Discovering the Relationship between Heat-Stress Gene Expression and Gene SNPs Features using Rough Set Theory

Heba Zaki[1], Ahmed Farouk Al-Sadek[4]
Agricultural Research Center (ARC)
Giza, Egypt

Mohammad Nassef[2], Amr Ahmed Badr[3]
Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt

***Abstract*—Over the years of applying machine learning in bioinformatics, we have learned that scientists, working in many areas of life sciences, call for deeper knowledge of the modeled phenomenon than just the information used to classify the objects with a certain quality. As dynamic molecules of gene activities, transcriptome profiling by RNA sequencing (RNA-seq) is becoming increasingly popular, which not only measures gene expression but also structural variations such as mutations and fusion transcripts. Moreover, Single nucleotide polymorphisms (SNPs) are of great potential in genetics, breeding, ecological and evolutionary studies. Rough sets could be successfully employed to tackle various problems such as gene expression clustering and classification. This study provides 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. 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. This strategy is applied to the analysis of SNPs in A. thaliana plant under heat-stress.**

Keywords*—RNA sequencing (RNA-seq); variant calling; Single Nucleotide Polymorphisms (SNPs) analysis; rough sets; gene expression

I. Introduction

RNA sequencing (RNA-seq) technology has resulted in exceptionally fast and wide scale analysis of the genetic information exists in all organisms. This mainly includes the concurrent study of alternative splicing, Single nucleotide polymorphisms (SNPs) and differential expression. The approach of genome-guided transcriptome has been the standard RNA-seq analysis method for model organisms like A. thaliana. Some existing software packages are available to perform this task [1]. New tools are continuously developed to be used for RNA-seq analysis task starting from reads alignment ending with the pathway analysis mission. Unfortunately, some non-expert users for those tools cannot get the full power and capabilities of them on wide scale [2].

SNPs are single nucleotide base variations, caused by transitions or transversions, in the same position between individual genomic sequences. Genetics and breeding are the most two important studies using SNPs as significant molecular markers. In genetic studies, SNPs are ideal genomic resources used for functional gene identification for traits and characterization of genetic resources because of their extensive genome distribution, wide density and, high scalability [3]. Fortunately, SNPs discovery can be accomplished on both approaches of genome-guided and de novo on variety of organisms [4]. This is applied on many plants, including those with little or no available genetic information.

Among the various benefits of performing SNPs analysis using RNA-seq data, there are two important ones [5]. First the reasonable cost for simultaneous discovery of thousands SNPs together with expression levels of functional genes at the same time. Second is involving phenotypes which can be predicted according to genotypes and, the location of SNPs in coding regions related to the possibly identified plant biological and agronomical traits.

RNA-seq is considered the ideal method for gene expression profiling [6] and, it is commonly used for precision medicine due to its high capability of measuring dynamic gene activity in the genome for a specific tissue type. Moreover, when applying RNA-seq on some disease tissue samples [7], it detects most of mutations exist in expressed genes that are related to disease biology.

Machine learning techniques can support very interesting and critical analysis applications dedicated for the fields of molecular biology and bioinformatics. Particularly, rough set method is considered very commonly used for this task of data analysis due to its flexibility in handling qualitative data. Rough Set theory was proposed in 1982 by Z. Pawlak [8] and, has been used as a methodology of database mining or knowledge discovery. It can contribute in many processes like attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction.

Rough Set uses information system or information table to represent data. This table consists of objects (rows) and attributes (columns) [9]. There are two types of attributes named as the condition attribute and decision attribute. Each row of an information table defines a decision rule, which specifies the decision attribute values when conditions are indicated by condition attributes are satisfied. Additionally, a set of objects is classified using rough set theory by finding dependencies and relations between attributes [10]; reduction of unnecessary attributes; discovering the most important attributes; or by decision rule generation.

Rough set-based rule generation provides easier explanations and descriptions for complex biological systems [8]. The challenges of those complex systems can be summarized into determining the features or attributes that can demonstrate the biological phenomenon, and what combinations of features' values can define that phenomenon and make a significant added value to the system study [11]. The possible decision values can be the participation of some particular genes in a biological process. Furthermore, learning the set of minimum features that can determine the gene involvement in this process may be interesting issue for some biologists. High throughput problems have a great care about discovering which features, in which order and, in which combinations define decisions.

The main motivation for this research is to provide aid to find a reasonable answer about the relationship between expressed genes and SNPs under the effect of heat-stress phenomenon. This relationship is achieved through determining the set of features that best describe this biological process. Rough-set based rule induction method is applied on RNA-seq data for the A. thaliana plant.

The rest of the paper is organized as follows. Related work is reviewed in Section II. Section III describes the data used in the research and tools adopted to perform SNPs detection and analysis. Methodology and techniques utilized for SNPs Detection phases as well as SNPs analysis are discussed in Section IV. The results of the experiments in the form of evaluation terms, tables and charts are discussed in Section V. Section VI provides our derivations, outcome on the study and, recommendations for the future work.

II. Related Work

Various research efforts in the literature have been targeted to the two main focal topics of this research; SNPs identification and Rough set theory in bioinformatics. This section lists a summarization of these efforts as follows.

