Pathway Clusters of Aging Genes using Data Mining Techniques

Vidya A
Department of Computer Science and Engineering,
University Visvesvaraya College of Engineering,
Bangalore, India
Research Scholar, Jawaharlal Nehru Technological University,
Hyderabad
vidyaananth16@gmail.com

Kalaivani M and Venugopal K R
Department of Computer Science
and Engineering, University
Visvesvaraya College of Engineering,
Bangalore, India

L M Patnaik
Honorary Professor
Indian Institute of Science
Bangalore, India

***Abstract*—Exploring and identifying novel aging genes has been the current area of interest in Gerontology. A variety of techniques have been proposed to identify the genes that affect the centenarians and the focus is on the study of genes of interest affecting older population. However the study of aging related pathways using computational methods has not been discussed explicitly so far. In this paper, an attempt is made to cluster the aging genes into different biological pathways using data mining techniques. Text mining is used to identify the most relevant keywords from different pathway databases, which is used as one of the feature describing a gene. K-means clustering is done on the aging pathway dataset. The clusters formed are in good agreement with the background knowledge about the aging genes and their pathways. The quality of the K-means clustering is quite promising as it well separates the different aging genes into their respective pathways.**

***Keywords*—Aging Genes, Clusters, K-means,Text mining, Pathway.**

I. Introduction

Aging, a system-level phenomenon is regarded as extremely complex mechanism, which is quite prevalent among the older population and has gained huge attention in medical research. Aging is defined as the accumulation of changes in a person over time that results due to the decline in the proliferative capacity of aging cells. In humans, the functioning of senescence is characterized by the declining ability to stress response, imbalance in homeostasis and prone to disease factor that gradually ends in death. As genes that have an effect on aging are identified both in-vitro and in-vivo, aging is increasingly being regarded in a similar fashion to other genetically influenced conditions, potentially treatable.

The study of aging genes trace back to the 19th century with the invention of daf2 gene responsible for the prolong of aging in C.elegans [1]. Protein coding for daf2 is similar to insulin receptors in humans. These proteins suspend their own development in a phase which is termed as Dauer formation. Daf2 is closely associated with calorie restriction, reduced food consumption results in the mutation of gene that encodes $PI3K$ gene, which results in the formation of daf2 the basis for aging. It is identified that these receptors play vital role in aging, stress resistance, metabolism and development. The major breakthrough in the aging research is the mimic of mammalian insulin pathway to daf2 signaling pathway. The functional genes controlled by daf2 are stress resistance collagens and metabolism, significantly turning on the functionality of daf2.

Specifically, Caloric Restriction(CR) [2] has been shown to increase lifespan in mice. CR works on many other species beyond mice and appears to increase lifespan in primates according to a study done on Rhesus monkeys at the National Institute of Health (US), although the increase in lifespan is only notable if the caloric restriction is started early in life. The molecular level of age is counted not as time but as the number of cell doublings progression that results in CR. This is supported by the fact that mTOR significantly contributes to the CR. mTOR, a protein that inhibits autophagy, which has been linked to aging through the insulin signaling pathway. It has been found, in various model species that CR leads to longer lifespans. When organisms restrict their diet their mTORs activity is reduced which allows for more autophagy, or cell self-eating. Autophagy is a cells way to clean house and recycle old or damaged cell parts, and keep the cells and the body running efficiently [3].

Overall it has been found that Insulin Signaling and Metabolism pathway contributes to the theory of aging. Pathways are largely the networks of metabolism which are referred to all chemical reactions that takes place in homosapiens, including some of the vital mechanisms such as digestion and the transport of solutes into and out of the cells, in which case the set of reactions within the cells sums up the total process in the body. It is important to understand how different pathways coordinate to perform aging related functions. However there are no published proposals describing how to build aging related pathway clusters that are functionally related so far. Different groups have recently shown that the usage of prior biological knowledge significantly improves the clustering results in terms of accuracy as well as reproducibility and interpretability of aging pathway lists. Several groups have proposed to adapt clustering methods in such a way that the algorithms can benefit from using prior biological knowledge [4].

