Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

Distribution of RNA 5-mers in Epigenetic Modification Regions and Genes Interactions

Dariush Salimi*[1], Mohaddese Salimi†[2] and Ali Moeini‡[3]

[1,2]Department of Computer Engineering, Faculty of engineering, University of Zanjan, Zanjan, Iran
[3]Department of Algorithms and Computation, Faculty of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran

ABSTRACT

The demand for extracting sophisticated features, capable of effectively predicting gene interaction networks, from DNA or RNA sequences has increased in computational biology. The epigenetic modifications along with their patterns have been intensely recognized as appealing features affecting on gene interaction. However, studying sequenced-based features highly correlated to this key element has remained limited. In this paper, classification of 23 genes in PPAR signaling pathway associated with muscle fat tissue in human was proposed based on statistical distributions of the specified RNA-5-mers abundance. Then, we suggested that these 5-mers highly correlated to epigenetic modifications can efficiently categorize the different gene interactions, particularly co-expression interaction and physical interaction. Our results were evaluated according to

Keyword: 5-mers distributes, epigenetics modifications, genes interactions

ARTICLE INFO

Article history:
Research paper
Received 09, May 2022
Received in revised form 14, August 2022
Accepted 11, October 2022
Available online 30, December 2022

AMS subject Classification: 92-08; 92C42; 92D10.

*dsalimi@znu.ac.ir
†salimimohaddese@gmail.com
‡moeini@ut.ac.ir

1 Abstract continued

GeneMania web interface and shows that the geometric distribution of 5 mers in the epigenetic modifications region indicates the proportion of most physical interactions and the Poisson distribution the proportion of most Co-expression between genes

2 Introduction

Undoubtedly, one of the fastest growing area in computational biology is extracting dominant features capable of dramatically differentiate gene interaction networks [13,16]. Truly, there are clear advantages of the use of these dominant features in developing a high-quality model, leading to more accurately discriminating results. Proceeding this track, it has been figured out epigenetic modification as one of the well-recognized class of these features, such as methylation and acetylation, and can be applied to predict gene expression [4,10]. From the statistical standpoint, considerably-correlated features to epigenetic modifications can be served as markers in prediction of gene expression[17], among which k-mers can be supposed to study as appealing markers to specify the location as well as level of epigenetic modification on chromosomes[11]. formatter will need to create these components, incorporating the applicable criteria that follow. However, studies on applying these K-mers in predicting gene interaction has fairly limited.

K-mers, small repetitive sequence with 2 up to 10 nucleotides, are treated as an appropriated tool to identify CpG islands, set propitious probes in microarray experiments, determine epigenetic modification patterns on the chromosome, assay relations between small segment sequences and different organisms genomes, identify of various breeds and organisms, fingerprint in bacteria for identification of diseases, prediction of genes interactions and study the association of genes with transcription factors[15,11]. Recently, Martin Sauk suggested a k-mer-based method to replace mapping reads in FASTQ formatted sequencing raw data[11]. In addition, a k-mer-based software for capturing local RNA variation from a set of standard RNA-seq libraries was suggested, independently of a reference genome or transcriptome [12].

Applying statistical distribution has been well-recognized as a powerful tool in many different fields in computational biology. The empirical distribution of DNA k-mers in different studies was examined even prior to the sequencing of large genomes, providing tenable probabilistic models for these k-mers [7]. One of the earliest research indicated that negative binomial is a distribution for biology data due to higher variability between biological replicates where an interesting variable having negative binomial distribution is the counts of RNA. [5]. In addition, Carlos A. C. Bastos et. al indicated that distribution of protein – coding inter-dinucleotide distances in the human genome is geometric distribution, providing an analysis method the whole genome [2]. As well, the negative binomial distribution, in duration modeling, was suggested for lengths in a DNA strand[9]. In another recent study, Trung Nghia Vu et.al proposed a Beta-Poisson distribution for gene expression data distribution[18]. Furthermore, the distribution of mutants emerging from single bursts in the RNA bacteriophage $\varphi 6$ was figured out a Poisson distribution[6].

Seemingly, statistical distribution of variable of interest can be an efficient approach to discover specified patterns in genome, proteins, and other fields.

3 Materials and methods

3.0.1 Datasets

The epigenetic modifications data were downloaded from
(http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx) for acetylation and l(http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx) for methylation. A widely used modification module including 13 modification sites revealed at 3286 promoters and identified, by Wang et al and Braski et al, was addresses Based on the data in these two authors" web sites These includes H2A-Z, H2BK12ac, H2BK120ac, H3K4me1 ,H3K4me2 ,H3K4me3 ,H3K9me1 ,H3K18ac ,H3K27ac ,H3K36ac ,H4K5ac ,H4K8ac and H4K91ac.This data set involved coordinates,having 23 up to 36 nucleoids length for all uniquely mapped rides for each of modifications [1,19].
The gene data involved 23 genes are available on
($http://www.kegg.jp/dbget-bin/www_bget?hsa03320$).
The gene set of interest in this study is a 23-gene set of PPAR signaling pathway associated to muscle fat tissue in human (Tables 1). Moreover, the DNA and RNA sequences of the corresponding genes were obtained from NCBI database[8]. This pathway is one of the dramatically dominant pathways effective on the fat metabolism, controlling gene expression network in adiposeness, lipid metabolism and metabolic homeostasis. The PPARs are member of unclose receptors and there are three PPARs in mammalians called: 1- PPAR α, 2-PPAR β, 3-PPAR γ . Furthermore, all computational works were performed using a toolkit developed in our lab in MATLAB, EASYFIT, and R.

3.0.2 The proposed method

In this research, we are interested classifying genes based on the best distribution for the 5-mers abundances on these genes, as it was indicated that 5-mers can be taught of as important feature in prediction of epigenetic modification and genes interactions[15,3,14].our research as abstract show in Figure 1.
Specifying 5-mers across whole RNA required thoroughly accounting critical considerations to achieve accurate results, developing two main stages to extract 5-mers related to epigenetic modifications. In the first stage, the RNA sequences were divided into 250-nucleotide segments, due to having different lengths. Then, we kept only the set of 5-mers that observed with at least two repeats in the RNA, owing to minimize the risk of including 5-mers involving sequencing errors. In the next step, the total frequencies of these 5-mers in all 23 RNA sequences were calculated. Finally, the 5-mers with more than 10 frequencies were selected for further analysis and mapped on the DNA sequences of all 23 genes and 13 modifications, separately.
Second, the intervals of epigenetic modifications on the genes were symmetrically extended

Gene	Ch	NCBI ID	Start	End
PPARD	6	NC_000006.12	35342558	35428191
PPARA	22	NC_000022.11	46150547	46243756
PPARG	3	NC_000003.12	12287850	12471013
TBL1X	X	NC_000023.11	9463295	9719740
TBL1XR1	3	NC_000003.12	177020754	177197260
HSD11B1	1	NC_000001.11	209686180	209734950
LIPA	10	NC_000010.11	89213569	89252039
PPARGC1A	4	NC_000004.12	23792021	24472771
NCOA4	10	NC_000010.11	46005088	46030714
FABP1	2	NC_000077.6	88122982	88128131
ANGPTL4	19	NC_000019.10	8364127	8374375
CYP4A11	1	NC_000001.11	46929174	46941499
ACOX1	11	NC_000077.6	8366263	116199045
ACOX3	4	NC_000004.12	8366263	8440725
FABP3	1	NC_000001.11	31360418	31373283
ACOX2	3	NC_000003.12	58505136	58537348
SAT1	x	NC_000023.11	23783158	23786223
PTGER2	14	NC_000080.6	44988111	45003820
ACAA1	3	NC_000003.12	38122710	38137242
RXRA	9	NC_000009.12	134326463	134440586
UGT2B4	4	NC_000004.12	69480165	69526014
HMGCS1	2	NC_005101.4	52427351	52445082
NCOA6	20	NC_000020.11	34714774	34825630

