ISSN 2714-7533

An Information Entropy Based to Identify Dominant Species for No Biological Data Genes

Nurul Ain Nazirah [a], Shahreen Kasim [a,1,*], Dwiny Meidelfi [b]

[a]Faculty of Computer Science and Technology, Universiti Tun Hussein Onn, Parit Raja, 86400, MALAYSIA
[b]Department of Information Technology, Politeknik Negeri Padang, West Sumatera, Indonesia

[1] shahreen@uthm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received September 16, 2020
Revised October 30, 2020
Accepted November 14, 2020

Keywords
Genes
Dominant species
Clustering
Entropy

ABSTRACT

This report discusses about the dominant species of no biological data genes. This genes are belong to animal species. Gene is one of a process where the biological data encoded in the gene that instructed by the DNA to convert into a functional product such as protein. In gene classification, with the growth of using gene expression database, there are not enough tools to extract the gene expression from these databases. There exist over 23,000 to 50,000 genes for animal genome. So, this might contribute to data redundancy as problems can happen while handling a huge database. Therefore, to overcome the problem, many approaches to cluster and determine the dominant species have been proposed in the previous literature. For this project, in order to determine the dominant species, the information entropy based method is used. As conclusion, the purpose of this research is to identify the dominant species of no biological genes using entropy method proposed.

1. Introduction

The development of the idea of entropy of random variables and processes by Claude Shannon provided the beginnings of information theory and of the modern age of ergodic theory. Entropy is simply the average (expected) amount of the information from the event. A fundamental pattern in genetics, even in the most diverse communities of animals species comprise a few highly dominant genes accompanied by many more genes that are uncommon or rare which is, represented by only a few species. The often terms ‘dominant’ or ‘common’ are used interchangeably to describe these highly dominant species. Dominant species is the species that have high abundance relative to other species in a community, and have proportionate effects on environmental conditions, community diversity and/or ecosystem function.

With the evolution of new expression of portraying techniques such as SAGE (Serial Analysis of Gene Expression), the gene expression databases are increasing day by day. Gene expression is one of a process where the biological data encoded in the gene that instructed by the DNA to convert into a functional product such as protein [1]. For years, many types of technology, technique and measurement is used to collect all the biological and genetics information regarding gene expression that lies under different conditions during a biological process and experiments of different tissue

samples. One of the techniques is clustering. The clustering techniques has allowed a fast progress in this biological research [2] and it also has further the study of the issues such as differential gene expression [3].

The main objective for this research is to study the method approach that are related to the classification and clustering based for the dominant species of no biological data genes. Next is to calculate the probability outcome of no biological data genes. The third objective is to identify the dominant species of no biological data genes.

The rest of the paper is organized as follows: Section 2 presents the proposed method for the no biological data genes. Section 3 shows the results of the finding. Finally, Section 4 will concludes all the related works and highlight the directions for future research.

2. Materials

In this section, the materials and methodology presents to describes all the necessary information that is required to obtain the results of this research.

2.1 Materials

There are five types of animal species genes dataset that are used to perform this research analysis. Those five types are belong to Bos Taurus, Canis Lupus Familiaris, Sus Scrofa, Gallus Gallus and Homo Sapiens. For this research one method of calculation is used which is from information entropy based method. Description of the entropy based method are as follows:

a) Information Entropy Based Method

Entropy is a measure of the uncertainty of a random variable. For a discrete random variable with limited states. For the sake of simplicity, p(ij) is used to represent probability of the data genes. Entropy is simply the average or the expected amount of the information from the event. The formula for this entropy method is combined from information formula. Information formula is shows as below.

$$I(p) = -\log_b(p) \quad (1)$$

The p in the formula indicated the probability of the how much certain event that happened. Next, for log calculation which is b indicated the base which base 2 is mostly used in information theory. Next, the entropy formula is derived as below.

$$I = -\sum_{i=1}^{n}(N * pi) * \log_b(pi) \quad (2)$$

Based on the formula above, I indicated the total information from N occurrences. N indicated the number of occurrences while (N*pi) indicated the approximated number that the certain result will come out in N occurrence.

$$\text{Entropy} = -\sum_{i=1}^{n} pi \log_b(pi) \quad (3)$$

Based on the formula above, there are difference between the total Information from N occurrences and the Entropy equation which is only thing that changed in the place of N. The N is moved to the right, which means that I/N is Entropy. Therefore, Entropy is the average or the expected amount of information in a certain event. For this research, the entropy formula is shown in section 3.1.2.

3. Methodology

In this section, the activities throughout in this research are explained based on the framework provided in the section 3.1 where it describes each phases of the framework. In the section 3.2 the hardware and software requirements are described.

3.1 Research Framework

Research framework is divided into three phases. In each phases there are activities are under taken and carried out which help to complete this research. The function of each phases are explained in this section. Figure 1 below shows the overall flow of the research methodology.