Common sources of such biological knowledge are databases that contain pathway information, Protein-Protein Interaction (PPI) or Gene Ontology (GO). It is a well-known fact that, genes do not work in isolation. Each gene is a part of

978-1-4799-6758-2/14/$31.00 ©2014 IEEE

overall biological pathways. Therefore, it is essential to know how different genes or pathways coordinate their activities. There are several data mining techniques available to group the data objects into clusters. Clustering is a technique for finding similar groups in data, called clusters. It groups the data objects that are similar to each other in one cluster and different from each other into different clusters. There exists a number of clustering algorithms like partitioning, hierarchical, density based and so on. The quality of a clustering result depends on the algorithm, the distance function and the application [5].

Motivation

Clustering analysis has become powerful tool for analyzing biological data. A huge number of methods have been developed for these problems. Discovering patterns hidden in gene data offers a tremendous potential for advanced investigation in computational biology. Because of the large number of genes and the complexity of biological systems, clustering is necessary exploratory technique for further analysis.

Contribution

Although several promising pathway databases exist that make machine learning based methods for aging based pathway prediction a very difficult task. In this paper, we propose a method called Pathway Clusters of Aging Genes (PCAG) to apply and evaluate text mining approach for aging pathway prediction through K-means clustering.

Organization

The organization of this paper is as follows. Literature Survey is discussed in Section II. Proposed method is defined in Section III and Results and Discussions were described in Section IV. Conclusions are presented in Section V.

II. Literature Survey

Barzilai et al., [6] have identified major metabolic pathways that regulate mammalian longevity. Amongst them CR represents the most robust intervention to extend both mean and maximum life span in humans due to the magnitude of pathways affected by CR, including reduced cytokine levels, adiposity, IIS signalling, thyroid hormone levels, and increased adiponectin. In response to these changes, numerous downstream cellular pathways are engaged, including $SIRT1$ activation, IIS/phosphatidylinositol 3-kinase (PI3K)/Akt signalling, AMPK/mTOR signalling, and extracellular signal-regulated kinase signalling. The collective response of these pathways to CR is believed to promote cellular fitness and ultimately longevity via activation of autophagy, stress defence mechanisms, and survival pathways while attenuating proinflammatory mediators and cellular growth.

Pang et al., [4] have proposed the pathway clusters construction from the pathway-based classification models. It identifies clusters of pathways sharing similar function. It helps in understanding of molecular mechanisms affecting gene of interest and to investigate the flow of informative genes within pathways and relation between the pathways within a cluster. Classification of pathway clusters are obtained from gene expression data collected from micro array studies. Pathway clusters are built from various databases using random forest classification and not targeted on specific genes or pathway. The results narrow down only to understanding the molecular mechanisms from gene expression data. The attributes are focused only on gene expression from microarray studies.

Ibrahim et al., [7] have discussed the gene selection method incorporating the prior biological knowledge of genetic pathways to find groups of strongly-correlated genes that accurately discriminate complex as well as simple disease traits. A specified number of differentially-expressed genes from relevant pathways are used for disease classification. However this feature selection based algorithm method outperformes other methods in terms of disease classification accuracy and used in diagnostic tests. These results were obtained by collecting the data samples from the diseased patients and verifying the classification of genes accurately. However the disease sample were limited to three major diseases and classifying these into pathways.

Nirmala et al., [8] discussed about the consistency in classification of cancers based on gene expression data over clinical markers data. Biological pathway-based feature selection integrates signaling and gene regulatory pathways with gene expression data to improve the accuracy. Feature selection improves the accuracy of cancer microarray data set.

Lee and Erik [9] mentioned the common pathways found among five species such as S.cerevisiae, Homo Sapiens, C.elegans, Arabidopsis thaliana and Drosophila Melanogoster. It was found that seven of 69 pathways found in all species. This species specific pattern of pathway clustering reflects adaptations or evolutionary events concerned with particular lineage. Although clustering of genes in a pathway was bulit in prokaryotes due to transcriptional operons which is absent in eukaryotes thereby finding difficult in clustering of genes in eukaryotic pathway. S.cerevisea and C.elegans showed significant clustering when compared to clustering of genes in human pathways.

Stephen et al., [10] proposed the unbiased approach for selecting features with diagnostic capacity from massive data sets. Refining the dataset to train the classifier is the basis of machine learning approach applied in computational biology for pathway analysis. It is a unique feature selection method to design parsimonious classifiers from microarray and high dimensional influenza datasets.