Table 1: The 23 human genes of PPAR signaling pathway

to 30 nucleotides in both directions, starting from the center of the original interval. Then, the abundances of these 5-mers in the corresponding modification interval were determined for each of epigenetic modifications and genes. Therefore, for each gene, a vector of elements according to the detected 5-mers was generated, including 719 different 5-mers.
Having extracted, we determined the empirical distribution of 5-mers frequencies under different epigenetic modifications including methylation and acetylation modifications, applying statistical tests. Different discrete distributions were examined including geometric, negative binomial, binomial, uniform, Poisson, logarithmic, and Hyper-geometric through Kolmogorov-Smirnov test. The analysis was performed on 719 different 5-mers found in each genes under each of epigenetic modifications, separately.

4 Results

Preliminary results exhibited that only three distributions including Poisson, negative binomial and geometric distribution were expedient and others were poor match to the

5-mers abundance. Interestingly, Poisson distribution is able to capture the variability of 5-mers frequencies in high number of genes under different epigenetic modifications. On the other hand, the resulting fitted distribution for most of remaining is negative binomial and geometric distribution, respectively.
On the other hand, to get more information, we further examined the distributions of 5-mers abundance in the specified genes under different epigenetic modifications. It can be seen that this feature in all genes but RXRA is not of the same distribution. Simply, for instance, the detailed results are provided the distribution plots of this feature in gene PPARA under epigenetic modifications of interest, Figure 2. As shown, Poisson distribution is the best fit for H2bk12ac, H3k27ac, H3k4me2, and H4k5ac and that of remaining is Negative binomial distribution. It should be pointed out that even in the same distribution, the means of 5-mers abundance under different modifications are not the same.

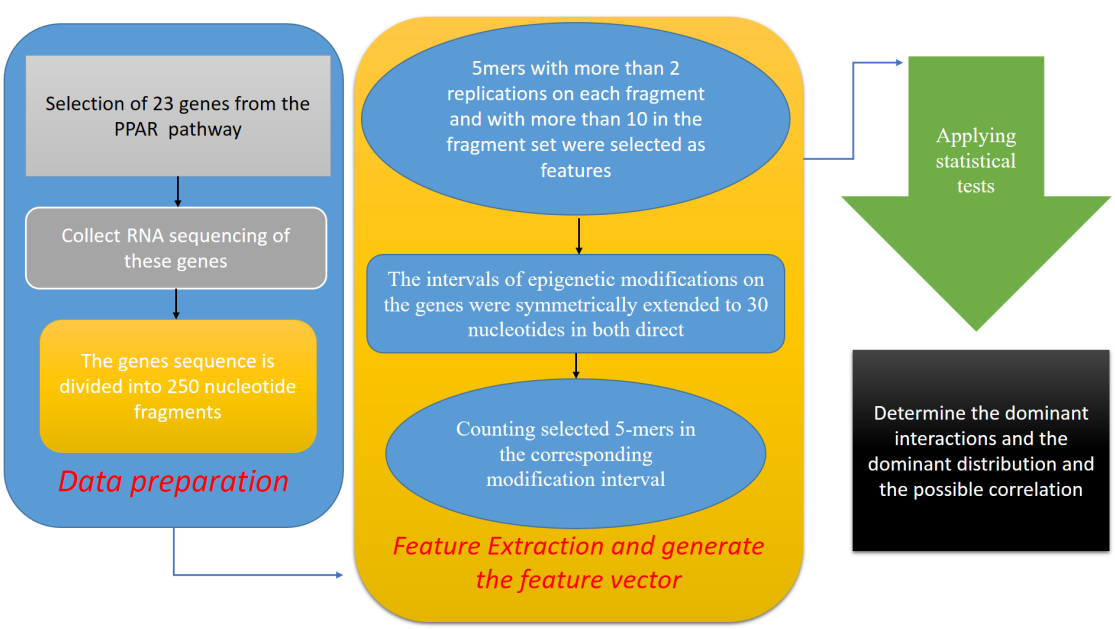
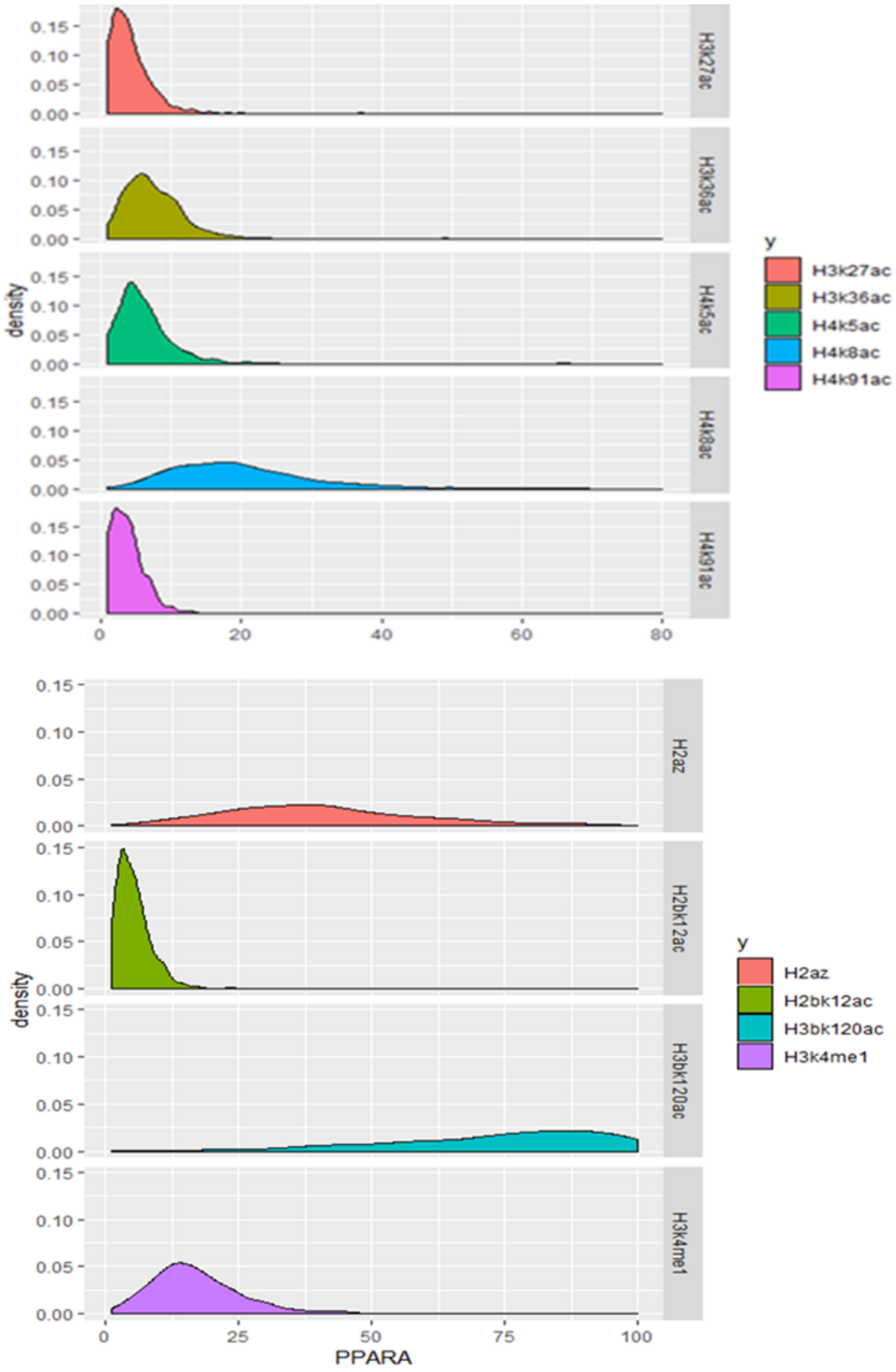