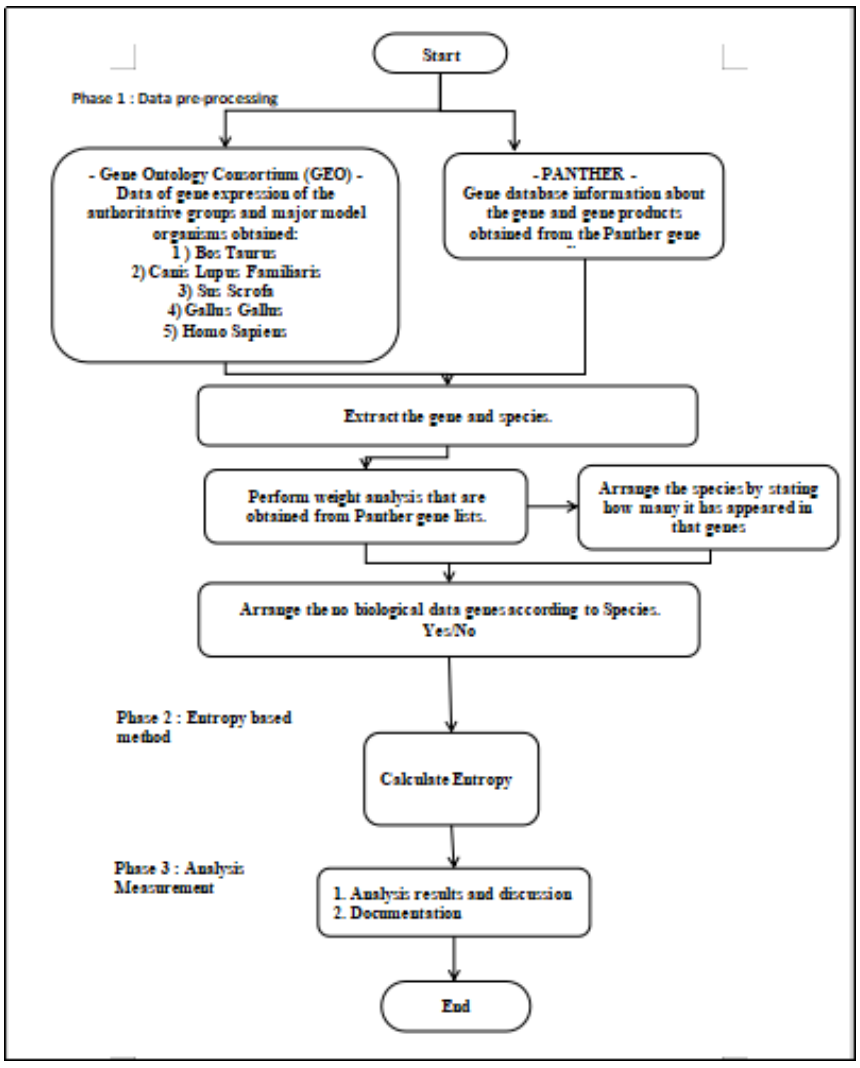


Figure 1. The flow of research methodology

3.1.1 Phase 1: Data Pre-processing

There are five datasets that are used in this research which are Bos Taurus, Canis Lupus Familiaris, Sus Scrofa, Gallus Gallus and Homo Sapiens. All data of gene expression of the authoritative groups and major model organisms are obtained from Gene Ontology Consortium (GEO) [4] and the gene database information about the gene and gene products obtained from the Panther gene lists. The other related information are collected from internet, journals and articles. Additional information that are related to this research are referred to previous research and all this information are gathered in details.

General information on the details of gene expression was reviewed after the data collection. Bos Taurus, Canis Lupus Familiaris, Sus Scrofa and Gallus Gallus genes informations is obtained from the GEO database. All these species contain three types of classes which are molecular function(F), cellular component(C) and biological process(P) [5]. For Bos Taurus, there are 44311 samples in this set of data. In this samples, there are 7 samples of no biological data genes. Next, for Canis Lupus Familiaris there are 121321 of raw samples data and 6 samples of no biological data genes. For Sus Scrofa, there are 129119 of raw samples data and 23 samples of no biological data genes. As for Gallus Gallus, there are 101399 of raw samples and 32 samples of no biological data genes.

	A	B	C	D	E	F	G	H	I	J	K
1	!gaf-version: 2.1										GENE
2	UniProtKB	A0A0A0MP	LGR4		GO:0004930	GO_REF:00(	IEA	UniProtKB-	F	G_PROTEIN_RECEP_	LGR4
3	UniProtKB	A0A0A0MP	LGR4		GO:0005886	GO_REF:00(	IEA	UniProtKB-	C	G_PROTEIN_RECEP_	LGR4
4	UniProtKB	A0A0A0MP	LGR4		GO:0007186	GO_REF:00(	IEA	UniProtKB-	P	G_PROTEIN_RECEP_	LGR4
5	UniProtKB	A0A0A0MP	LGR4		GO:0016021	GO_REF:00(	IEA	UniProtKB-	C	G_PROTEIN_RECEP_	LGR4
6	UniProtKB	A0A0A0MP	LGR4		GO:0016500	GO_REF:00(	IEA	InterPro:IPI	F	G_PROTEIN_RECEP_	LGR4
7	UniProtKB	A0A0A0MP	SERPINA3-1		GO:0005615	GO_REF:00(	IEA	InterPro:IPI	C	SERPIN domain-con	SERPINA3-1
8	UniProtKB	A0A140G0D	GALNTL5		GO:0016021	GO_REF:00(	IEA	UniProtKB-	C	N-acetylgalactosam	GALNTL5
9	UniProtKB	A0A140G0D	GALNTL5		GO:0016740	GO_REF:00(	IEA	UniProtKB-	F	N-acetylgalactosam	GALNTL5
10	UniProtKB	A0A140T82	IMPDH1		GO:0000166	GO_REF:00(	IEA	UniRule:UR	F	Inosine-5'-monophc	IMPDH1|IMPDH
11	UniProtKB	A0A140T82	IMPDH1		GO:0003677	GO_REF:00(	IEA	UniProtKB:I	F	Inosine-5'-monophc	IMPDH1|IMPDH
12	UniProtKB	A0A140T82	IMPDH1		GO:0003938	GO_REF:00(	IEA	EC:1.1.1.20	F	Inosine-5'-monophc	IMPDH1|IMPDH
13	UniProtKB	A0A140T82	IMPDH1		GO:0005634	GO_REF:00(	IEA	UniProtKB-	C	Inosine-5'-monophc	IMPDH1|IMPDH
14	UniProtKB	A0A140T82	IMPDH1		GO:0005829	GO_REF:00(	IEA	UniProtKB:I	C	Inosine-5'-monophc	IMPDH1|IMPDH
15	UniProtKB	A0A140T82	IMPDH1		GO:0006177	GO_REF:00(	IEA	UniRule:UR	P	Inosine-5'-monophc	IMPDH1|IMPDH
16	UniProtKB	A0A140T82	IMPDH1		GO:0046651	GO_REF:00(	IEA	UniProtKB:I	P	Inosine-5'-monophc	IMPDH1|IMPDH
17	UniProtKB	A0A140T82	IMPDH1		GO:0046872	GO_REF:00(	IEA	UniRule:UR	F	Inosine-5'-monophc	IMPDH1|IMPDH
18	UniProtKB	A0A140T82	IMPDH1		GO:0055114	GO_REF:00(	IEA	UniProtKB-	P	Inosine-5'-monophc	IMPDH1|IMPDH
19	UniProtKB	A0A140T83	ACHE		GO:0004104	GO_REF:00(	IEA	InterPro:IPI	F	Carboxylic ester hyc	ACHE
20	UniProtKB	A0A140T83	ACHE		GO:0005576	GO_REF:00(	IEA	UniProtKB-	C	Carboxylic ester hyc	ACHE
21	UniProtKB	A0A140T83	UVRAG		GO:0005789	GO_REF:00(	IEA	UniProtKB-	C	Acyltransferase	UVRAG|DGAT2
22	UniProtKB	A0A140T83	UVRAG		GO:0016021	GO_REF:00(	IEA	UniRule:UR	C	Acyltransferase	UVRAG|DGAT2
23	UniProtKB	A0A140T83	UVRAG		GO:0016747	GO_REF:00(	IEA	InterPro:IPI	F	Acyltransferase	UVRAG|DGAT2
24	UniProtKB	A0A140T84	BCAR3		GO:0005085	GO_REF:00(	IEA	UniProtKB-	F	Breast cancer anti-e	BCAR3