Johannes et al., [11] have introduced a collection of different Support Vector Machine based classification methods for improved gene selection and classification performance. The method contained in pathClass do not merely rely on gene expression data as seen in other trained classifiers but also exploits the information that is carried in gene-network data.

Khatri et al., [12] reviews on different approach followed by the researchers on pathway analysis to explore the possible understanding on differentially expressed genes and proteins for about a decade. No pathway analysis is complete without studying the features of pathway such as GO, PPI, microarray gene expression data. The researchers have manipulated their approach towards study of pathway analysis by meddling with either of these attributes as the prime feature in their analysis. The methods followed to study the pathway analysis ranges from statistics to next generation sequence(ngs)data analysis with its own advantages and disadvantages. Despite the limitations in ORA and FCS methods, these techniques still emerge as a tool to explore abundant number of novel genes and to study its relation in a pathway with the advent of fea-

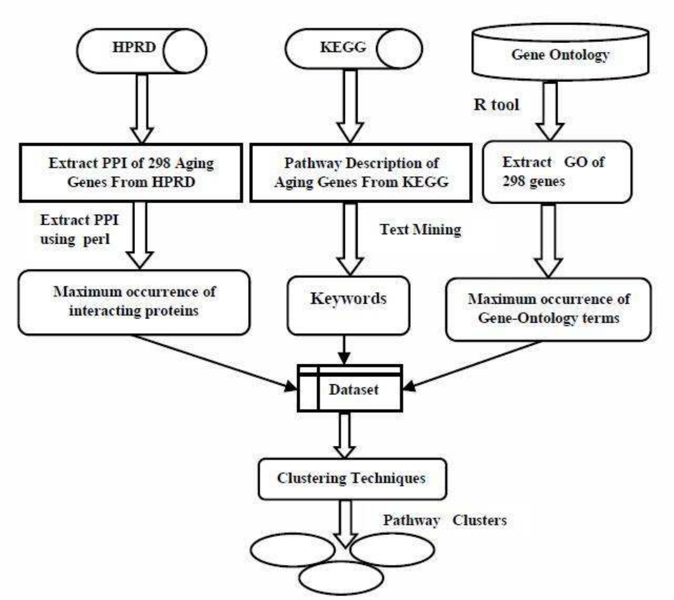
HPRD
KEGG
Gene Ontology
R tool
Extract PPI of 298 Aging Genes From HPRD
Pathway Description of Aging Genes From KEGG
Extract GO of 298 genes
Extract PPI using perl
Text Mining
Maximum occurrence of interacting proteins
Keywords
Maximum occurrence of Gene-Ontology terms
Dataset
Clustering Techniques
Pathway Clusters

Fig. 1. System Model for Pathway Clusters of Aging Genes(PCAG)

tures. By converting these attributes by statistical support, the researchers could challenge to provide a larger understanding of validated pathway analysis with more specificity, sensitivity and reliability.

III. Proposed Method

A. Problem Definition

The objective is to propose aging pathways that correctly identifies the unknown gene that would be prone to aging or not-aging pathways. This pathway information would serve as a platform for further research on pathways exclusively and other unknown genes that could be classified as aging gene in future.

B. System Model

In this paper, we are proposing a method called Pathway Clusters of Aging Genes(PCAG) that groups the aging genes into different pathway clusters. Fig. 1 depicts the system model of pathway clusters of aging genes.

C. Dataset Preparation

We focused our study on aging genes. 298 aging genes were collected from the GenAge which are already curated as aging [13]. These aging genes description are annotated from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14], as it is commonly called as a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. The pathway database records networks of molecular interactions in cells and their variants pertaining to particular organisms.

The following information such as type of pathway a gene belongs to and the KEGG classification of the particular gene and its detailed description are appended with each of the 298 aging gene and a database is created. The other attributes that are retrieved are Gene-Ontology(GO) [15] derived from R tool, Protein Protein Interaction(PPI)from HPRD database [16] *via* perl script and keywords from text mining. We have considered these attributes because PPI and GO forms platform for any pathway identification. Considering the extensive use of PPIs [17] and their receptors, the interaction between them assumes immense importance in classifying any proteins or genes to pathway. The PPI is used for drug discovery; it forms a crucial factor in identification of the drug targeted against specific receptors. Hence with aging too the identification of proteins interacting with the aging genes will serve as the basis for drug discovery.