The authors in [7] investigated the most suitable method that can provide the greatest number of SNP calls with high specificity and sensitivity. Following the steps of alignment sequence reads to the genome, removing duplicates, and using SAMtools to call SNPs had achieved the required purpose. SAMtools proved higher consistency than GATK with 8–10% more variants identification.

Plant functions, related to climate adaptation, have leading genes involved in transcriptional mechanisms. In the study [12], they realized that neat and strong peaks of association were identified in expected functional variants in the extreme tail of genetic differentiation. Those results proved that climate adaptation can mainly cause the genomic variation when applied on A. thaliana at a small scale.

SNP-ML (SNP machine learning) suggested in [13], a novel tool, predicted true SNPs from sequence data using machine learning. It was designed for calling more trusted SNPs from polyploids. Moreover, it provided SNP machine learner (SNP-MLer), a functionality to train new models for customized use. Tetraploid peanut SNPs were identified using SNP-ML, and the validated true- and false-positive SNP mapping data improved the discovery process.

Another research [2] suggested Visualization Pipeline for RNA-seq analysis (VIPER) that combined stages of an RNA-seq analysis workflow. This workflow graded from raw RNA-seq data, then quality control and genome alignment, reaching to the differential expression and pathway analysis. VIPER listed the most popular tools used in the workflow like, RSEM for quantification, and SNPeff for annotating identified SNPs.

A reasonable amount of work has been performed on the usage of rough set methodology in solving bioinformatics issues and challenges [14]. These studies have focused on problems of classification and reduction of bioinformatics data. Some other literatures have dealt with topics related to selection of genes, classification of protein sequence and, prediction of protein structure.

A novel approach for tumor classification was proposed by [15]. This approach was based on Wavelet Packet Transforms (WPT) and Neighborhood Rough Sets (NRS) as tools for effective features extraction and selection. WPT performed features extraction, and then decision tables are formed. High classification with few attributes was reduced by NRS. The proposed method was applied on three gene expression datasets and experimental results showed feasibility and effectiveness.

A feature selection algorithm based on rough set theory had been suggested in [16]. It depended on selecting reduced set of genes from microarray data based on relevance and significance criteria of the selected genes. The importance of rough set theory here was computing both criteria to produce theoretical analysis justification. The proposed algorithm performance, along with a comparison with other related methods had obtained 100% predictive accuracy for three cancer and two arthritis data sets.

Suitable solutions were provided in [17] to solve two important issues exist in the data represented in information table. Those two solutions were applied based on concepts exist in Rough Set Methodology. The first issue was the indiscernible objects that were represented several times and solved by data reduction. This reduction included eliminating the unnecessary attributes and deletion of identical rows. The second issue was the existence of many redundant attributes and solved by dimensionality reduction. This solution used simplifying discernibility function to get reducts which used for generating if-then rules for classification.

A Promising framework was introduced in [18] to handle the complexities of protein structure prediction. Rough set improved harmony search quick reduct algorithm to be used for selecting the optimum number of features. More compact rules were generated via Rough set classification which showed a higher overall accuracy rates compared with classification algorithms in Weka.

III. Datasets and Materials

This section lists the datasets with their types and full description of the experiment conditions. Moreover, the tools, and computational power needed for accomplishing this study are presented. It involves Data Sources, Software Packages and Tools and, Computational Requirements.

A. Data Sets

The A. thaliana reference genome FASTA sequences have been downloaded from the Ensemble FTP (https://plants.ensembl.org/info/website/ftp/index.html). The RNA-seq FASTQ data files for A. thaliana under heat-stress were downloaded from the NCBI website. These data files represent an experiment that is performed on A. thaliana plants in Moscow, Russia. A Third leaf was collected from 15 plants of age 21 days after heat treatment at 42°C for 1, 3, 6, 12, and 24 hours. The experiment was accomplished with 2 replicates for each of the mentioned 5 different time points to resolve false-positive calls at the low end of signal detection [5]. This experiment was SINGLE stranded – Illumina Hi-Seq 2000 – RNA-seq libraries from TRANSCRIPTOMIC PolyA RNA. Ten files were downloaded and their attributes and description are listed in Table I.

B. Software Packages and Tools

The following list contains software tools and packages that were integrated with custom code to carry out the execution of the various processes along the presented workflow.

- STAR (Spliced Transcripts Alignment to a Reference): 2.5.3a [March 17, 2017] version available on BA-HPC.
- SAMtools (Sequence Alignment/Map tool): 1.9-intel-b [2018] version available on BA-HPC.
- BCFtools (Binary Counterpart Format tools): 1.9-foss-b [2018] version available on BA-HPC.
- SnpEff (variant annotation and effect prediction tool): 4.1d using (Java-1.7.0_80) [2015] version available on BA-HPC.
- Rosetta: version 1.4.41 [May 27 2001].

C. Computational Requirements

The experiments conducted in this research are based on the Unix-type operating systems (primarily Linux); it provides a command-line interface and is best run on a high-memory, multicore computer or in a high-performance computing environment. In general, having ~1 GB of RAM per 1 million paired-end reads is recommended. A typical configuration is a multicore server with 256 GB to 1 TB of RAM.