Figure 2. Raw samples of gene expression data for Bos Taurus

For example, Figure 2 above shows the raw samples of Bos Taurus datasets that are obtained from GEO [6]. There are 44311 genes in this datasets. This raw data will then scaled down or filtered to the no biological datasets as shown in Figure 3 below. Next, the filtered datasets will then uploaded to the Panther Classification System to obtained the Panther gene lists [7] as shown in Figure 4.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	!gaf-version: 2.1										GENE	PRODUCT NAME	ORGANISM	PANTHER FAMILY	TYPE	SOURCE
2	UniProtKB	A0JNP5	OTOS		GO:00036	GO_REF:0	ND		F	OTOS prot	OTOS	Otospiralin	Bos Taurus	Otospiralin pthr35073	protein	UniProtKB
3	UniProtKB	A0JNP5	OTOS		GO:00055	GO_REF:0	ND		C	OTOS prot	OTOS	Otospiralin	Bos Taurus	Otospiralin pthr35073	protein	UniProtKB
4	UniProtKB	A0JNP5	OTOS		GO:00081	GO_REF:0	ND		P	OTOS prot	OTOS	Otospiralin	Bos Taurus	Otospiralin pthr35073	protein	UniProtKB
5	UniProtKB	A5D7T7	MACROD2		GO:00036	GO_REF:0	ND		F	MACROD2	MACROD2	Uncharacterized prot	Bos Taurus	ganglioside induced differ	protein	UniProtKB
6	UniProtKB	A5D7T7	MACROD2		GO:00055	GO_REF:0	ND		C	MACROD2	MACROD2	Uncharacterized prot	Bos Taurus	ganglioside induced differ	protein	UniProtKB
7	UniProtKB	A5D7T7	MACROD2		GO:00081	GO_REF:0	ND		P	MACROD2	MACROD2	Uncharacterized prot	Bos Taurus	ganglioside induced differ	protein	UniProtKB
8	UniProtKB	A6QLD7	ADNP2		GO:00036	GO_REF:0	ND		F	C2H2-type	ADNP2	C2H2-type domain- c	Bos Taurus	neuroprotective peptide c	protein	UniProtKB
9	UniProtKB	A6QLD7	ADNP2		GO:00055	GO_REF:0	ND		C	C2H2-type	ADNP2	C2H2-type domain- c	Bos Taurus	neuroprotective peptide c	protein	UniProtKB
10	UniProtKB	E1BIY3	PKHD1L1		GO:00036	GO_REF:0	ND		F	PA14 dom	PKHD1L1	PKHD1 like 1	Bos Taurus	plexin pthr22625	protein	UniProtKB
11	UniProtKB	E1BIY3	PKHD1L1		GO:00055	GO_REF:0	ND		C	PA14 dom	PKHD1L1	PKHD1 like 1	Bos Taurus	plexin pthr22625	protein	UniProtKB
12	UniProtKB	E1BIY3	PKHD1L1		GO:00081	GO_REF:0	ND		P	PA14 dom	PKHD1L1	PKHD1 like 1	Bos Taurus	plexin pthr22625	protein	UniProtKB
13	UniProtKB	F2X2F1	IGF1		GO:00036	GO_REF:0	ND		F	Insulin-lik	IGF1|IGF-I	Insulin-like growth fa	Bos Taurus	Insulin-like growth factor	protein	UniProtKB
14	UniProtKB	Q0P5G9	SPRYD4		GO:00036	GO_REF:0	ND		F	B30.2/SPR	SPRYD4	B30.2/SPRY domain-c	Bos Taurus	e3 ubiquitin-protein ligas	protein	UniProtKB
15	UniProtKB	Q0P5G9	SPRYD4		GO:00081	GO_REF:0	ND		P	B30.2/SPR	SPRYD4	B30.2/SPRY domain-c	Bos Taurus	e3 ubiquitin-protein ligas	protein	UniProtKB
16	UniProtKB	Q0VCH2	IMMP1L		GO:00036	GO_REF:0	ND		F	Mitochon	IMMP1L	Mitochondrial inner r	Bos Taurus	protease family s26 mitoc	protein	UniProtKB
17	UniProtKB	Q0VCH2	IMMP1L		GO:00055	GO_REF:0	ND		C	Mitochon	IMMP1L	Mitochondrial inner r	Bos Taurus	protease family s26 mitoc	protein	UniProtKB
18	UniProtKB	Q0VCH2	IMMP1L		GO:00081	GO_REF:0	ND		P	Mitochon	IMMP1L	Mitochondrial inner r	Bos Taurus	protease family s26 mitoc	protein	UniProtKB

