Clustering Some MicroRNAs Expressed in the Breast Tissue Using Shannon Information Theory and Comparing the Results With UPGMA, Neighbor-joining, and Maximum-likelihood Methods

Arezou Askari Rad[1], **Jamal Fayazi**[1*], **Houshang Dehghanzadeh**[2]

1. Department of Animal Science, Faculty of Animal Science and Food Industry, Agricultural Sciences and Natural Resources University of Khuzestan, Ahvaz, Iran.
2. Department of Animal Science Research, Guilan Agricultural and Natural Resources Research and Education Center, (AREEO), Rasht, Iran.

Use your device to scan and read the article online

Citation Askari Rad A, Fayazi J, Dehghanzadeh H. Clustering Some MicroRNAs Expressed in the Breast Tissue Using Shannon Information Theory and Comparing the Results With UPGMA, Neighbor-joining, and Maximum-likelihood Methods. Research in Molecular Medicine. 2020; 8(4):179-188. https://doi.org/10.32598/rmm.8.4.3

https://doi.org/10.32598/rmm.8.4.3

Article Type:
Research Paper

Article info:
Received: 24 Jul 2020
Revised: 25 Aug 2020
Accepted: 22 Sep 2020

Keywords:
Information theory, Kullback-Leibler divergence, MicroRNA, Clustering, Entrop

ABSTRACT

Background: Because milk and milk products play a vital role in human nutrition, dairy cattle farmers are working in increasing milk production or changing its composition. For this reason, researching the genes which play an important role in milk production and its composition is of high value. Information theory is an interdisciplinary branch of mathematics which overlaps with communications engineering, biology, and medicine. It has been used in genetic and bioinformatics analyses such as the biological structures and sequences.

Materials and Methods: In this study, a total of 20 microRNAs from those affecting the breast tissue and mammary glands have been extracted from the microRNA database. For each microRNA sequence, the entropy values of the first- to third-order were calculated and the Kullback-Leibler divergence criteria were estimated. Then, the Kullback-Leibler divergence matrix of the microRNAs was considered as the inputs for clustering methods. All calculations were performed in the R program. The biological pathway of each target was predicted using the KEGG server.

Results: MicroRNAs are divided into two main groups based upon comparing and analyzing all the created clusters. The first group contains 18 microRNA and the second group contains 2 microRNAs at the first- and third-order entropies. The second-order entropy contains 19 microRNA in the first group and only 1 microRNA in the second group. The clustering topology changes as the entropy order changes from 1 to 3, with the most significant changes being seen in the clustering resulted from the third-order entropy.

Conclusion: In the proposed method of clustering, we obtained a biological grouping of genes. There is a good concordance between most of the microRNAs within one cluster and their biological pathway. The algorithm is applicable for clustering a range of genes and even genomes based on their DNA sequences entropy. Our method can help assign and predict the biological activity of those genes that lack robust annotations because it relies only on the DNA sequence and length of the genes.

*** Corresponding Author:**
Jamal Fayazi, PhD.
***Address:** Department of Animal Science, Faculty of Animal Science and Food Industry, Agricultural Sciences and Natural Resources University of Khuzestan, Ahvaz, Iran.*
***Phone:** +98 (916) 6124162*
***E-mail:** j_fayazi@asnrukh.ac.ir*

Introduction

Increasing milk production or changing its composition has attracted the attention of dairy cattle breeders. For this aim, investigating the genes affecting milk production and its composition can be a key step towards identifying and developing marker-assisted selection and developing breeding programs to improve productive traits [1, 2].

MicroRNAs are a group of short, non-coding single-stranded conserved RNAs (between 19 and 23 nucleotides) that act as post-transcriptional regulators for the gene expression in a wide range of animals, plants, and viruses [3, 4]. MicroRNAs are present in the cells as well as in exosomes found in biological fluids such as milk. Most investigated microRNAs in the mammary tissue and glands have immune-related functions and different patterns have been observed in their expression during the lactation period [5]. MicroRNAs can regulate milk production from the mammary glands through epigenetic mechanisms. They can also, directly and indirectly, control the activity of agents involved in the epigenetics through inhibiting the transcription initiation, intermediate microRNA degradation, target genes, as well as inhibiting the positive regulation and methylation adjustment of the target genes [6-8].

The information theory, a branch of mathematics which overlaps with communications engineering, biology, and medicine, plays a key role in this regard. This theory, proposed by Claude Shannon in 1948, explores the mathematical laws governing the data behavior during the transfer, storage, and retrieval processes. Shannon entropy is at the core of the information theory and it is sometimes known under topics such as the degree of uncertainty or randomness, disorganization, and unpredictability. Information is the scale of uncertainty or entropy in a situation, so a system with more entropy has more information. When a situation is completely predictable, there is no information about it. This condition is called solidity (negentropy) [9]. The unit of entropy is bit, and a system's entropy is related to the amount of available information in it. A system with a higher order can be described by fewer bits of information, while a system with a lower order requires more bits [10].