TABLE. I. FILES OF RNA-SEQ READS

Symbol	Replicate Name	Accession
H1_R1	1 hour Replicate 1	SRX1881868
H1_R2	1 hour Replicate 2	SRX1881876
H3_R1	3 hours Replicate 1	SRX1881880
H3_R2	3 hours Replicate 2	SRX1881883
H6_R1	6 hours Replicate 1	SRX1881886
H6_R2	6 hours Replicate 2	SRX1881888
H12_R1	12 hours Replicate 1	SRX1881889
H12_R2	12 hours Replicate 2	SRX1881897
H24_R1	24 hours Replicate 1	SRX1881908
H24_R2	24 hours Replicate 2	SRX1881912

For the research problem presented in this article, the lack of the required computing resources to accomplish the required work could be a challenge. In this study, the used computational resources were provided by The Bibliotheca Alexandrina (bibalex)[1]. The super computer BA-HPC capabilities are used to achieve this work.

IV. METHODOLOGY

This study proposes a promising framework to illustrate how SNPs can be discovered, annotated and, analyzed from RNA-seq data in order to be used to describe genes expression. Methodologies are divided mainly into two phases: (A) SNPs Detection, and (B) SNPs Analysis. Details of both phases and needed resources are being described below.

A. SNPs Detection

This phase shows an overview of the steps and methods that are employed to identify the most suitable performing of RNA-seq SNPs detection pipeline in Fig. 1. The employed steps start from creating genome indices, and go along till finding out annotated SNPs.

1) Creating genome indices: Using the A. thaliana reference genome (*.fna) file from (Data Sets) section, Indices are created using STAR tool.

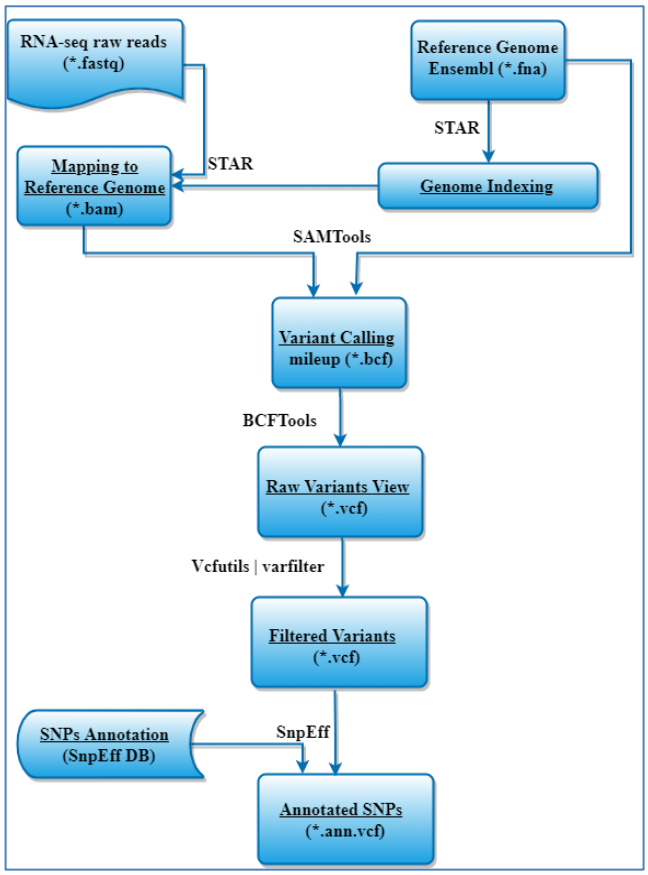

Fig. 1. A Framework for RNA-Seq SNP Detection.

[1] BA-HPC group, Alexandria Library, https://www.bibalex.org/, Egypt, (accessed Oct 2019).

2) Mapping raw reads to reference genome: This step aims to find matches between the reference genome and the sequences of the sampled RNA-seq short reads. During mapping, using existing gene models obtains the maximum advantage to some read mapper in order to map the coordinates accurately.

Using indices files generated by *STAR*, each *individual* read with the reference genome is mapped. Convert mapped reads from SAM to BAM, sort, and index. BAM files are generated sorted by coordinates, so they can be loaded much more quickly.

3) Variant calling: Find deviation from reference genome, the output of both previous steps (RefGenome indices and mapped reads) are used together to perform the variant calling task. This step is done using SAMtools which uses the mpileup command to compile information about the bases mapped to each reference position. It collects summary information in the input BAMs, computes the likelihood of data given each possible genotype and stores the likelihoods in the BCF format Output BCF file is a binary form of the text Variant Call Format (VCF).

4) Obtaining raw variants: BCF file came from the previous step is converted into VCF file using *BCFtools.* It is a collection of utilities to call SNPs and manipulate VCF files. Those utilities are calling SNPs and small indels, annotating and sub-selecting entries from VCF files, querying, filtering, merging VCF files, and converting BCF to human-readable VCF. VCF file has a nice header explaining what the columns mean. Below that header, there are rows of data describing potential genetic variants. Fig. 2 shows a sample for one of the produced (*.vcf) files header and content.

Header contains mandatory lines like the first line (1) and the line containing columns' headers (30). Lines (4) and (5) in the shown sample include data about the reference genome and bam files used to get that vcf variant calling. While lines from (6-12) have the contigs or chromosomes of the reference genome and length of each of them. Moreover, there are optional lines that describe some meta-data about the information in the VCF body shown in Table II.