Figure 3. No biological datasets of Bos Taurus

	A	B	C	D	E	F
1						
2	HORSE|Ensembl=EI	PKHD1L1	PKHD1 lik	FIBROCYSTIN-L (PTHI		Equus caballus
3	CANLF|Ensembl=EI	SPRYD4	B30.2/SPR	SPRY DOM	ubiquitin-	Canis lupus familiaris
4	FELCA|Ensembl=EN	OTOS	Otospirali	OTOSPIRALIN (PTHR:		Felis catus
5	RAT|RGD=708465|U	OTOS	Otospirali	OTOSPIRALIN (PTHR:		Rattus norvegicus
6	ANOCA|Ensembl=E	ADNP2	C2H2-type	ACTIVITY-DEPENDEN		Anolis carolinensis
7	HUMAN|HGNC=203	PKHD1L1	Fibrocysti	FIBROCYSTIN-L (PTHI		Homo sapiens
8	GORGO|Ensembl=E	IMMP1L	Mitochon	MITOCHO	protease(	Gorilla gorilla gorilla
9	PANTR|Ensembl=EI	OTOS	Otospirali	OTOSPIRALIN (PTHR:		Pan troglodytes
10	BOVIN|Ensembl=EI	SPRYD4	B30.2/SPR	SPRY DOM	ubiquitin-	Bos taurus
11	ANOCA|Ensembl=E	OTOS	Otospirali	OTOSPIRALIN (PTHR:		Anolis carolinensis
12	ORYLA|Ensembl=EN	MACROD2	Macro dor	ADP-RIBOSE GLYCOH		Oryzias latipes
13	RAT|RGD=1310406|	PKHD1L1	PKHD1-lik	FIBROCYSTIN-L (PTHI		Rattus norvegicus
14	HUMAN|HGNC=161	MACROD2	ADP-ribos	ADP-RIBOSE GLYCOH		Homo sapiens
15	MACMU|Ensembl=I	IMMP1L	Mitochon	MITOCHO	protease(	Macaca mulatta
16	CANLF|Ensembl=EI	OTOS	Otospirali	OTOSPIRALIN (PTHR:		Canis lupus familiaris
17	PANTR|Ensembl=EI	SPRYD4	B30.2/SPRY domain-containing			Pan troglodytes
18	CANLF|Ensembl=EI	IMMP1L	Mitochon	MITOCHO	protease(	Canis lupus familiaris
19	PIG|Ensembl=ENSS	ADNP2	C2H2-type	ACTIVITY-DEPENDEN		Sus scrofa
20	HUMAN|HGNC=274	SPRYD4	SPRY domain-containing protei			Homo sapiens
21	PANTR|Ensembl=EI	MACROD2	Mono-AD	ADP-RIBOSE GLYCOH		Pan troglodytes
22	DANRE|ZFIN=ZDB-(	MACROD2	Macro dor	ADP-RIBOSE GLYCOH		Danio rerio
23	ORYLA|Ensembl=EN	IMMP1L	Mitochon	MITOCHO	protease(	Oryzias latipes
24	RAT|RGD=1306649|	SPRYD4	SPRY dom	SPRY DOM	ubiquitin-	Rattus norvegicus

BOS TAURUS ND | BOS TAURUS 7 pantherGeneList | TASK 1 | TASK 2

Figure 4. Datasets that are obtained from Panther gene lists.

Next, after datasets is obtained as shown in Figure 4, this datasets is extracted to carry out two tasks in order to determine how many no biological data gene for each species and how many species for each genes. The first task will be shown as in Figure 5 below and second task is shown as in Figure 3.6 below as well.

	A	B
1	GENE	
2	ADNP2	15
3	IMMP1L	18
4	MACROD2	13
5	OTOS	16
6	PKHD1L1	13
7	SPRYD4	17
8		

Figure 5. Task one to identify how many no biological data gene for each species.

As shown in Figure 5, this activity is to identify how many no biological data genes that appeared in each species. For example, ADNP2 is one of Bos Taurus's genes that appeared in other 15 species.

	A	B
1	**SPECIES**	
2	Anolis carolinensis	4
3	Bos taurus	6
4	Canis lupus familiaris	6
5	Danio rerio	4
6	Equus caballus	6
7	Felis catus	6
8	Gallus gallus	4
9	Gorilla gorilla gorilla	6
10	Homo sapiens	6
11	lepisosteus oculatus	1
12	Macaca mulatta	6
13	Monodelphis domestica	5
14	Mus musculus	6
15	Ornithorhynchus anatinus	4
16	Oryzias latipes	4
17	Pan troglodytes	6
18	Rattus norvegicus	5
19	Sus scrofa	5
20	Xenopus tropicalis	2

Figure 6. Task two to identify how many species for each genes

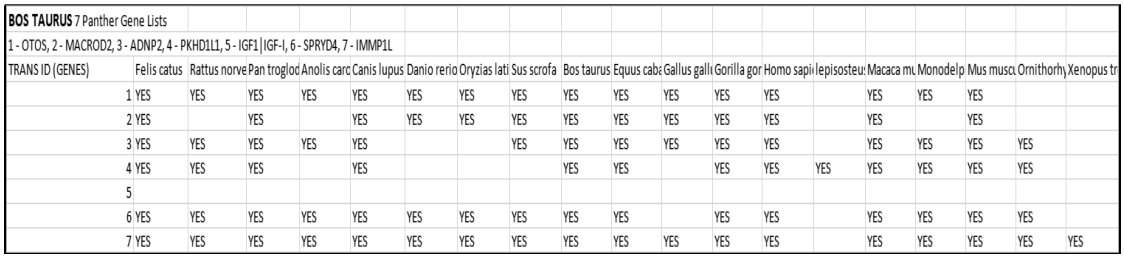
As shown in Figure 6, this activity is done to identify how many species that appeared for each genes. For example, species Anolis carolinensis has appeared in other 4 types of genes. Next, a weight analysis for no biological data genes of Bos Taurus is carried out. This activity is done in order to identify to classify which species have appeared in each genes. Figure 3.7 below will show the analysis weight for Bos Taurus.