Figure 1: The graphical abstract of our research

Afterwards,classifying genes was mainly performed by attributing the genes of which 5-mer frequency distributions are the same distribution. The results are presented in Table 2.


density
0.15
0.10
0.05
0.00
H3k27ac
H3k36ac
H4k5ac
H4k8ac
H4k91ac
0
20
40
60
80
y
H3k27ac
H3k36ac
H4k5ac
H4k8ac
H4k91ac
H2az
H2bk12ac
H3bk120ac
H3k4me1
25
50
75
100
PPARA

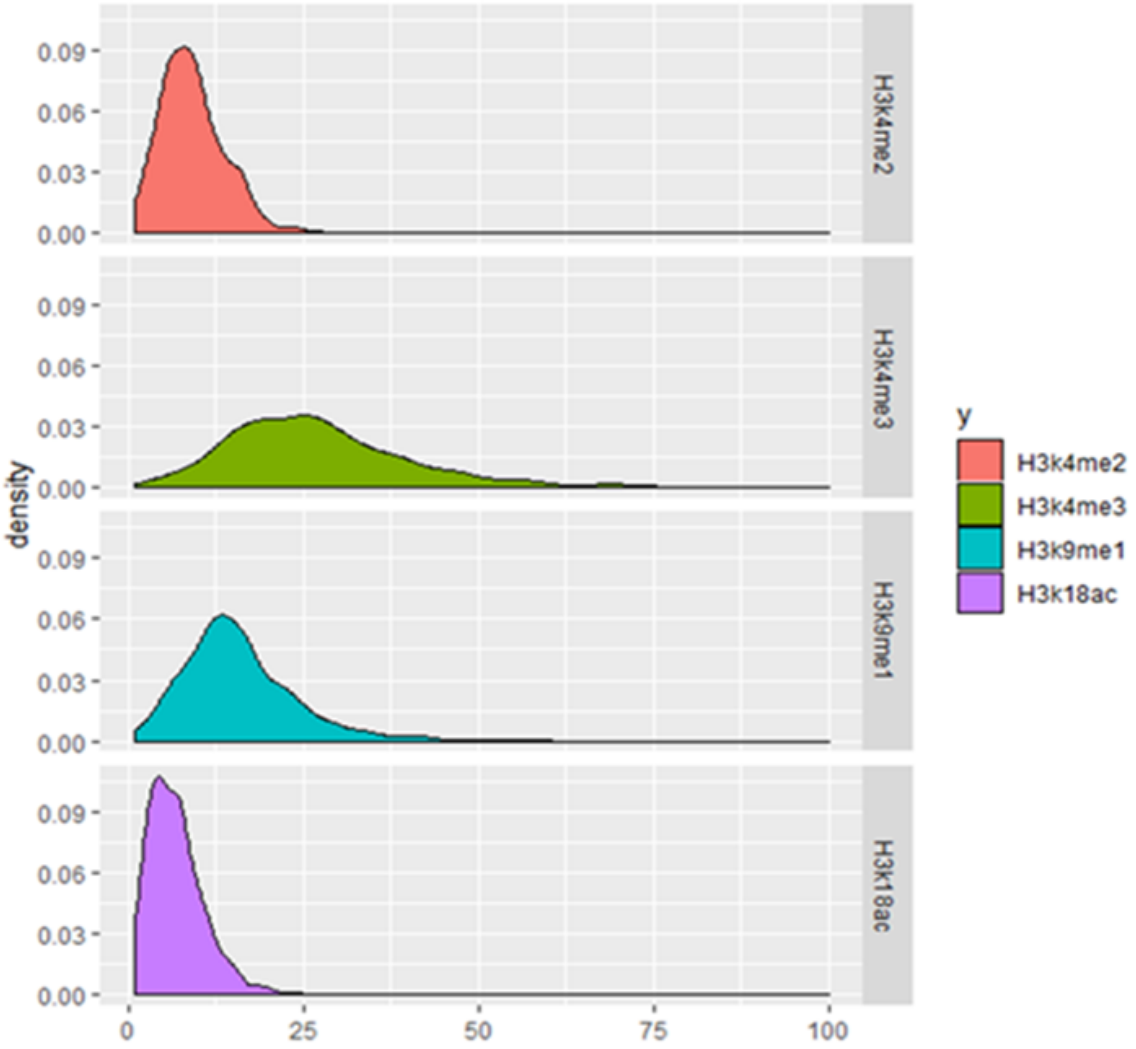


Figure 2: The distribution plot of 5-mers abundance in gene PPARA under 13 different epigenetic modifications.

These findings motivated us to biologically evaluate these resulted clusters of genes. In light of this, we applied GeneMania[20] and the accuracy results are shown in Table 3 and Figure 3.We found a quite acceptable accuracy in high number of compositions. As can be seen, under H2az epigenetic modification, the best performance achieved by Poisson distribution in predicting Co-expression network with network with 99.64 % precision.This high level of prediction can be seen in other epigenetic modifications including H2bk5ac, H3bk120ac, H3k36ac, H3k4me1, and H4k8ac. This means that if 5-mers were randomly sampled from this RNA sequence, the shape of 5-mers abundances follow a Poisson distribution where the mean and variance are the same. However, the distribution of interest in gene PPARA under other epigenetic modifications is Negative binomial, a two-parameter

distribution	Cluster1 (Negative Binomial)	Cluster2 (Geometric)	Cluster3 (Poisson)
H2az	PPARA PPARGC1A TBL1X TBL1XR1 HSD11B1	ACOX2 NCOA6 PPARD PPARG RXRA	ACOX3 CYP4A11 FABP3 LIPA NCOA4 UGT2B4 ACAA1
H3bk120ac	PPARD PPARA PARG TBL1X HSD11B1 PPARGC1A NCOA6	ACOX3 RXRA	TBL1XR1 LIPA NCOA4 FABP1 ANGPTL4 CYP4A11 FABP3 ACOX2 SAT1 ACAA1 UGT2B4
H2bk12ac	PPARD TBL1X ACOX3	PPARGC1 RXRA NCOA6	PPARA PPARG TBL1XR1 HSD11B1 LIPA FABP1 ANGPTL4 CYP4A11 FABP3 ACOX2 SAT1 ACAA1 UGT2B4
H2bk5ac	PPARD PPARA PPARG TBL1X PPARGC1 ACOX3	RXRA	TBL1XR1 HSD11B1 LIPA FABP1 ANGPTL4 CYP4A11 FABP3 ACOX2 ACAA1 UGT2B4 NCOA6
H3k18ac	PPARD PPARA TBL1X PPARGC1A ACOX3	RXRA	PPARG TBL1XR1 HSD11B1 LIPA FABP1 ANGPTL4 CYP4A11 FABP3 ACOX2 ACAA1 UGT2B4 NCOA6

distribution more efficient capable of explanation an over-dispersion variable than Poisson distribution.