GO of 298 genes were obtained by running the R. The three terms in GO, such as cellular components, molecular function and biological process of the genes were collected. The GO terms chosen as attribute are protein binding, DNA binding, ATP binding, sequence-specific DNA binding transcription factor activity, identical protein binding, metal ion binding and transcription factor binding. The binding factors essentially constitute molecular function which is critical in the identification of pathways as pathways are the network of genes, its binders and its interaction with receptors and gene products.

GO terms describes the functionality of aging genes and hence involvement in the pathway networks. These terms were taken due to maximum occurrence among aging genes. The frequency is set to 30 and above. The PPI is obtained for 298 genes from HPRD database. Protein interaction for 21 proteins was missing. However these missing values were substituted with 0 value. The aim is to focus on the number of interacting partners. The primary secondary proteins that interact with the aging gene results are tabulated.

D. Text Mining

In order to prepare the dataset, we have considered keywords identified from text mining as one of the attributes. Eight pathways of aging genes are manually identified with detailed study from different pathway databases. Pathways are invariably large in number but our target is aging gene clusters and it is found that aging largely affects metabolic activity [1] and hence pathways were concised to metabolism and its descendents. Although it should be noted that, there are several pathways but our study of interest is refined to eight pathways, because 298 aging genes were clustered largely to each of the selected eight pathways.

Initially the 13 pathways were taken but the grouping of genes overlapped between the similar pathways. Hence 8 pathway clusters were taken to overcome such pathway distribution which has appropriate aging genes relevant to that particular pathway. Keywords relevant to these 8 aging pathways are curated using text mining approach using R [18]. The KEGG pathway description for 298 aging genes is consolidated. The pathways summed up to 354 after eliminating the redundancy because a particular gene will have the tendency to form network in different pathways and the same applies to 298 genes. This forms a series of repeated pathway which should thus be eliminated.

A five step procedure is followed to identify the keywords.

- Extracting text from different pathway databases.

- Converting the extracted text to a data frame and then to a corpus.
- Subject a corpus to a couple of transformations including removing punctuations, stopwords and steming words to build Document Term Matrix(DTM).
- Term Document Matrix(TDM) is built from a processed corpus.
- Frequent terms are found out with frequency no less than 5.

The minimum word length is selected to identify the desired keyword applicable to the pathway. A total of 60 keywords are short listed and selected as an attribute. The uniqueness of these keywords are in such a way that it has common feature among various classes that are described. These keywords are treated as binary attributes to each of the pathway classes. The GO terms, PPI and keywords were selected based on its frequency of occurrence. A total of 116 attributes are fed to 298 genes and presence or absence of these attributes are checked for individual gene. The pivot table is drawn to find the frequency of occurrence and macro code is written to prepare the data set.

E. Algorithm

A number of clustering techniques are available to group the data objects into different clusters. In order to cluster the genes, we are using k-means clustering technique as it forms a small number of clusters from a large number of obeservations [5]. The K-means algorithm for pathway clusters of aging genes is as shown in algorithm 1.

Algorithm 1 KPCAG: K-means Pathway Clustering of Aging Genes

```
input : A dataset $D$, of $n$ aging genes, $g_1, g_2, ..., g_n$
        and a desired number of pathway clusters $k$.
output: Set of $k$ pathway clusters.
Begin
        (i) randomly select $k$ aging genes as the
            centroids for $k$ aging pathway clusters;

repeat
        (ii) assign each aging gene $g_i$, to a pathway cluster
             such that the distance between the gene $g_i$ and
             the cluster center is comparably minimum
             among all the $k$ pathway clusters;
        (iii) update the centroids for each cluster based on
              the genes assigned to the clusters.

until no change;
End
```

IV. Results and Discussions

Thirty one genes were clustered under Repair pathway class. Twenty one under insulin signaling, 54 under apoptosis, 38 mitochondria, 35 under hamper in immune systems, 29 under disease, 41 under metabolism. The eighth cluster has a network of different pathways, which constitutes 49 aging genes. The resulting pathway clusters of aging genes from K-means algorithm are listed as in the Table I.

- Cluster 1 (Repair Pathway): GenAge has identified 33 genes under repair group [13]. Applying textmining, PPI and Go attributes, 31 aging genes falls into repair pathway cluster, showing an accuracy of 92%. For example the family of ERCC falls under repair pathway cluster. Further consolidation in text mining attribute can give appropriate results on subclasses of Repair pathway.