SAMtools/BCFtools may write the following fields in the 'INFO' tag in VCF/BCF.

- DP: The number of reads covering or bridging POS.
- INDEL: Indicating the variant is an indel.
- I16: contains 16 integers like; sum of reference base qualities, sum of ref mapping qualities, sum of tail distance for ref bases, etc.

5) Variant filtering: This step applies the prior and does the actual calling. It performs filtering short variants using vcfutils.pl varFilter. This filtration contains; delete duplication, remove low-quality reads (defined by sequencing device), filter unmapped reads and, filter low quality reads/mappings.

6) Finding out annotated SNPs: The last step in the phase of SNPs Detection is to discover the categorization of the variants effects in genome sequences. SnpEff (an abbreviation of "SNP effect") tool is able to analyze and annotate thousands of variants per second and predict their possible genetic effects. Since many databases containing genomic annotations are available with SnpEff distribution, the SNPs annotation is called through SnpEff DB. The output information provided using SnpEff (*.ann.vcf) includes some additional lines at the end of the header which are concerned with the annotation part, as presented in Fig. 3. In this research, 'ANN' field is the main target which includes the information needed to determine the value of the variant as shown in Fig. 4.

```
1  ##fileformat=VCFv4.2
2  ##FILTER=<ID=PASS,Description="All filters passed">
3  ##samtoolsVersion=1.9+htslib-1.9
4  ##samtoolsCommand=samtools mpileup -g -f Arabidopsis_thaliana.TAIR10.dna.toplevel.fa /home/cairo02lul/data/HeatMapping/M_SRR3724768/Aligned.sortedByCoord.out.bam
5  ##reference=file://Arabidopsis_thaliana.TAIR10.dna.toplevel.fa
6  ##contig=<ID=1,length=30427671>
7  ##contig=<ID=2,length=19698289>
8  ##contig=<ID=3,length=23459830>
9  ##contig=<ID=4,length=18585056>
10 ##contig=<ID=5,length=26975502>
11 ##contig=<ID=Mt,length=366924>
12 ##contig=<ID=Pt,length=154478>
13 ##ALT=<ID=*,Description="Represents allele(s) other than observed.">
14 ##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
15 ##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
16 ##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
17 ##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
18 ##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
19 ##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
20 ##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
21 ##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
22 ##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
23 ##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
24 ##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
25 ##INFO=<ID=I16,Number=16,Type=Float,Description="Auxiliary tag used for calling, see description of bcf_callret1_t in bam2bcf.h">
26 ##INFO=<ID=QS,Number=R,Type=Float,Description="Auxiliary tag used for calling">
27 ##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
28 ##bcftools_viewVersion=1.9+htslib-1.9
29 ##bcftools_viewCommand=view M_SRR3724768.bcf; Date=Thu Oct 10 12:13:05 2019
30 #CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  /home/cairo02lul/data/HeatMapping/M_SRR3724768/Aligned.sortedByCoord.out.bam
31 1   3673    .   C   <*> 0   .   DP=2;I16=2,0,0,0,70,2458,0,0,40,800,0,0,7,49,0,0;QS=1,0;MQ0F=0   PL  0,6,36
32 1   3684    .   A   <*> 0   .   DP=2;I16=2,0,0,0,79,3121,0,0,40,800,0,0,29,445,0,0;QS=1,0;MQ0F=0    PL  0,6,36
33 1   3695    .   A   <*> 0   .   DP=2;I16=2,0,0,0,79,3121,0,0,40,800,0,0,42,884,0,0;QS=1,0;MQ0F=0    PL  0,6,36
```

Fig. 2. VCF File Header and Content.

```
##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"
##SnpEffCmd="SnpEff  Arabidopsis_thaliana /home/heba/Arab_Data/SNP_detection/M_SRR3724768/M_SRR3724768.vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type |
Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS /
INFO' ">
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene_Name | Gene_ID |
Number_of_transcripts_in_gene | Percent_of_transcripts_affected'">
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Format: 'Gene_Name | Gene_ID |
Number of transcripts in gene | Percent of transcripts affected'">
```

Fig. 3. Header of (*.ann.vcf) File.

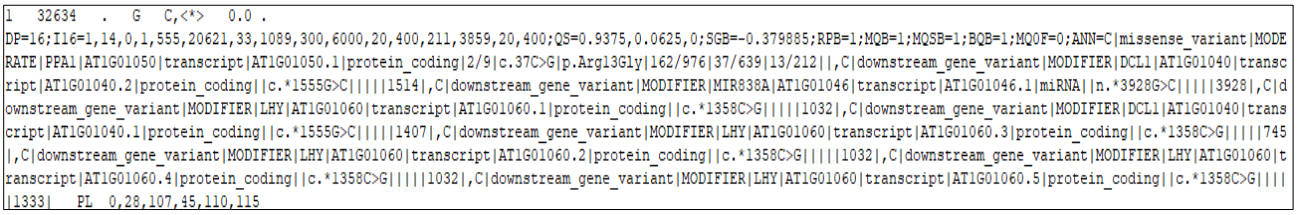

Fig. 4. Content Sample of 'ANN' Field.