When comparing the plots, we found that in situations where 5-mers abundances have more dispersion, negative binomial distribution has a better fit. In addition, the accuracy for negative binomial and geometric clusters are moderately acceptable, 66.88%, in most

distribution	Cluster1 (Negative Binomial)	Cluster2 (Geometric)	Cluster3 (Poisson)
H3k27ac	PPARD TBL1X	PPARGC1A RXRA	PPARA PPARG TBL1XR HSD11B1 LIPA FABP1 ANGPTL4 CYP4A11 ACOX3 FABP3 ACOX2 ACAA1 UGT2B4 NCOA6
H3k36ac	PPARD PPARA PPARG TBL1X	TBL1XR1 PPARGC1A RXRA	HSD11B1 LIPA FABP1 ANGPTL4 CYP4A11 ACOX3 FABP3 ACOX2 ACAA1UG T2B4 NCOA6
H3k4me1	PPARA TBL1X CYP4A11 ACOX3 ACOX2 NCOA6	PPARD PPARG TBL1XR1 PPARGC1A ANGPTL4 RXRA	HSD11B1 LIPA FABP1 FABP3 SAT1 ACAA1 UGT2B4
H3k4me2	PPARD TBL1X PPARGC1A ACOX3	PPARG RXRA	PPARA TBL1XR1 HSD11B1 LIPA NCOA4 FABP1 ANGPTL4 CYP4A11 FABP3 ACOX2 SAT1 ACAA1 UGT2B4
H3k4me3	PPARD PPARA TBL1X HSD11B1 PPARGC1A ACOX3 NCOA6	PPARG RXRA TBL1XR1	LIPA FABP1 ANGPTL4 CYP4A11 FABP3 ACOX2 ACAA1 UGT2B4

of epigenetic modifications.

Finally, to deeper understand how this feature is able of classifying the genes, we applied GeneMania to represent gene interaction network for each of epigenetic modifications.(Figure 3)

Interestingly, as it can be seen, removing hub genes such as PPARA and PPARG from

distribution	Cluster1 (Negative Binomial)	Cluster2 (Geometric)	Cluster3 (Poisson)
H4k5ac	PPARD PPARGC1A	RXRA NCOA6	PPARA PPARG TBL1X TBL1XR1 HSD11B1 NCOA4 FABP1 ANGPTL4 CYP4A11 ACOX3 ACOX2 SAT1 ACAA1 UGT2B4
H4k8ac	PPARD PPARA PPARG TBL1X HSD11B1 NCOA6	PPARGC1A RXRA	TBL1XR1 LIPA FABP1 ANGPTL4 CYP4A11 ACOX3 FABP3 ACOX2 SAT1 ACAA1
H4k91ac	PPARD TBL1X PPARGC1A ACOX3 NCOA6	RXRA	PPARA PPARG TBL1XR1 HSD11B1 ANGPTL4 CYP4A11 FABP3 ACOX2 SAT1 ACAA1

Table 2: Clustering genes under different epigenetic modifications based on distribution of 5-mers abundance in the genes

cluster 3 leads to 100 % accuracy in prediction. From biologically standpoint, this result states that these genes are of special importance in gene interaction network, indicating that different epigenetic modifications place these genes in different clusters.
To sum up, despite of that the prediction performances for both Pathway and Genetic interaction were not satisfactory, the high consistency with GeneMania results can be substantially seen for Physical interaction and in particular Co-expression interaction, supporting our proposed feature as a powerful tool to predict these two important networks.

5 Discussion

In this paper, we suggested a 5-mer-based feature to predict two key gene interactions, Co-expression and physical interaction networks in fat tissue. Our main approach was to find the best distribution for the 5-mer abundance to classify genes. We found that three commonly-used statistical distributions in biological fields, Negative binomial, geometric

and Poisson distribution, were efficiently match to the 5-mers abundance, predicting the interaction networks of which the genes are likely to be involved. The frequencies of 5-mers within RNA sequences were differently distributed under various different epigenetic modifications. As the Poisson distribution was proposed for rare DNA k-mers by Benny Chor et al[7], it is a satisfied distribution to predict Co-expression network. In addition, David Williams et al suggested the abundance distribution of DNA k-mers as a negative binomial[21] which one of proposed distribution in our research. These two distributions were suggested in other biological research [7,18].

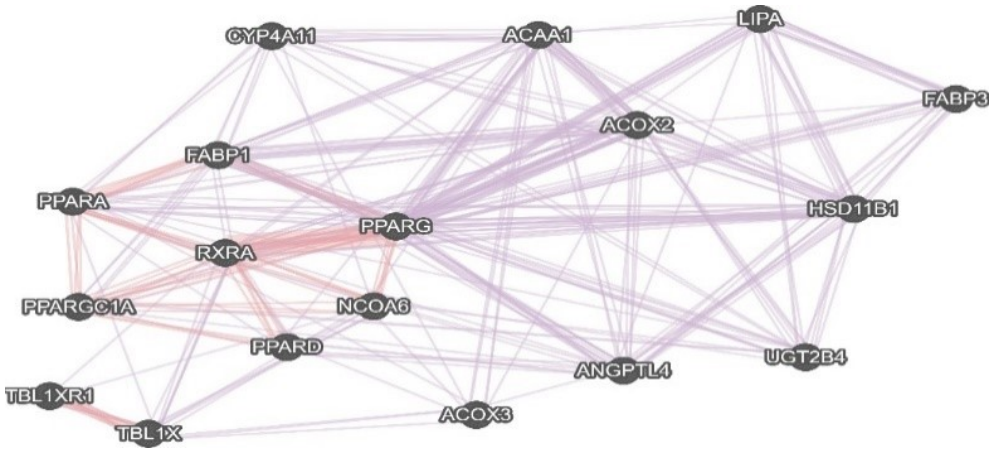
The findings proposed that the structure within the specified 5-mers space exposing various epigenetic modifications can be further involved in different networks. In fact, our study proposed that epigenetic modifications take place in regions of genes with different distributions of 5-mers abundance. This means that these 5-mers have beneficial information in predicting the gene interaction network where the genes of interest are involved.

Epigenetic Modification	Distribution	Physical interactions	Co-expression	Pathway	Genetic interactions
H2az	Geometric	66.88	27.38	4.39	1.35
H2az	Geometric	66.88	27.38	4.39	1.35
H2bk12ac	Geometric	66.88	27.38	4.39	1.35
H3bk120ac	Geometric	66.88	27.38	4.39	1.35
H2bk12ac	Geometric	66.88	27.38	4.39	1.35
H3k27ac	Geometric	66.88	27.38	4.39	1.35
H3k36ac	Geometric	66.88	27.38	4.39	1.35
H3k4me1	Geometric	39.6	38.83	21.58	0
H3k4me2	Geometric	66.88	27.38	4.39	1.35
H3bk120ac	Geometric	66.88	27.38	4.39	1.35
H3k4me3	Geometric	66.88	27.38	4.39	1.35
H4k8ac	Geometric	66.88	27.38	4.39	1.35
H2az	Poisson	0	99.64	0	0.36
H2bk12ac	Poisson	7.48	58.15	34.38	0
H2bk5ac	Poisson	0	100	0	0
H3bk120ac	Poisson	0	100	0	0
H3k18ac	Poisson	6.72	76.63	16.64	0
H3k27ac	Poisson	9.59	56.63	33.78	0
H3k36ac	Poisson	0	99.95	0	0.05
H3k4me1	Poisson	0	100	0	0
H3k4me2	Poisson	22.33	58.63	23.12	0
H3k4me3	Poisson	23.12	54.55	41.37	0
H4K5ac	Poisson	26.65	39.12	34.23	0
H4k91ac	Poisson	8.31	41.85	49.84	0
H2az	Negative Binomial	66.88	27.38	4.39	1.35
H2bk12ac	Negative Binomial	66.88	27.38	4.39	1.35
H2bk5ac	Negative Binomial	8.34	35.16	56.5	0
H3bk120ac	Negative Binomial	21	39.68	39.32	0
H3k18ac	Negative Binomial	66.88	27.38	4.39	1.35
H3k27ac	Negative Binomial	66.88	27.38	4.39	1.35
H3k36ac	Negative Binomial	66.88	27.38	4.39	1.35
H3k4me1	Negative Binomial	0	48.12	51.73	0.14
H3k4me2	Negative Binomial	66.88	27.38	4.39	1.35
H3k4me3	Negative Binomial	2.72	38.9	58.37	0
H4K5ac	Negative Binomial	66.88	27.38	4.39	1.35