- Cluster 2 (Insulin Signaling Pathway): $KEGG$ database has identified above 200 genes as participating in insulin signaling pathway [14]. There are about 21 aging genes grouped under same cluster and the functionality of all these genes significantly play a vital role in insulin signaling. For example INS, which is the critical protein and whose receptors form a network of other protein or gene products has been placed under this cluster. Other genes that are crucial in insulin signaling are Growth Harmone Receptor (GHR) family.

- Cluster 3 (Apoptosis Pathway): The Apoptosis database affiliated to NCBI has a collection of genes that take part in cell death and differentiation. Majority of 54 genes have been grouped under this cluster. The TOP family of genes play important role in cell senescence and they are rightly clustered under functioning of apoptotic class.

- Cluster 4 (Mitochondria Pathway): Oxidative stress and calorie restrictions are key factors for the functioning of ATP phosphorylation in mitochondria. 38 genes have been clustered under this class, whose functionality reflects either stress or oxidative response. Family of Heat Shock proteins (HSPA1A) falls under this cluster, whose major role is activation during stress or calorie restrictions.

- Cluster 5 (Immune Pathway): Aging genes have been identified in hampering the immune system. These genes have predominantly effected in centenarians. The complement system, t-cells, macrophages, interleukins and b-cells together constitute the effective functioning of immunity in human body. The interleukins family of genes (ILs) are significantly fallen under this cluster.

- Cluster 6 (Disease Pathway): It is learnt that, aging genes gradually lead to disease due to various factors such as age, weight, mutation, expose to drugs and some by birth default. One of the notable gene, APOE known to cause Alzheimer disease in humans has grouped under this cluster. Other various progeria syndrome identified were Hutchinson-Gilford progeria syndrome, dementia causes and bloom syndrome etc., all of which are aging disorders.

TABLE I. **Pathway Clusters of Aging Genes Obtained from K-means Algorithm**

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
A2M	GH1	ABL1	ADCY5	AGTR1	APO E	APP	AKT1
APEX1	GHR	AIFM1	AGPAT2	APOC3	BCL2	ATP50	CCNA2
APTX	GHRH	AR	BAK1	ARHGAP1	BMI1	BSCL2	CDC42
ATM	GHRHR	ARNTL	CREB1	BDNF	BRCA2	CACNA1A	CDK1
ATR	GSK3A	ATF2	EEF2	C1QA	CISD2	CAT	CLU
CHEK2	GSK3B	BAX	EGFR	CETP	EFEMP1	CNR1	COQ7
ERCC1	IGF1	BLM	EPOR	DGAT1	FGF21	CTGF	CSNK1E
ERCC2	IGF2	BRCA1	ERBB2	EGF	FOXM1	DBN1	CTNNB1
ERCC3	IGFBP2	BUB1B	HRAS	EIF5A2	GSTP1	DLL3	E2F1
ERCC4	IKBKB	BUB3	HSP90AA1	EMD	H2AFX	ELN	FLT1
ERCC5	INS	CDK7	HSPA1A	FGF23	HDAC1	EPS8	FOXO1
ERCC6	KCNA3	CDKN1A	HSPA1B	FGFR1	HDAC2	GCLC	FOXO3
ERCC8	MTOR	CDKN2A	HSPA8	GCLM	HIC1	GPX1	FOXO4
FEN1	PAPPA	CDKN2B	HSPA9	GSR	HTT	GPX4	HBP1
GTF2H2	PARGC1A	CEBPA	HSPD1	IL2	IGHBP3	GRB2	HESX1
HELLS	SIRT1	CEBPB	IGF1R	IL2RG	MAX	GRN	HMGB1
NBN	SIRT6	CLOCK	INSR	IL6	MDM2	GSS	HMGB2
PARP1	SLC13A1	CREBBP	IRS1	IL7	MLH1	GSTA4	HOXB7
PCNA	TCF3	DDIT3	IRS2	IL7R	PCMT1	KL	HOXC4
POLB	TGFB1	EEF1A1	JAK2	LEP	PLAU	MAPT	HSF1
POLD1	UCHL1	EEF1E1	NFKB2	LEPR	PRDX1	MIF	HTRA2
POLG	-	EGR1	NGF	LMNB1	RELA	MSRA	AP3K5
PRKDC	-	EP300	NGFR	LRP2	SIRT3	MT-CO1	APK14
RECQL4	-	ESR1	NOG	PDGFB	SIRT7	NRG1	MAPK3
RPA1	-	FAS	PDPK1	PIK3CA	SNCG	PCK1	MAPK8
TP53	-	FOS	PIK3CB	PMCH	TNF	PEX5	MAPK9
TP53BP1	-	HDAC3	PIK3R1	PTGS2	TP63	PON1	MYC
WRN	-	HIF1A	PIN1	RAE1	TP73	PPM1D	NCOR1
XPA	-	JUN	PLCG2	SDHC	UBB	PTPN1	FE2L2
XRCC5	-	JUND	PPARG	SOD1	-	PYCR1	NFKB1
XRCC6	-	LMNA	PRKCA	SOD2	-	RAD51	NR3C1
-	-	MED1	PRKCD	UCP1	-	RAD52	DGFRA
-	-	MT1E	PSEN1	UCP3	-	RGN	DGFRB
-	-	MXD1	PTK2	VCP	-	SST	PML
-	-	MXI1	PTK2B	ZMPSTE24	-	STK11	POLA1
-	-	NCOR2	PTPN11	-	-	STUB1	PPARA
-	-	NFKBIA	RET	-	-	SUN1	PPP1CA
-	-	NUDT1	SOCS2	-	-	TXN	PTEN
-	-	POU1F1	-	-	-	UCP2	RB1
-	-	PROP1	-	-	-	VEGFA	SIN3A
-	-	S100B	-	-	-	YWHAZ	QSTM1
-	-	SHC1	-	-	-	-	SSTR3
-	-	SP1	-	-	-	-	STAT3
-	-	SUMO1	-	-	-	-	STAT5A
-	-	TBP	-	-	-	-	STAT5B
-	-	TERC	-	-	-	-	TAF1
-	-	TERF1	-	-	-	-	TERT
-	-	TERF2	-	-	-	-	TFDP1
-	-	TFAP2A	-	-	-	-	UBE2I
-	-	TOP1	-	-	-	-	-
-	-	TOP2A	-	-	-	-	-
-	-	TOP2B	-	-	-	-	-
-	-	TOP3B	-	-	-	-	-
-	-	TPP2	-	-	-	-	-