TABLE. II. VCF META-DATA

Tag	Description
CHROM	No. of chromose that variant belongs to
POS	Position of that variant on that chromosome
REF	Reference sequence at POS involved in the variant
ALT	Comma delimited list of alternative sequence(s)
QUAL	Phred-scaled probability of all samples being homozygous reference
INFO	Semicolon delimited list of variant information

B. SNPs Analysis

The input of this phase is the 10 (*.ann.vcf) files created in the first phase, which include analysis ready variants. This phase includes set of well-ordered processes which are applied to determine the relationship between SNPs biological features and gene expression.

1) Adjustment: This process handles the 10 (*.ann.vcf) files and put them into another flexible form enabling the separation of some specific information to be analyzed (INFO, ANN) tags.

2) Separation of variants from indels: The main goal is analyzing SNPs and their effect on the genome sequence. So, indels are removed to focus on SNPs only.

3) SNPs selection: Choose only SNPs located in the common heat-stress genes of A. thaliana, published in reference databases; DRASTIC[2] and TAIR10[3].

4) Detection of SNPs biological features: Some biological features of SNPs mainly describe the biological value of the detected SNP. They are isolated and been prepared for analysis. Table III lists some of those features, description and, their possible values.

5) Discovery of relationship between SNPs' features and genes differential expression using Rough Set: The most suitable technique to represent this kind of relation is Rough set. Rough set theory has been a methodology of database mining or knowledge discovery in relational databases [9]. The target is to find the set of rules that translate the relationship between the values of the biological features of detected SNPs for some gene and the differential expression of the same gene.

[2] Gary Lyon, The DRASTIC gene expression database, http://www.drastic.org.uk, (accessed Nov 2018).

[3] The Arabidopsis Information Resource (TAIR), http://www.Arabidopsis.org, (accessed Oct 2019).

Rough Set Analysis approach has many important advantages like; Discovery of hidden patterns in data, Data reduction (finds minimal sets of data), Evaluating the importance of data, Representing data as sets of decision rules and, Providing the interpretation of obtained result [19].

Rosetta is a general-purpose tool that is not geared towards any particular application domain. The name ROSETTA can be construed as an acronym, for a Rough Set Toolkit for Analysis of Data. It has been put to use by a large number of researchers world-wide, and has resulted in scientific publications in a wide variety of areas. Moreover, it implements features relevant to build and evaluate rough set models in different domains, and offers a highly user friendly environment in which to conduct experiments. In this study, Rosetta is used to generate rough set rules for the predicted SNPs. This will be discussed obviously in the (Generation of Rough set rules) section [20].

6) Measuring the generated rules: To quantify the generated rules, several numerical measures for the rules are illustrated in Definition 1, 2, and 3 and described in Table IV [21] [22].

S is called a decision table, which is denoted by $S = (U, C, D)$. They are called C, D condition and decision attributes, respectively.

Definition 1: Let $S = (U, C, D)$ be a decision table, $\Phi \in For(C)$ and $\Psi \in For(D)$ .The expression *if* Φ *then* Ψ is called a *decision rule* and is denoted by $\Phi \rightarrow \Psi$.

Definition 2: Let $S = (U, C, D)$ be a decision table and $\Phi \rightarrow \Psi$ a decision rule in S. The *certainty factor* of this rule is defined as:

$$Cer_S(\Phi, \Psi) = \frac{card(\| \Phi \wedge \Psi \|_S)}{card(\| \Phi \|_S)}$$

It is obvious that $0 \leq Cer_S(\Phi, \Psi) \leq 1$ for every $\Phi \rightarrow \Psi$. This coefficient is widely used in data mining and is called confidence coefficient too.

Definition 3: Let $S = (U, C, D)$ be a decision table and $\Phi \rightarrow \Psi$ a decision rule in S. The coverage factor of this rule is defined as:

$$Cov_S(\Phi, \Psi) = \frac{card(\| \Phi \wedge \Psi \|_S)}{card(\| \Psi \|_S)}$$

It is obvious that $0 \leq Cov_S(\Phi, \Psi) \leq 1$ for every $\Phi \rightarrow \Psi$.

TABLE. III. SNPS BIOLOGICAL FEATURES

Feature	Description, and Possible value(s)
Gene Name	Common gene name (PPA1)
Gene ID	Gene ID (AT1G01050)
Annotation (effect or consequence)	Annotated using Sequence Ontology (SO) terms (e.g. chromosome_number_variation, exon_loss_variant, stop_gained, stop_lost, start_lost, etc.)
Annotation_impact	A simple estimation of putative impact / deleteriousness : {HIGH, MODERATE, LOW, MODIFIER}
Feature type	Which type of feature is in the next field (e.g. transcript, motif, miRNA, etc.). It is preferred to use SO terms
Transcript biotype	The bare minimum is at least a description on whether the transcript is {Coding, Noncoding}
cDNA_position / (cDNA_len)	Position in cDNA and trancript's cDNA length
Protein_position / (Protein_len)	Position and number of AA

TABLE. IV. EVALUATION MEASURES FOR ROUGH SET RULES

Measure	Description
Rule Support	The number of samples that represent this rule
Rule Strength	The Rule Support divided by the total number of samples. (The more cases support a rule, the stronger it is)
Rule Certainty (accuracy)	the frequency of objects having Ψ in the set of objects with the property Φ
Decision Coverage	the frequency of objects with the property Φ in the set of objects with the property Ψ

V. RESULTS AND EVALUATION

A. Identification of SNPs in Heat-Stress Genes

Continuous decision values may cause a challenge. In most practical approaches, there are about two to five decision classes. So, if the problem has continuous decision values, they can be split into 2 or 3 intervals [11]. In another example of exon expression values, the decision experimentally was split into three classes by taking 20%: 60%: 20% corresponding to highly expressed, medium expressed and low expressed exons [23]. Similarly, in this study the decision is divided only into two classes, by taking the highest expressed (Yes): the lowest expressed (No) genes.