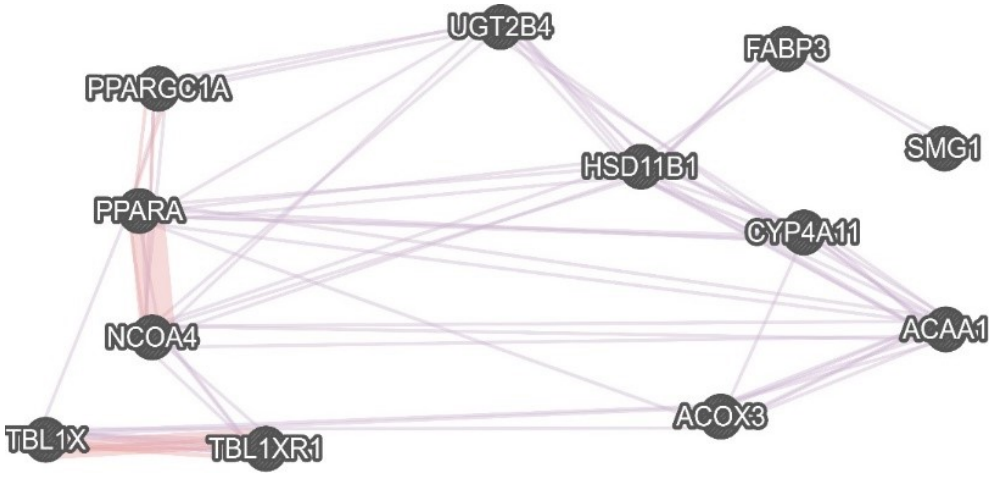
Epigenetic Modification	Distribution	Physical interactions	Co-expression	Pathway	Genetic interactions
H4k8ac	Negative Binomial	4.76	39.66	55.59	0
H4k91ac	Negative Binomial	66.88	27.38	4.39	1.35

Table 3: Clusterring genes under diffetent epigenetic modifications based on distribution of 5-mers abundance in the genes

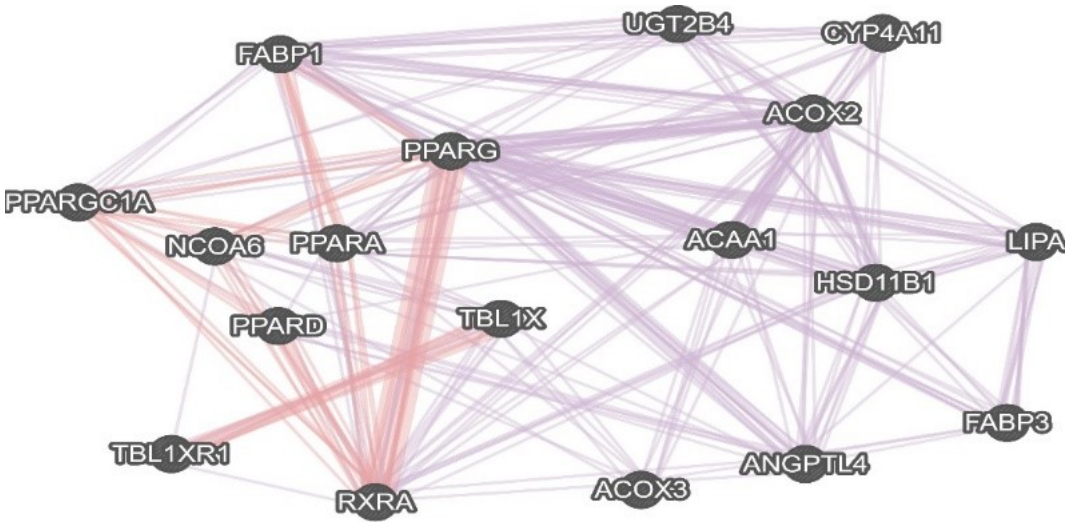
Furthermore, we figured out that genes such as PPARA and PPARG contributed in three different clusters are of special biologically importance in gene interaction network. In other words, these genes have distinct roles in different gene interactions, under different epigenetic modifications, and can be treated as hub genes.