- Cluster 7 (Metabolism Pathway): 41 genes are grouped under a single cluster, whose functions are associated with carbohydrate, protein, fatty acid, vitamins and mineral metabolism.

- Cluster 8 (Others): A chunk of 49 genes were grouped under single cluster which had overlapped aging pathways. These genes and its products play role in multi pathways. Hence grouping these genes into single cluster will not justify its role in single pathway.

The aging genes in the obtained eight pathways are cross checked with its presence in the master database for metabolism(KEGG) to accurately cluster the genes into its corresponding pathways under metabolism and its descendents clusters. However BioCarta [19] and GeneCards [20] also supports our results. This validates the clustering of aging genes to its corresponding pathways.

A. Comparison with other Method

Pang's method [4] proposes clustering through Random forest classification and it focuses on varied number of informative genes. The sample size being diffusive and presents possible crosstalk between the similar function of the pathways, whereas in our method clustering focuses on aging gene whose sample size is fixed to 298. The objective is to extract the features for these studied aging genes and lay platform for future novel gene that will have the probability to be clustered under any of these pathways strictly on the basis of feature extraction such as GO, PPI and occurrence frequency of characteristics of pathways. Pang's method is not biased with feature extraction terms but its analysis is based on gene expression studies and provides a conceptual framework for the overall understanding of molecular traits of pathways in human. However considering gene expression values as one more attribute in the feature selection will improve the efficiency is our estimate, but that is in store for future research work.

V. Conclusion

Our proposed method PCAG allow bioinformaticians and biologists to investigate how aging genes within pathway are related to each other and understand possible cross talk between aging pathways in a cluster. These attributes information in the dataset and final clustered pathway inference will also be helpful in pharmacology essentially in the drug discovery. However the work is limited in terms of prerequisite issue where in without prior biological knowledge, it would be difficult to cluster the newly gene into identified pathways. The additional feature selection incorporation such as gene expression and SNP values can remarkably improve the pathway identification to greater heights.