DRASTIC and TAIR10 reference databases are used as trusted sources for the highest expressed heat-stress genes for A. thaliana. The union of heat-stress genes in those two databases are (225) unique genes. Next, all SNPs that are located into those set of genes exist in the resulting 10 (*.ann.vcf) files are being selected. The number of heat-stress genes detected in each replicate and also numbers of their identified SNPs are listed in details in Table V.

For balance, an equalized set of the lowest expressed heat-stress genes are selected from work presented in [24], to analyze their SNPs features too. About (200) genes are selected and their SNPs are got from the 10 (*.ann.vcf) files. The number of the lowest expressed heat-stress genes detected in each replicate and, number of their identified SNPs are listed in details in Table VI.

B. Capturing Biological Features of Identified SNPs

After that, the biological features of SNPs for both groups of genes, for the highest and lowest expressed heat-stress genes, are picked up from the 10 (*.ann.vcf) files. The most effective biological features exist in these annotation files due to the rough set are (Gene Name, Gene ID, Annotation, Annotation Impact, Feature Type and, Transcript BioType). The top-ranked attributes (biological features) are used to build a rule-based classifier using the Rosetta system.

C. Generation of Rough Set Rules

An information system or information table can be viewed as a table, consisting of objects (rows) and attributes (columns). The captured set of SNPs are used to discover finite set of rules that can describe whether the genes, those SNPs belong to, are heat-stress or not. Rules were generated in *Rosetta* [20] with the manual reducer which determines decision rules (Heat Expressed: Yes; No) based on characterization of a set of objects in terms of attribute values (SNPs biological features). Table VII explores the given replicates and their total number of objects, number of (Yes) decision, number of (No) decision, and the number of resulting rules. Total number of Rules (251) represents the sum of all rules generated over the 10 replicates. However, the set of non-repeatable rules shared between the 10 replicates is (32) rules.

Table VIII presents samples of the resulting rules after applying the rough set characterization. It shows the values of the chosen condition attributes based on the Rough set reduction, and the decision attribute for each rule.

D. Rules Evaluation

To quantify the generated rules, three main numerical parameters for the rules are defined: Rule Strength, Rule Certainty (accuracy) and, Decision Coverage.

Those parameters are calculated for the generated set of rules by applying Definition 1, 2, 3 and, Table IV mentioned in section (Measuring the generated rules). The pie chart shown in Fig. 5 displays the Rule Strength of all rules, showing rules that have the highest strength percentages. Moreover, Fig. 6 shows the different Rule Certainty in a bar chart. Rules that have the same condition values but different decision value (Yes, No) are represented in adjacent bars. Finally, Decision Coverage of rules is represented in Fig. 7 for (Yes) decision rules and Fig. 8 for (No) decision rules.

TABLE. V. THE HIGHEST EXPRESSED HEAT-STRESS GENES AND THEIR SNPS IN ALL REPLICATES

Rep. Name	H1_R1	H1_R2	H3_R1	H3_R2	H6_R1	H6_R2	H12_R1	H12_R2	H24_R1	H24_R2
No.Genes	140	64	172	171	184	186	186	190	183	173
No. SNPs	2353	440	5158	2971	4860	2799	3930	3020	3446	3041

TABLE. VI. THE LOWEST EXPRESSED HEAT-STRESS GENES AND THEIR SNPS IN ALL REPLICATES

Rep. Name	H1_R1	H1_R2	H3_R1	H3_R2	H6_R1	H6_R2	H12_R1	H12_R2	H24_R1	H24_R2
No.Genes	95	22	134	144	157	159	165	170	158	144
No. SNPs	424	58	889	898	1181	1013	1241	1195	1177	853

TABLE. VII. REPLICATES OBJECTS AND THEIR RULES

Rep. Name	Objects	(Yes)	(No)	Rules
H01_R1	2775	2351	424	24
H01_R2	498	440	58	16
H03_R1	6047	5158	889	23
H03_R2	3863	2965	898	27
H06_R1	6037	4856	1181	24
H06_R2	3806	2793	1013	26
H12_R1	5164	3923	1241	27
H12_R2	4210	3015	1195	28
H24_R1	4622	3445	1177	29
H24_R2	3891	3038	853	27
Total	**40913**	**31984**	**8929**	**251**