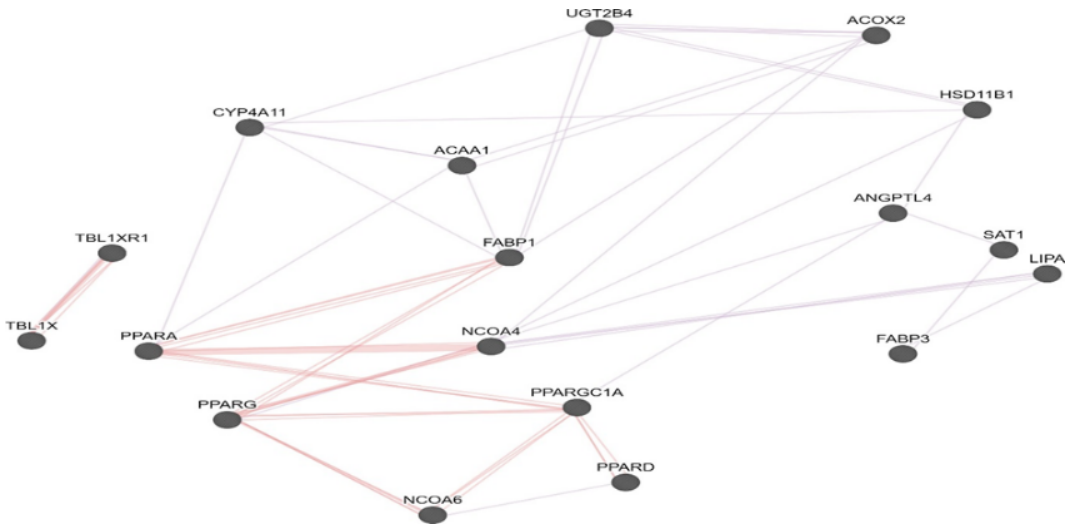
H3K27ac



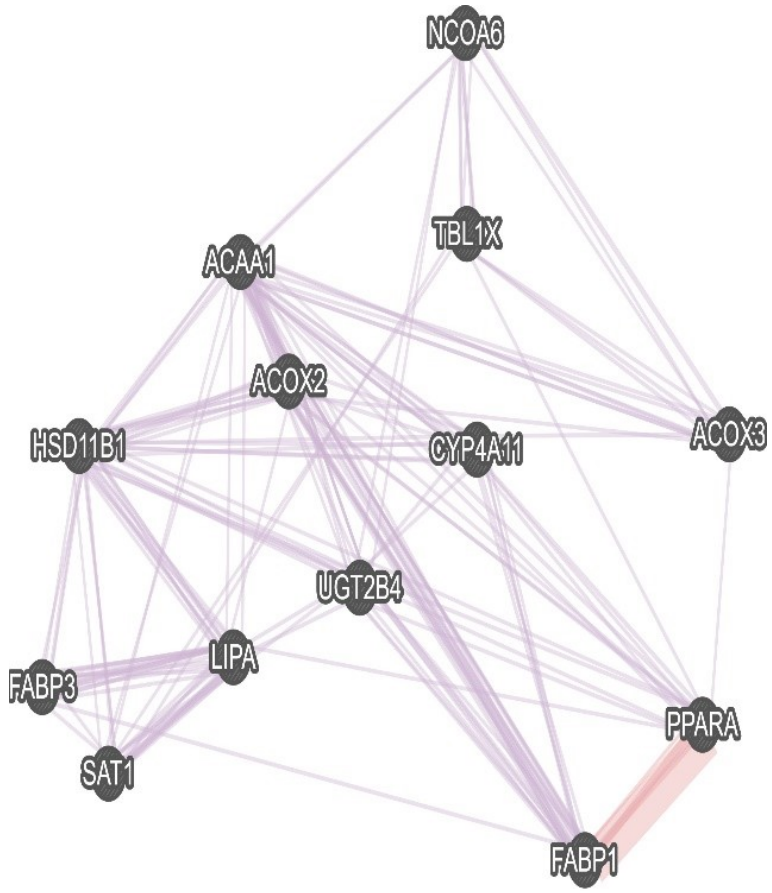
H2az



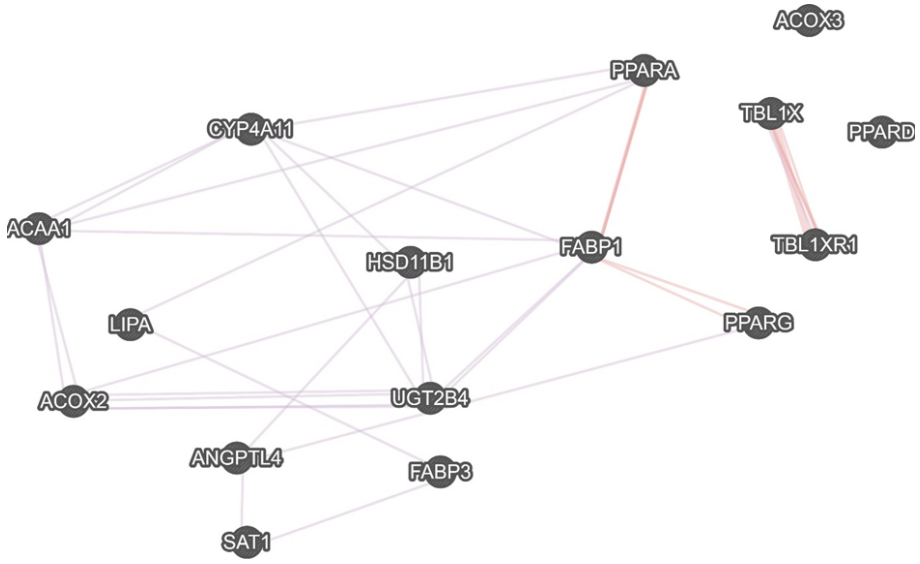
H3K36ac



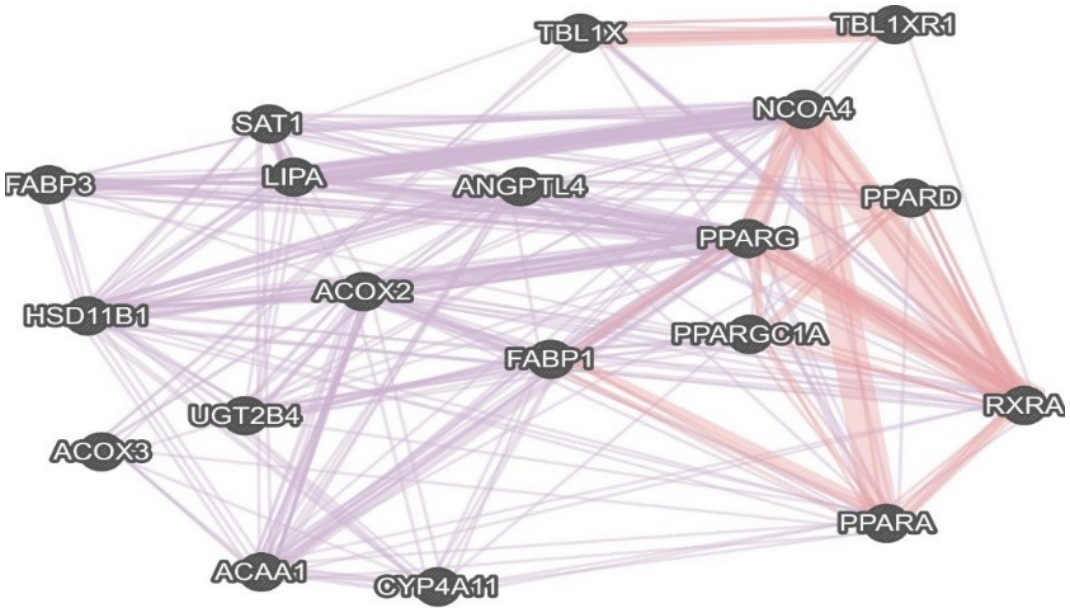
H2BK120ac



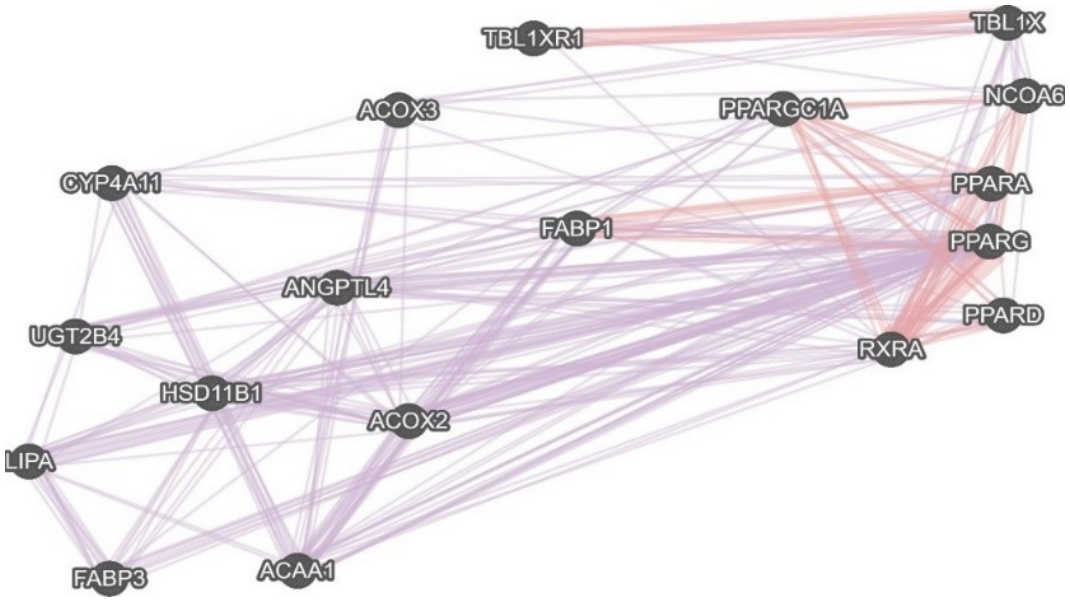
H3K4me1



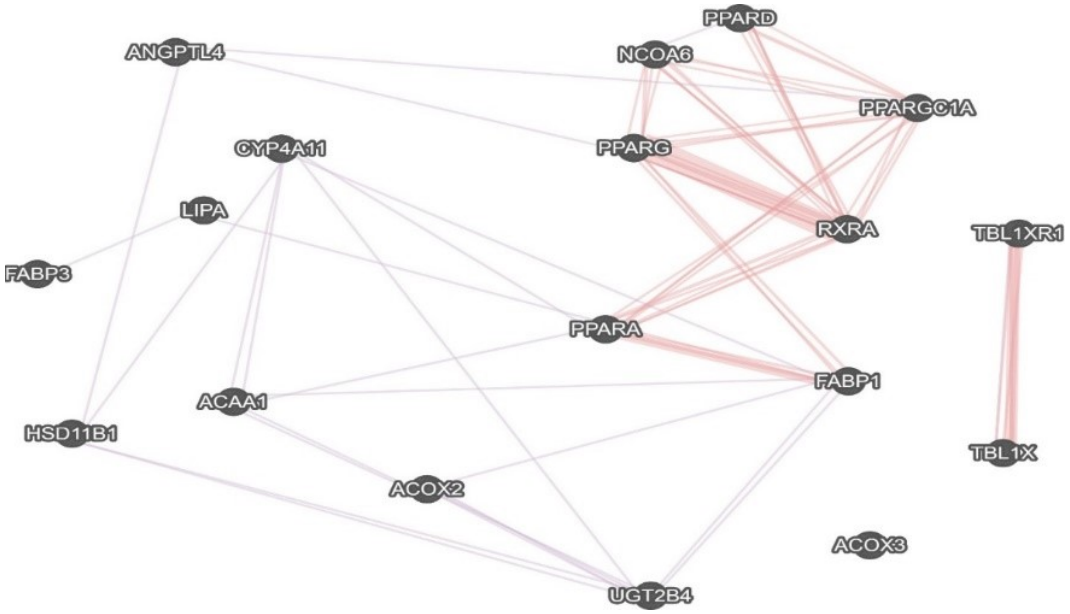
H2BK12ac



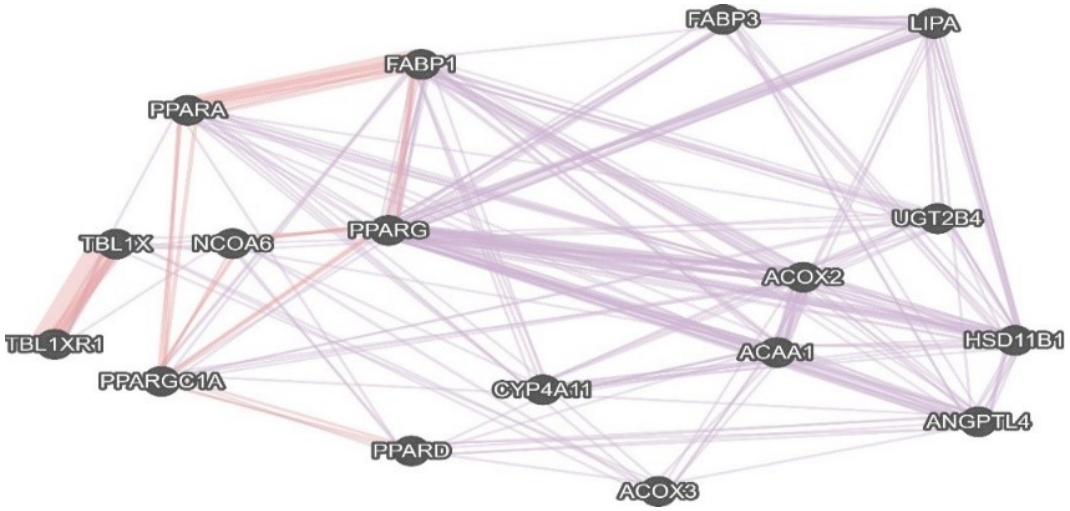
H3K4me2



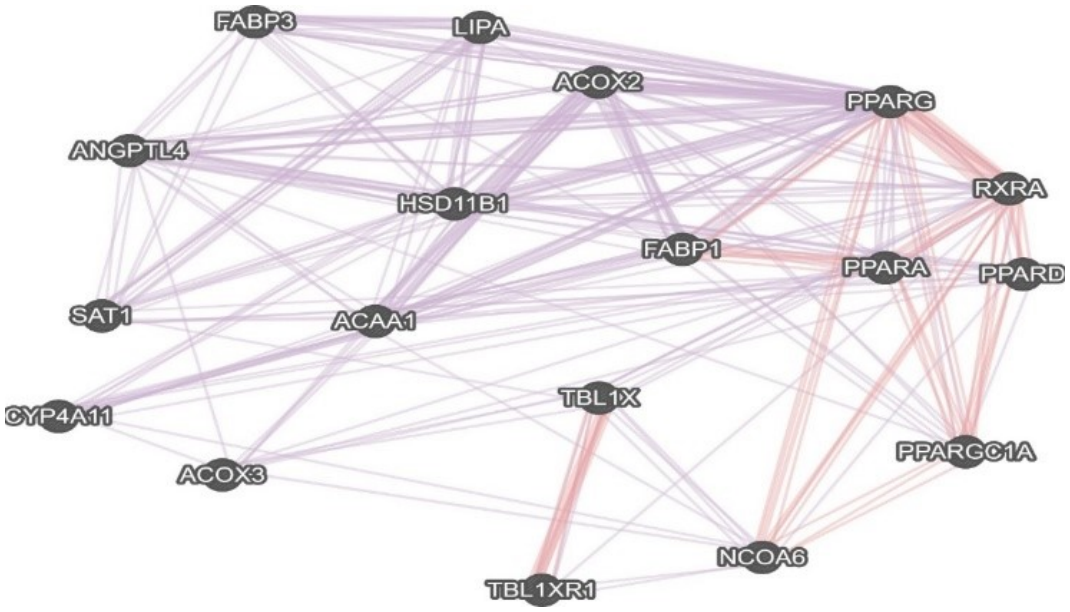
H2BK5ac



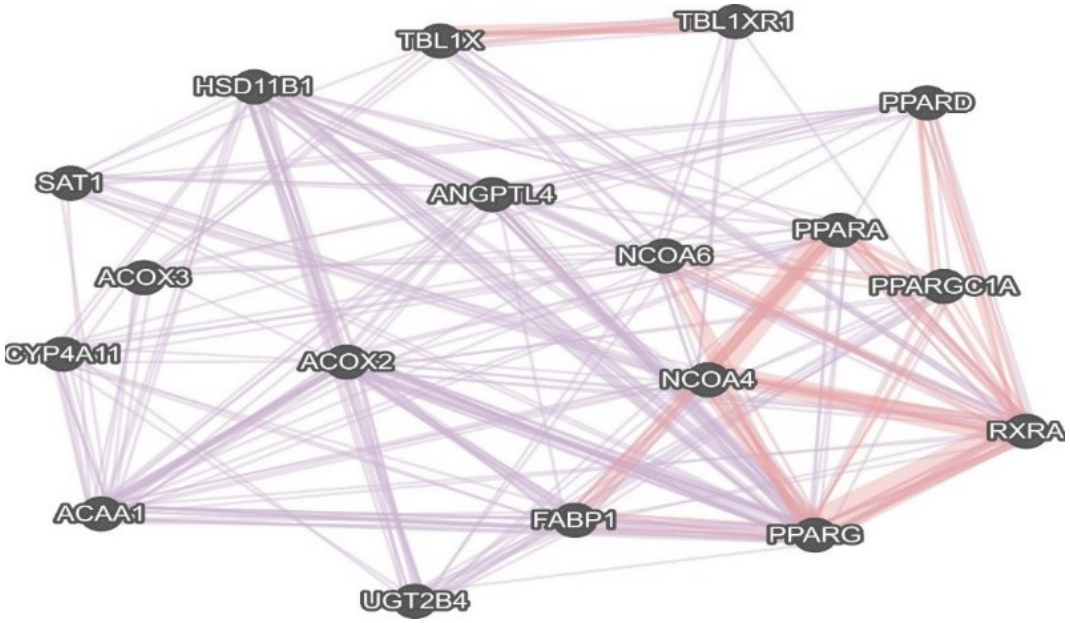
H3K4me3



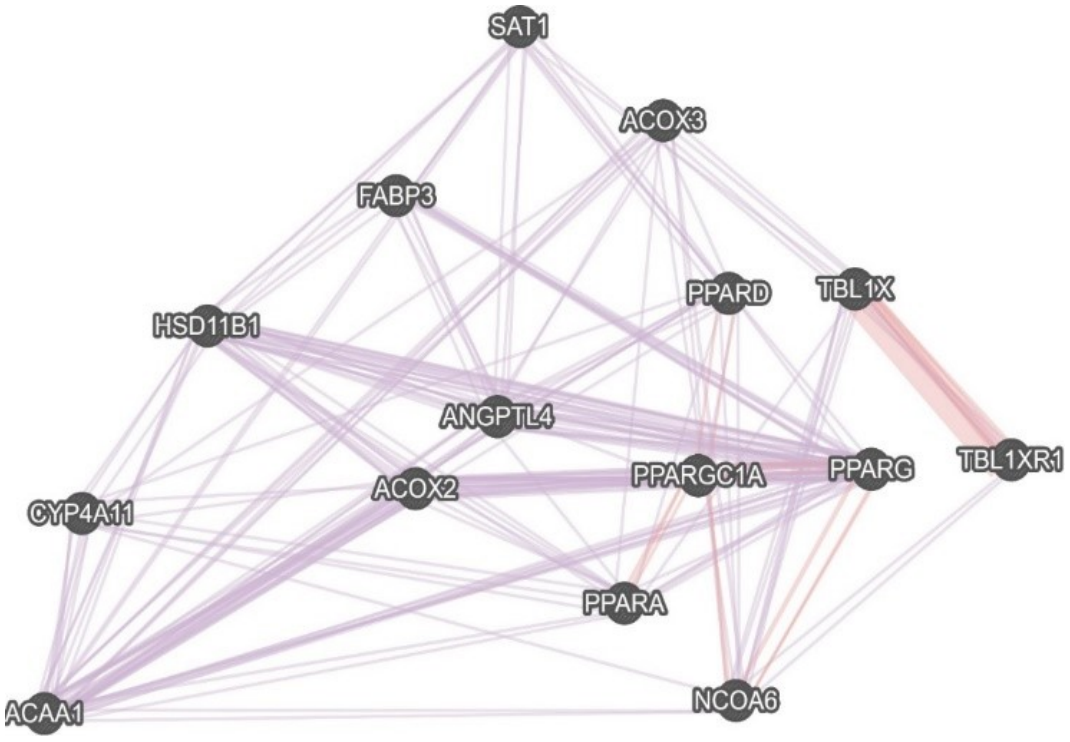
H3K18ac



H4K8ac



H4K5ac



H4K91ac

Figure 3: Physical and co-expression networks for the epigenetic modifications based on distribution of 5-mers abundance in the genes

References

[1] Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K., High-resolution profiling of histone methylations in the human genome. Cell,129(4)(2007) 823-837.

[2] Bastos, C.A., Afreixo, V., Pinho, A.J., Garcia, S.P., Rodrigues, J.M. and Ferreira, P.J., Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions. Journal of integrative bioinformatics, 8(3) (2011) 31-42.

[3] Benveniste, D., Sonntag, H.J., Sanguinetti, G. and Sproul, D., Transcription factor binding predicts histone modifications in human cell lines. Proceedings of the National Academy of Sciences, 111(37 (2014)13367-13372.

[4] Bird, A., DNA methylation patterns and epigenetic memory. GENES and DEVELOPMENT, 16(2002)6-21.