K	L	M	N	O	P	Q	R	S	T	U	V	W
GENE	**PRODUCT NAME**	**ORGANISM**	**PANTHER FAMILY**	**TYPE**	**SOURCE**	**SYNONYM**						**WEIGHT ANALYSIS**
OTOS	Otospiralin	Bos Taurus	Otospiralin pthr350	protein	UniProtKB	UniProtKB: M	taxor	2E+07	AgBase	UniProt		Felis catus -1, Rattus norvegicus -1, Pan troglodytes -1, Anolis carolinensis -1, Canis
OTOS	Otospiralin	Bos Taurus	Otospiralin pthr350	protein	UniProtKB	UniProtKB: M	taxor	2E+07	AgBase	UniProt		lupus familiaris -1, Danio rerio -1, Oryzias latipes -1, Sus scrofa -1, Homo sapiens -1,
OTOS	Otospiralin	Bos Taurus	Otospiralin pthr350	protein	UniProtKB	UniProtKB: M	taxor	2E+07	AgBase	UniProt		Gallus gallus -1, Monodelphis domestica -1, Gorilla gorilla gorilla -1, Mus musculus
MACROD2	Uncharacterized prot	Bos Taurus	ganglioside induced	protein	UniProtKB	UniProtKB: G	taxor	2E+07	UniPro	UniProt		Oryzias latipes -1, Homo sapiens -1, Pan troglodytes -1, Danio rerio -1, Gallus gallus -
MACROD2	Uncharacterized prot	Bos Taurus	ganglioside induced	protein	UniProtKB	UniProtKB: G	taxor	2E+07	UniPro	UniProt		1, Sus scrofa -1, Canis lupus familiaris -1, Bos taurus -1, Macaca mulatta -1, Mus
MACROD2	Uncharacterized prot	Bos Taurus	ganglioside induced	protein	UniProtKB	UniProtKB: G	taxor	2E+07	UniPro	UniProt		musculus -1, Felis catus -1, Gorilla gorilla gorilla -1, Equus caballus -1
ADNP2	C2H2-type domain- c	Bos Taurus	neuroprotective pep	protein	UniProtKB	-	taxor	2E+07	UniPro	UniProt		Anolis carolinensis -1, Sus scrofa -1, Gorilla gorilla gorilla -1, Monodelphis
ADNP2	C2H2-type domain- c	Bos Taurus	neuroprotective pep	protein	UniProtKB		taxor	2E+07	UniPro	UniProt		domestica -1, Equus caballus -1, Pan troglodytes -1, Ornithorhynchus anatinus -1,
PKHD1L1	PKHD1 like 1	Bos Taurus	plexin pthr22625	protein	UniProtKB	UniProtKB:AC	taxor	2E+07	UniPro	UniProt		Equus caballus -1, Homo sapiens -1, Rattus norvegicus -1, Canis lupus familiaris -1,
PKHD1L1	PKHD1 like 1	Bos Taurus	plexin pthr22625	protein	UniProtKB	UniProtKB:AC	taxor	2E+07	UniPro	UniProt		lepisosteus oculatus -1, Gorilla gorilla gorilla -1, Pan troglodytes -1, Macaca mulatta
PKHD1L1	PKHD1 like 1	Bos Taurus	plexin pthr22625	protein	UniProtKB	UniProtKB:AC	taxor	2E+07	UniPro	UniProt		-1, Mus musculus -1, Bos taurus -1, Monodelphis domestica -1, Felis catus -1,
IGF1|IGF-I	Insulin-like growth fa	Bos Taurus	Insulin-like growth f	protein	UniProtKB	-	taxor	2E+07	AgBase	UniProt		-
SPRYD4	B30.2/SPRY domain-c	Bos Taurus	e3 ubiquitin-protein	protein	UniProtKB	-	taxor	2E+07	UniPro	UniProt		Canis lupus familiaris -1, Bos taurus -1, Pan troglodytes -1, Homo sapiens -1, Rattus
SPRYD4	B30.2/SPRY domain-c	Bos Taurus	e3 ubiquitin-protein	protein	UniProtKB	-	taxor	2E+07	UniPro	UniProt		norvegicus -1, Danio rerio -1, Sus scrofa -1, Anolis carolinensis -1, Macaca mulatta -
IMMP1L	Mitochondrial inner	Bos Taurus	protease family s26	protein	UniProtKB	-	taxor	2E+07	AgBase	UniProt		Gorilla gorilla gorilla -1, Macaca mulatta -1, Canis lupus familiaris -1, Oryzias
IMMP1L	Mitochondrial inner	Bos Taurus	protease family s26	protein	UniProtKB	-	taxor	2E+07	AgBase	UniProt		latipes -1, Equus caballus -1, Anolis carolinensis -1, Gallus gallus -1, Danio rerio -1,
IMMP1L	Mitochondrial inner	Bos Taurus	protease family s26	protein	UniProtKB	-	taxor	2E+07	AgBase	UniProt		Mus musculus -1, Monodelphis domestica -1, Bos taurus -1, Sus scrofa -1, Pan

Figure 7. Analysis weight for Bos Taurus datasets

A shown in the figure 7 above, an analysis weight is carried out. This activity is by arranging the species by stating how many it has appeared in that genes. For example, in gene OTOS, the species have appeared are Felis catus -1, Rattus norvegicus -1, Pan troglodytes -1, Anolis carolinensis -1, Canis lupus familiaris -1, Danio rerio -1, Oryzias latipes -1, Sus scrofa -1, Homo sapiens -1, Gallus gallus -1, Monodelphis domestica -1, Gorilla gorilla gorilla -1, Mus musculus -1, Macaca mulatta -1, Equus caballus -1 and Bos taurus -1. Number 1 indicate that these species appeared one time in this gene.

Next, after analysis weight is done, new datasets is carried out which is to arrange the no biological data genes according to the species. Figure 3.8 below shows the arrangement of the datasets.



BOS TAURUS 7 Panther Gene Lists

1 - OTOS, 2 - MACROD2, 3 - ADNP2, 4 - PKHD1L1, 5 - IGF1|IGF-I, 6 - SPRYD4, 7 - IMMP1L

TRANS ID (GENES)	Felis catus	Rattus norve	Pan troglod	Anolis carc	Canis lupus	Danio rerio	Oryzias lati	Sus scrofa	Bos taurus	Equus cabe	Gallus galli	Gorilla gor	Homo sapi	lepisosteu	Macaca mu	Monodelp	Mus muscu	Ornithorhy	Xenopus tr
1	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES	YES		
2	YES		YES		YES	YES	YES	YES	YES	YES	YES	YES	YES		YES		YES		
3	YES	YES	YES	YES	YES			YES	YES	YES	YES	YES	YES		YES	YES	YES	YES	
4	YES	YES	YES		YES				YES	YES		YES	YES	YES	YES	YES	YES	YES	
5																			
6	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES		YES	YES	YES	YES	
7	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES	YES	YES	YES

Figure 8. Arrangement of no biological data genes according to species.

As shown in Figure 8, the no biological data genes is arranged according to which species is appeared. In the figure, there are many YES words is stated in the columns. YES is stated to indicated that the gene has appeared in the species. For example, gene 1 is OTOS. OTOS has appeared in other 16 species which are Felis catus, Rattus norvegicus, Pan troglodytes, Anolis carolinensis, Canis lupus familiaris, Danio rerio, Oryzias latipes, Sus scrofa, Homo sapiens, Gallus gallus, Monodelphis domestica, Gorilla gorilla gorilla, Mus musculus, Macaca mulatta, Equus caballus, and Bos taurus.

3.1.2 Entropy based method.

In order to calculate the probability for the dominant species of no biological data genes, entropy based approach is used. Below shows the formula that is used in order to obtained the entropy value for each genes and species.

$$H(vi) = -\sum_{j=1}^{di} Pij \log_2 Pij \quad (4)$$

Based on the equation 4 above, the H(vi) indicated the entropy value for the species and genes. The Pij indicated the probability of each species and genes.

3.1.3 Phase 3 : Analysis Measurement

The analysis measurement of this research is to show results of the calculation. By this the total value of entropy is obtained and measured.

For this study, documentation of paper works the soft materials, code and formula that used is in the form of paper work. Documentation is meant to provide readers with a clear understanding of the overall flow research analysis.