TABLE. VIII. SAMPLE OF GENERATED RULES

Annotation	Annotation_Impact	Feature_Type	Transcript_BioType	Heat_Expressed
missense_variant	MODERATE	Transcript	protein_coding	Yes
stop_lost	HIGH	Transcript	protein_coding	Yes
downstream_gene_variant	MODIFIER	Motif	protein_coding	No
intergenic_region	MODIFIER	intergenic_region	Noncoding	Yes
start_lost	HIGH	Transcript	protein_coding	No

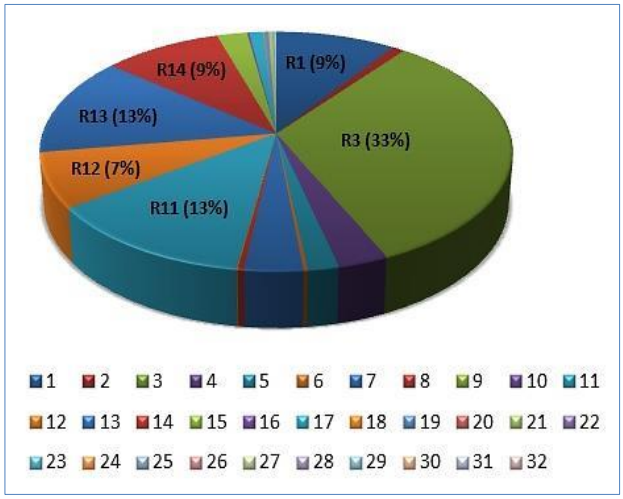

Fig. 5. Rules Strength.

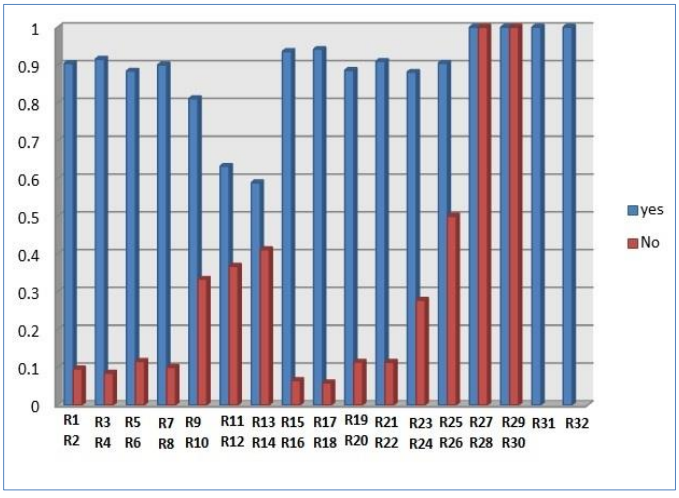

Fig. 6. Rules Certainty.

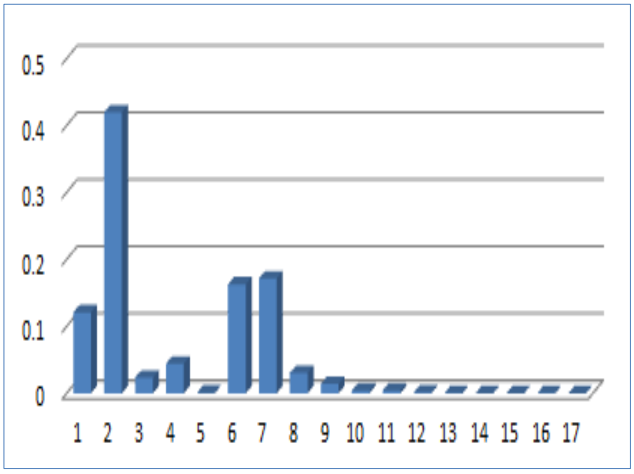

Fig. 7. Decision Coverage (Yes).

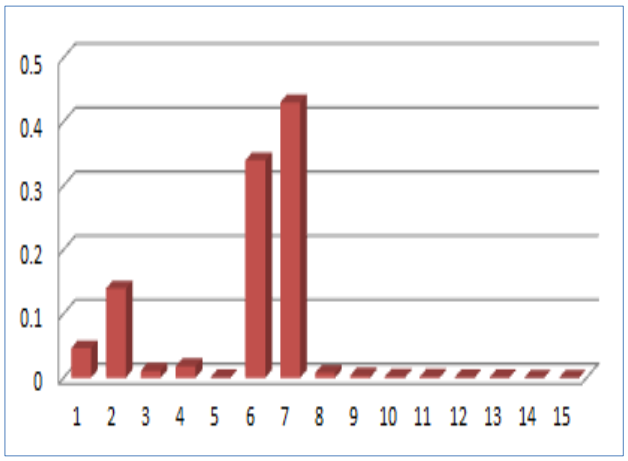

Fig. 8. Decision Coverage (No).

VI. Conclusion

The ultimate goal of this research is to find the relationship between the set of heat-stress expressed genes and their detected SNPs biological features in A. thaliana RNA-seq raw reads. Utilizing rough set-based rule induction resulted in set of descriptive rules which can draw the correlation between those two significant concepts; genes and SNPs. A promising analysis framework was presented to detect SNPs in RNA-seq raw reads then using annotations of those SNPs to figure out their biological features. Additionally, about (225) unique genes got from DRASTIC and TAIR10 databases were used to represent the highly expressed heat-stress genes for A. thaliana. However, (200) genes were selected to represent the lowly expressed heat-stress genes for the same plant. The top-ranked biological features of SNPs for both groups of genes with decision rules (Heat Expressed: Yes; No) were utilized to build a rule-based classifier using the Rosetta system.