[5] Bliss, C.I. and Fisher, R.A., Fitting the negative binomial distribution to biological data. Biometrics, 9(2) (1953) 176-200.

[6] Chao, L., Rang, C.U. and Wong, L.E., Distribution of spontaneous mutants and inferences about the replication mode of the RNA bacteriophageφ6. Journal of virology, 76(7) (2002) 3276-3281.

[7] Chor, B., Horn, D., Goldman, N., Levy, Y. and Massingham, T., Genomic DNA k-mer spectra: models and modalities. Genome biology, 10(10)(2009)1-10.

[8] Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W., GenBank. Nucleic acids research, 44(D1)(Database issue)(2016) 67-72.

[9] Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G., Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press,(1998).

[10] Larson, J.L. and Yuan, G.C., Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. BMC bioinformatics, 11(1) (2010) 1-12.

[11] Pham, T.H., et al., Prediction of Histone Modifications in DNA sequences. in IEEE 7th International Symposium on BioInformatics and BioEngineering, (2007) 959-966.

[12] Phaml, T.H., Tran, D.H., Ho, T.B., Satou, K. and Valiente, G., Qualitatively predicting acetylation and methylation areas in dna sequences. Genome Informatics,16(2) (2005) 3-11.

[13] Rabindra Kumar Singh , D.M.S., Feature Selection of Gene Expression Data for Cancer Classification: A Review. Procedia Computer Science, 50 (2015) 52-57.

