Cairo University

Faculty of Computers and Artificial Intelligence

Computer Science Department

Computational Analysis for RNA-Seq of Plant Organisms

PHD. Thesis

Submitted by

Heba Mohammed Zaki

Supervised by

Prof. Dr. Hesham Ahmed Hassan
Computer Science Department
Faculty of computers and Artificial Intelligence, Cairo University

Prof. Amr Anwar Badr
Computer Science Department
Faculty of computers and Artificial Intelligence, Cairo University

Dr. Mohammad Nassef
Computer Science Department
Faculty of computers and Artificial Intelligence, Cairo University

Dr. Ahmed Farouk Al-Sadek
Agricultural
Research Center, ARC

A thesis submitted to the Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science.

November / 2020

Table of Contents

List of Figures

List of Tables

Abstract

As a revolutionary technology for life sciences, RNA-seq has many applications and the computational pipeline has also many variations. RNA-seq combines simultaneous transcript identification and quantification of a large number of genes in a single assay. Consequently, the actual RNA-seq data analysis also has many variations, depending on the applications and studied organisms. Computational techniques are used widely to help in the RNA-seq analysis process for better understanding genes behavior regarding different biotic and abiotic stress conditions. Usage of computational methods has led to saving a lot of time, effort and money for biologists.

This thesis aims at maximizing benefit from RNA-seq data analysis, via usage of computational methods which can be applied on any organism. The work presented in this thesis is divided into two parts: the first part towards enhancing the differential expression analysis using edgeR and Fisher criterion (FC) analysis methods to obtain more reliable expressed genes, and second part investigates the relationship between the expression level of genes and the biological features of their SNPs using Rough set theory.

First part suggests a workflow for RNA-seq analysis to identify differentially expressed genes. This workflow is applied on the analysis of *A. thaliana* plant under heat-stress conditions. The identified candidate genes are validated via two popular references; DRASTIC and TAIR10. Results suggest that edgeR and FC methods can be combined to perform differential expression analysis for RNA-Seq data, without strong assumptions. Moreover, new promising (23) genes have been identified through comparison of results, which are unreported as heat-stress genes.

Second part suggests general guidelines for accurate SNP discovery from RNA-seq data. Those SNPs annotations are used to find relation between their biological features and the differential expression of the genes to which those SNPs belong via Rough set. This strategy is applied on the analysis of SNPs in *A. thaliana* plant under heat-stress conditions. Rough sets are utilized to define this kind of relationship into a finite set of rules. Set of (32) generated rules proved good results with strength, certainty and coverage evaluation terms. The result increases the amount of knowledge for SNPs discovery and analysis in functional genomics research.