The system stated set of (32) non-repeatable rules. Results showed acceptable outcomes and, evaluation had been applied to check the suitability of the generated rules using Rule Strength, Rule Certainty and Decision Coverage. In conclusion, relation between SNP calls and expressed genes in RNA-seq data can be a very useful by-product and increases the amount of knowledge for SNPs discovery and analysis in functional genomics research. With this important result in mind, this method can be verified using in vivo tests to improve the work results. Moreover, generating rules for more species of the same plant may improve complete and well-defined base for machine learning approach to researchers of all expertise levels.

References

[1] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), 1494.

[2] Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., & Pun, M. (2018). VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC bioinformatics, 19(1), 135.

[3] Zhao, Y., Wang, K., Wang, W. L., Yin, T. T., Dong, W. Q., & Xu, C. J. (2019). A high-throughput SNP discovery strategy for RNA-seq data. BMC genomics, 20(1), 160.

[4] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., & McKenna, A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 43(5), 491.

[5] Yu, Y., Wei, J., Zhang, X., Liu, J., Liu, C., Li, F., & Xiang, J. (2014). SNP discovery in the transcriptome of white Pacific shrimp Litopenaeus vannamei by next generation sequencing. PloS one, 9 (1), e87218.

[6] Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., & Kocher, J. P. A. (2016). Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Briefings in bioinformatics, 18(6), 973-983.

[7] Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., Corvin AP & Morris, D. W. (2013). Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PloS one, 8(3), e58815.

[8] Hvidsten, T. R., & Komorowski, J. (2007). Rough sets in bioinformatics. In Transactions on rough sets VII (pp. 225-243). Springer, Berlin, Heidelberg.

[9] Rissino, S., & Lambert-Torres, G. (2009). Rough set theory—fundamental concepts, principals, data extraction, and applications. In Data mining and knowledge discovery in real life applications. IntechOpen, (pp. 35-60).

[10] Dong, J., Zhong, N., & Ohsuga, S. (1999). Probabilistic rough induction: the GDT-RS methodology and algorithms. In International Symposium on Methodologies for Intelligent Systems (pp. 621-629). Springer, Berlin, Heidelberg.

[11] Komorowski, J. (2014). Learning rule-based models-the rough set approach. Amsterdam: Comprehensive Biomedical Physics.

[12] Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M.,Rody, D. & Roux, F. (2018). A genomic map of climate adaptation in Arabidopsis thaliana at a micro-geographic scale. Frontiers in plant science, 9, 967.

[13] Korani, W., Clevenger, J. P., Chu, Y., & Ozias-Akins, P. (2019). Machine Learning as an Effective Method for Identifying True Single Nucleotide Polymorphisms in Polyploid Plants. The Plant Genome, 12(1).

[14] Hassanien, A. E., Al-Shammari, E. T., & Ghali, N. I. (2013). Computational intelligence techniques in bioinformatics. Computational biology and chemistry, 47, (pp. 37-47).

[15] Zhang, S. W., Huang, D. S., & Wang, S. L. (2010). A method of tumor classification based on wavelet packet transforms and neighborhood rough set. Computers in biology and medicine, 40(4), (pp. 430-437).

[16] Maji, P., & Paul, S. (2011). Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning, 52(3), (pp. 408-426).

[17] Jain, K., & Kulkarni, S. (2020), Multi-reduct Rough Set Classifier for Computer-Aided Diagnosis in Medical Data. In Advancement of

Machine Intelligence in Interactive Medical Image Analysis. (pp. 167-183). Springer, Singapore.

[18] Bagyamathi, M., & Inbarani, D. H. H. (2017). Prediction of Protein Structural Classes using Rough Set based Feature Selection and Classification Framework. Journal of Recent Research In Engineering and Technology (JRRET), 4.

[19] Pawlak, Z. (1997). Vagueness - a Rough Set View, In: Lecture Notes in Computer Science- 1261, Mycielski, J; Rozenberg, G. & Salomaa, A. (editors), (pp. 106-117), Springer, ISBN 3-540- 63246-8, Secaucus-USA.

[20] Komorowski J., Øhrn A., Skowron A., (2002) The ROSETTA Rough Set Software System. In: W. Klo¨sgen, Zytkow J, eds. Handbook of Data Mining and Knowledge Discovery, Oxford University Press.

[21] Pawlak, Z. (2002). Rough sets and intelligent data analysis. Information sciences, 147(1-4), 1-12.

[22] Vashist, R., & Garg, M. L. (2011). Rule generation based on reduct and core: A rough set approach. Int. J. Comput. Appl, 29(9), 0975-8887.

[23] Enroth, S., Bornelöv, S., Wadelius, C., & Komorowski, J. (2012). Combinations of histone modifications mark exon inclusion levels. PloS one, 7(1), e29911.

[24] Zaki. H., Nassef M., Farouk A., & Badr A. (2019). A proposed RNA-seq analysis workflow to study heat-stress genes in Arabidopsis thaliana. Bioscience Research, 16(3), (pp. 2641-2654),