[14] Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. and Sokhansanj, B., Metagenome Fragment Classification Using N-Mer Frequency Profiles. Advances in bioinformatics, 2008(2008).

[15] Salimi, D., moeini, A., Masoudi-Nejad, A., Sequence-based 5-mers highly correlated to epigenetic modifications in genes interactions.Genes and Genomics, 40(12)(2018)1363-1371.

[16] Sega, E., Barash, Y., Simon, I., Friedman, N., Koller, D.,. From Promoter Sequence to Expression: A Probabilistic Framework. in Sixth Annual International Conference on Computational Biology. (2002).

[17] Tran, D.H., Pham, T.H., Satou, K. and Ho, T.B., Conditional random fields for predicting and analyzing histone occupancy, acetylation and methylation areas in DNA sequences. In Workshops on Applications of Evolutionary Computation, (2006) 221-230.

[18] Vu, T.N., Wills, Q.F., Kalari, K.R., Niu, N., Wang, L., Rantalainen, M. and Pawitan, Y., Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics, 32(14) (2016) 2128-2135.

[19] Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q. and Zhao, K., Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics, 40(7) (2008) 897-903.

[20] Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T. and Maitland, A., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic acids research, 38 (2010) 214-220.

[21] Williams, D., Trimble, W.L., Shilts, M., Meyer, F. and Ochman, H., Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC genomics, 14(1) (2013) 1-11.