References

[1] B. P. Braeckman, K. Houthoofd, and J. R. Vanfleteren, "Intermediary Metabolism," *WormBook: the Online Review of C.elegans Biology*, pp. 1–24, February 2009.

[2] S. K. Kim, "Common Aging Pathways in Worms, Flies, Mice and Humans," *Journal of Experimental Biology*, vol. 210, no. 9, pp. 1607–1612, March 2007.

[3] C. H. Marica and A. Y. Bruce, "The Aging Stress Response," *Molecular Cell*, vol. 40, no. 2, pp. 333–344, October 2010.

[4] H.Pang and H. Zhao, "Building Pathway Clusters From Random Forests Classification Using Class Votes," *BMC Bioinformatics*, vol. 9, no. 1, February 2008.

[5] J.Han and M. Kamber, "Data Mining Concepts and Techniques," in *Second Edition, Morgan Kaufman*.

[6] N. Barzilai, D. M. Huffman, R. H. Muzumdar, and A. Bartke, "The Critical Role of Metabolic Pathways in Aging," *Diabetes*, vol. 61, no. 6, pp. 1315–1322, June 2012.

[7] M.A. Ibrahim, S. Jassim, M.A. Cawthrone, and K. Langlands, "A Pathway-based Gene Selection Method Provides Accurate Disease Classification," *International Journal of Digital society(IJDS)*, vol. 2, no. 1, pp. 566–573, December 2011.

[8] B. Nirmala, T. Kahveci, S. Goodison, Y. Sun, and S. Ranka, "Pathway-Based Feature Selection Algorithm for Cancer Microarray Data," *Advances in Bioinformatics*, 2009.

[9] J. M. Lee and L. L. S. Erik, "Genomic Gene Clustering Analysis of Pathways in Eukaryotes," *Genomic Research*, vol. 13, no. 5, pp. 875–882, 2003.

[10] Stephen O'Hara, K. Wang, R.A. Slayden, A.R. Schenkel, G.Huber, C.S. O'Hern, M.D. Shattuck and M. Kirby, "Iterative Feature Removal yields Highly Discriminative Pathways," *BMC Genomics*, vol. 14, no. 1, 2013.

[11] M. Johannes, H. Frohlich, H. Sultmann and T.Beibbarth, "pathclass: An R-Package for Integration of Pathway Knowledge into support Vector Machines for Biomarker Discovery," *Bioinformatics*, vol. 27, no. 10, pp. 1442–1443, March 2011.

[12] P. Khatri, M. Sirota and A. J. Butte, "Ten Years of Pathway Analysis: current Approaches and Outstanding Challenges," *PLoS Computational Biology*, vol. 8, no. 2, February 2012.

[13] R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. costa, V. E. Fraifeld and J. P. de Magalhaes, "Human Ageing Genomic Resources: Integreted Databases and Tools For the Biology and Genetics of Ageing," *Nucleic Acid Research*, vol. 41, pp. 1027–1033, November 2012.

[14] K. Minoru, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, "Data, Information, Knowledge and Principle: Back to Metabolism in KEGG," *Nucleic Acid Research*, vol. 42, pp. 199–205, November 2013.

[15] Ashburner, Michael, A. B. Catherine, A.B. Judith, b. David, H. Butler, J.C. Michael, Allan P. Davis et al., "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[16] Prasad, T.S. Keshava, G. Renu, K. Kumaran, K. Shivakumar, K. Sameer, M. Suresh, Deepthi Telikicherla et al., "Human Protein reference Database-2009 Update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[17] D. M. Suresh, R. T. Hiren, M. P . Dinesh, S. D. Prashanth and J. Chacko, "Impact of Proton Pump Inhibitors On Efficacy of Clopidogrel: Review of Evidence," *Indian Journal of Pharmacology*, vol. 43, no. 2, pp. 183–186, March-April 2011.

[18] Zhao, Yanchang,, "R and Data Mining: Examples and Case Studies," 2012.

[19] Nishimura, Darryl, "BioCarta: Biotech Software and Internet Report," *The Computer Software Journal for Scient*, vol. 2, no. 3, pp. 117–120, November 2013.

[20] Rebhan, Michael, C. Vered, P. Jaime, and L. Doron, "GeneCards:Integrating Information about Genes, Proteins and Diseases," *Trends in Genetics*, vol. 13, no. 4, p. 163.