes, the graphic of the gene's chromosomal regions, etc.
4. Find two (one human and one of another species) protein sequences of the selected disease gene in FASTA format and record the relevant information. {You will find it easier to do this in conjunction with task 2.1.}

Task 2 (30 marks)

1. Using Blast to compare the sequences related to the chosen human genetic disease. *The sequences are selected from Task 1.4.*
   Please record your findings from the experiments into the assignment report.
2. Use Needleman-Wunsch algorithm for sequence alignment. Use the Needleman-Wunsch algorithm based Pairwise alignment tool (http://www.ebi.ac.uk/Tools/psa/) to compare the similarity between the two selected protein sequences from Task 1.4.
3. Compute the longest common sequence (LCS) between two protein sequences. These two protein sequences are chosen from Task 1.4.

Task 3 (50 marks)

Find a bioinformatics (gene expression or another biological) dataset for a supervised classification problem from a public repository and apply at least three classification models on it for **knowledge discovery**.

Investigation tasks:

- Describe the data (number of variables, number of samples, dimensionality).
- Understand the problem space and how you can connect it to real world.
- Pre-processing of the data, if necessary (Missing Values, Normalisation, Rebalancing)
- Data Visualisation (correlation evidence from visual analysis of attributes)
- Feature selection/extraction (PCA, LDA). Show at least two and plot the features in 2D/3D space.
- Classification (Use three classification algorithms and describe them in brief. The limitations and strengths. Perform parameter tuning. **Reasons for selecting the algorithms**).
- Performance Metrics used for Data Analysis and Knowledge Discovery.
- What new knowledge you were able to discover from the data? How you can develop on it further? (Hint: Fuzzy Rules, Rule Extraction)
- Inference (What is the reason behind doing Classification? **Motivation** behind selecting the dataset and performing data analysis tasks).

*In order to receive feedback on this assessment you are ***encouraged to submit a draft version (does not take any marks!).***