3.2 Hardware and Software Requirements.

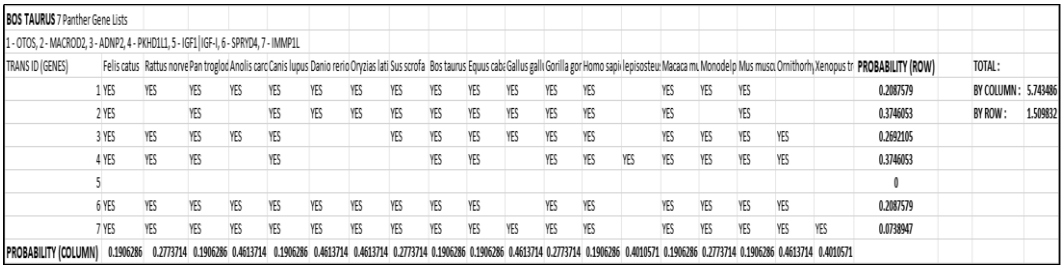
Some basic requirements are required for this research to be conducted successfully. A computer laptop with relatively high processing power and storage is used for hardware.

This study was carried out using Windows 10. Next, WPS Office Word is used to do the proposal and thesis writing while WPS Office Spreadsheet is used to sort the no biological data genes.

4. Results and Discussions

This section shows the results for the no biological data genes of four species. Figure 9, 10, 11, 12 and 13 are the results of the entropy calculation.

4.1 Results of Bos Taurus

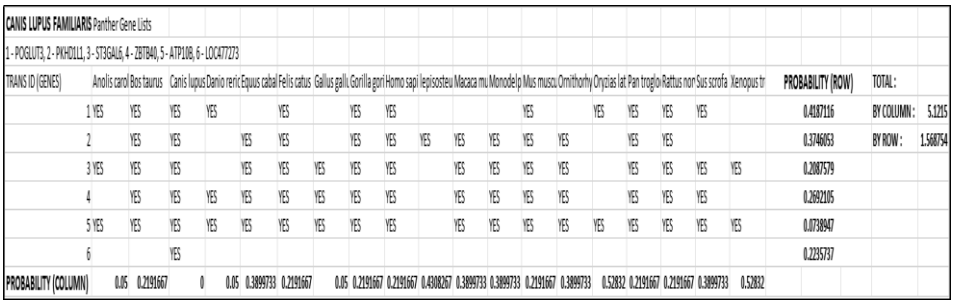


BOS TAURUS 7 Panther Gene Lists																									
1 - OTOS, 2 - MACROD2, 3 - ADNP2, 4 - PKHD1L1, 5 - IGFL1 IGF-I, 6 - SPRYD4, 7 - IMMP1L																									
TRANS ID (GENES)	Felis catus	Rattus norve	Pan troglod	Anolis carc	Canis lupus	Danio rerio	Oryzias lati	Sus scrofa	Bos taurus	Equus cabi	Gallus galli	Gorilla gori	Homo sapi	Ictalurus pu	Macaca mu	Monodelp	Mus muscu	Ornithorhy	Xenopus tr	PROBABILITY (ROW)	TOTAL:	
1 YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES	YES			0.2087579	BY COLUMN :	5.743486
2 YES		YES		YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES		YES			0.3746053	BY ROW :	1.509832
3 YES	YES	YES	YES	YES			YES	YES	YES	YES	YES	YES	YES		YES	YES	YES	YES		0.2692105		
4 YES	YES	YES		YES				YES	YES		YES	YES	YES	YES	YES	YES	YES	YES		0.3746053		
5																				0		
6 YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES		YES	YES	YES	YES			0.2087579		
7 YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES	YES	YES	YES		0.0738947		
PROBABILITY (COLUMN)	0.1906286	0.2773714	0.1906286	0.4613714	0.1906286	0.4613714	0.4613714	0.2773714	0.1906286	0.1906286	0.4613714	0.2773714	0.1906286	0.4010571	0.1906286	0.2773714	0.1906286	0.4613714	0.4010571			

Figure 9. Results of Bos Taurus using entropy based formula

As shown in Figure 9, this is the results that are obtained from the calculation by using entropy based formula. The probability outcome from this table is the entropy value. This entropy value is calculated for each genes and species. For example for trans id (genes) number one that is belong to OTOS, the entropy value is 0.2087579 while for the first species which is Felis catus, the entropy value is 0.1906286. From this calculation for each genes and species, the total of entropy value are obtained. Total entropy value for genes is 1.5098316 while total entropy value for species is 5.743486.

4.2 Result of Canis Lupus Familiaris



CANIS LUPUS FAMILIARIS Panther Gene Lists																									
1 - POGLUT3, 2 - PKHD1L1, 3 - ST3GAL6, 4 - ZBTB40, 5 - ATP10B, 6 - LOC477273																									
TRANS ID (GENES)	Anolis carol	Bos taurus	Canis lupus	Danio rerio	Equus caball	Felis catus	Gallus galli	Gorilla gori	Homo sapi	Ictalurus pu	Macaca mu	Monodelp	Mus muscu	Ornithorhy	Oryzias lat	Pan troglo	Rattus norv	Sus scrofa	Xenopus tr	PROBABILITY (ROW)	TOTAL:	
1 YES	YES	YES	YES		YES		YES	YES				YES		YES	YES	YES	YES		0.4187116	BY COLUMN :	5.1215
2	YES	YES		YES	YES		YES	YES	YES	YES	YES	YES	YES		YES	YES			0.3746053	BY ROW :	1.560754
3 YES	YES	YES		YES	YES	YES	YES	YES		YES	YES	YES	YES		YES	YES	YES	YES		0.2087579		
4	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES	YES	YES		YES	YES	YES			0.2692105		
5 YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES	YES	YES	YES	YES	YES	YES	YES		0.0738947		
6		YES																		0.2235737		
PROBABILITY (COLUMN)	0.05	0.2191667	0	0.05	0.3899733	0.2191667	0.05	0.2191667	0.2191667	0.4908267	0.3899733	0.3899733	0.2191667	0.3899733	0.52832	0.2191667	0.2191667	0.3899733	0.52832			

Figure 10. Results of Bos Taurus using entropy based formula

As shown in Figure 10, this is the results that are obtained from the calculation by using entropy based formula. The probability outcome from this table is the entropy value. This entropy value is calculated for each genes and species. For example for trans id (genes) number one that is belong to POGLUT, the entropy value is 0.4187116 while for the first species which is Anolis carolinensis, the entropy value is 0.05. From this calculation for each genes and species, the total of entropy value are obtained. Total entropy value for genes is 1.5687537 while total entropy value for species is 5.1215001.

4.3 Results of Sus Scrofa

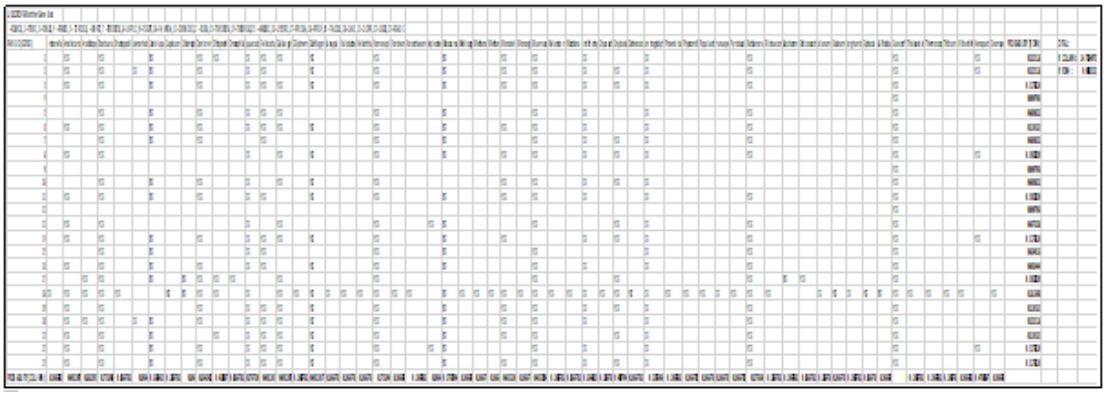


Figure 11. Results of Sus Scrofa using entropy based formula

As shown in Figure 11, this is the results that are obtained from the calculation by using entropy based formula. The probability outcome from this table is the entropy value. This entropy value is calculated for each genes and species. For example for trans id (genes) number one that is belong to ROMO1, the entropy value is 0.5225186 while for the first species which is Amborella trichopoda, the entropy value is 0.1966783. From this calculation for each genes and species, the total of entropy value are obtained. Total entropy value for genes is 9.9922522 while total entropy value for species is 14.7814972.

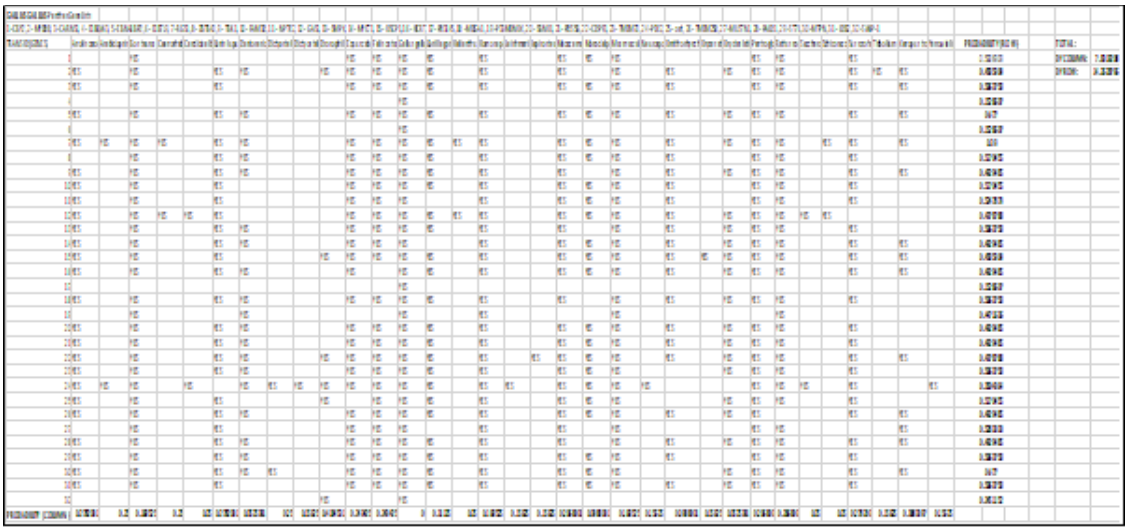


Figure 12. Results of Gallus Gallus using entropy based formula

As shown in Figure 12, this is the results that are obtained from the calculation by using entropy based formula. The probability outcome from this table is the entropy value. This entropy value is calculated for each genes and species. For example for trans id (genes) number one that is belong to CBY2, the entropy value is 0.5306909 while for the first species which is Anolis carolinensis, the entropy value is 00.2782031. From this calculation for each genes and species, the total of entropy value are obtained. Total entropy value for genes is 14.2620796 while total entropy value for species is 7.8319208.

4.4 Results of Gallus Gallus

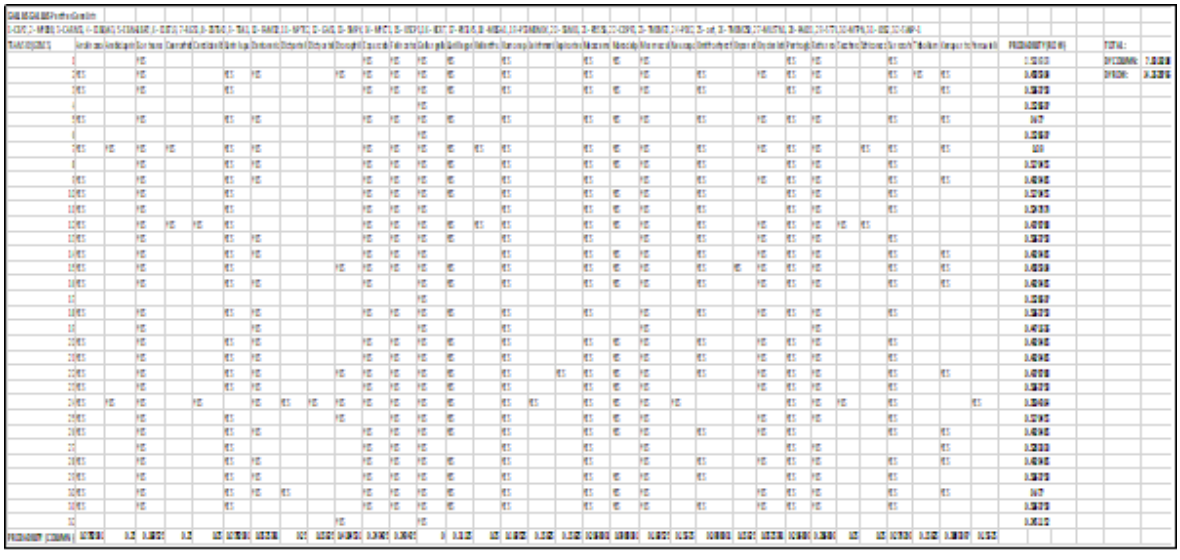


Figure 13. Results of Gallus Gallus using entropy based formula

As shown in Figure 13, this is the results that are obtained from the calculation by using entropy based formula. The probability outcome from this table is the entropy value. This entropy value is calculated for each genes and species. For example for trans id (genes) number one that is belong to CBY2, the entropy value is 0.5306909 while for the first species which is Anolis carolinensis, the entropy value is 00.2782031. From this calculation for each genes and species, the total of entropy value are obtained. Total entropy value for genes is 14.2620796 while total entropy value for species is 7.8319208.

4.5 Results of Homo Sapiens

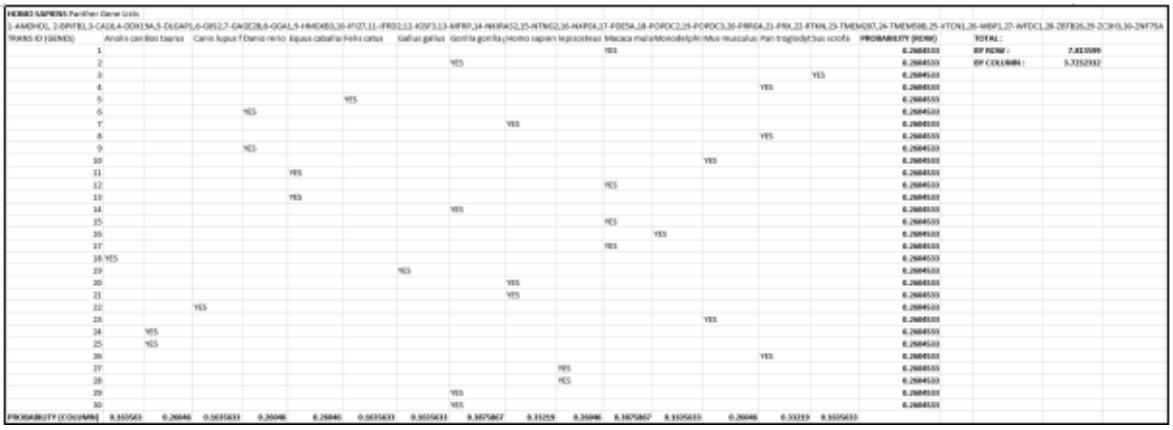


Figure 14. Results of Homo Sapiens using entropy based formula

As shown in Figure 14, this is the results that are obtained from the calculation by using entropy based formula. The probability outcome from this table is the entropy value. This entropy value is calculated for each genes and species. For example for trans id (genes) number one that is belong to AMDHD1, the entropy value is 0.2604533 while for the first species which is Anolis carolinensis, the entropy value is 0.1635633. From this calculation for each genes and species, the total of entropy value are obtained. Total entropy value for genes is 7.813599 while total entropy value for species is 3.7232332.

5. Conclusions

To prevent common errors in statistical modeling, many methods have been introduced, and entropy is one of the most widely used concept in medical and genetic sciences. Entropy was introduced by Nicholas Georgescu-Roegen in 1971 and later developed by scientists based on the principles established by Shannon [8]. Shannon had a major role in introducing entropy information, which has been widely used in high-dimensional studies [9]. One of the advantages of entropy is that calculation of values is based on theoretical forms, not the empirical and personal concepts. These values give small or large weights, proportional to the small or large actual values [10].

As conclusion, the results information entropy of dominant species for no biological data genes are discussed. This results is to determine the entropy value for each genes and species. The objective of this research study is to determine the entropy value of each genes and species. In order to know the dominant species of no biological data genes, this entropy values is referred. The entropy of a message is defined as the expected amount of information to be transmitted about the random variable defined in the previous chapter. This entropy method is concerned with data compression and transmission and builds upon probability and supports machine learning. The information provides a way to quantify the amount data that is measured in bits. Finally, entropy also provides a

measure of the average amount of information needed to represent an event drawn from a probability distribution for a random variable.

Acknowledgment

I would also like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support and encouragement throughout the process of conducting this study.

References

[1] Yourgenome (2016). What is gene expression. Retrieved on January 1, 2016, from https://www.yourgenome.org/facts/what-is-gene-expression.

[2] Giancarlo, R., Scaturro, D., & Utro, F. (2016). Valworkbench: An open source java library for cluster validation, with applications to microarray data analysis. Computer methods and programs in biomedicine, 118(2), 207-217.

[3] Vaes, E., Khan, M., & Mombaerts, P. (2016). Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes. BMC bioinformatics, 15(1), 39.

[4] Gene Ontology Consortium (2020). Annotations. Retrieved on October 9, 2020, from http://current.geneontology.org/products/pages/downloads.html.

[5] Gene Ontology Consortium (2020). Annotations. Retrieved on October 9, 2020, from http://current.geneontology.org/products/pages/downloads.html.

[6] Gene Ontology Consortium (2020). Annotations. Retrieved on October 9, 2020, from http://current.geneontology.org/products/pages/downloads.html.

[7] PANTHER Classification System (2021), Panther Gene Lists. Retrieved on December 18, 2020 from http://pantherdb.org/.

[8] Soltanian AR, Rabiei N, Bahreini F. Feature Selection in Microarray Data Using Entropy Information. In: Husi H, editor. Computational Biology [Internet]. Brisbane (AU): Codon Publications; 2019 Nov 21. Chapter 10. Available from: https://www.ncbi.nlm.nih.gov/books/NBK550347/ doi: 10.15586/computationalbiology.2019.ch10.

[9] Soltanian AR, Rabiei N, Bahreini F. Feature Selection in Microarray Data Using Entropy Information. In: Husi H, editor. Computational Biology [Internet]. Brisbane (AU): Codon Publications; 2019 Nov 21. Chapter 10. Available from: https://www.ncbi.nlm.nih.gov/books/NBK550347/ doi: 10.15586/computationalbiology.2019.ch10.

[10] Soltanian AR, Rabiei N, Bahreini F. Feature Selection in Microarray Data Using Entropy Information. In: Husi H, editor. Computational Biology [Internet]. Brisbane (AU): Codon Publications; 2019 Nov 21. Chapter 10. Available from: https://www.ncbi.nlm.nih.gov/books/NBK550347/ doi: 10.15586/computationalbiology.2019.ch10.