Information theory has been used as an important and versatile tool to search for patterns in the DNA sequences [11], the role of amino acids in the protein structures of the yeast [12], analysis of the quantitative traits and epistasis, investigation of the global genome information, analysis of the DNA microarray data, classification of the genes involved in cancer [13], comparison of the complexity amount for the DNA sequence analyses and phylogenetic trees reconstruction without aligning bases [14], evolutionary research [15], genetic diversity [16], comparison of the information contents of the genes' intron and exon areas [17], and analysis of the genome's CpG islands [18].

In the probability and information theories, the Kullback- Leibler divergence or relative entropy is an asymmetric criterion for measuring the difference between the two probabilistic distributions of Q and P. This divergence is a member of a wider class of divergences called F divergence [6]. This divergence was first introduced by Solomon Kullback and Richard Leibler as an oriented divergence between two distributions [19].

So far many microRNAs have been discovered, and categorized into biological groups. A better understanding of their features is needed to organize information for the upgrading. Clustering of microRNA sequences can be a confirmation of the microRNAs belonging to one family and also lead to the discovery of new sequences. Given that the microRNAs located in a cluster act on the same biological pathway, a well-organized clustering can lead to the discovery of new microRNAs which have not been categorized yet [20, 21].

The present paper illustrates the application of the Kullback-Leibler divergence-based algorithm presented for clustering several microRNAs affecting milk production for the first time. According to the authors' knowledge, no research has so far clustered the effective microRNAs on the mammary tissue and glands using the information theory. It is expected that the extraction of gene patterns from this clustering be used in the biological, pharmaceutical, and breeding research.

Materials and Methods

Extraction of microRNAs

In this study, a total of 20 microRNAs from those affecting the breast tissue and mammary glands of the dairy cattle have been extracted from the microRNA database (http://www.mirbase.org/) and examined as well. The microRNA precursor sequence has been implemented here. The precursor sequence has about 60-90 base pairs with a pin-like structure and is often preserved evolutionarily [22]. The sequences were saved after extraction as FASTA format. R programming language was

Table 1. The calculated first to third-order entropy values of the effective microRNAs on the breast tissue of the dairy cattle

No.	MicroRNA Symbol	Order of entropy‡		
		$H(x)_i$ First-Order	$H(x)_{ii}$ Second-Order	$H(x)_{iii}$ Third-Order
1	bta-mir-10[a]	1.9531	3.8298	5.3723
2	bta-mir-15[b]	1.9308	3.8032	5.4428
3	bta-mir-16[b]	1.9529	3.7931	5.2495
4	bta-mir-21	1.9785	3.7731	5.0861
5	bta-mir-23[a]	1.9696	3.7072	5.0224
6	bta-mir-24-2	1.9924	3.783	5.0632
7	bta-mir-33b*	1.836	3.434	4.7147
8	bta-mir-125b-2	1.9917	3.8291	5.3475
9	bta-mir-141	1.9806	3.8482	5.2897
10	bta-mir-145	1.9805	3.8327	5.3751
11	bta-mir-146b**	1.9967	3.8605	5.5111
12	bta-mir-155	1.9577	3.8774	5.0739
13	bta-mir-181a-1***	1.9893	3.9126	5.5626
14	bta-mir-199b	1.9815	3.8571	5.4279
15	bta-mir-205	1.986	3.7185	5.0804
16	bta-mir-221	1.9906	3.8781	5.3856
17	bta-mir-223	1.9955	3.7793	5.3077
18	bta-mir-484	1.9442	3.6305	4.7525
19	bta-mir-486	1.8577	3.6478	5.221
20	bta-mir-500	1.9891	3.731	5.046

‡ Maximum criteria for amounts in the first-, second-, and third-order entropy are 2, 4, and 6, respectively;

* Minimum microRNA entropy in the first-, second-, and third-order entropy;

** Maximum microRNA entropy in the first-order entropy;

*** Maximum microRNA entropy in the second- and third-order entropy.

used to extract the sequence features and computational algorithms.

Calculation of the entropy orders

In this study, for each microRNA sequence, the entropy parameters were calculated from the first to the third orders. In this regard, we used the Markov chain up to the third degree. The following formula is used to calculate the first-order entropy (zeroth-order Markov chain) (Formula 1):

1. $H(x)_i = -\sum p_i \, log_2 \, p_i$

in which, pi is the probability of the ith nucleotide from the set {A, U, G, C} in the microRNA chain. This type of entropy assumes that the appearance of each nucleotide is independent of the other one in the strand and does not depend on the type of the adjacent nucleotide.

The second-order entropy (first-order Markov chain) is given by the following relation (Formula 2):

2. $H(x)_{ii}=-\sum p_i \sum p_i \, (j) \, log_2 \, p_i(j)$

Table 2. The Kullback-Leibler (relative entropy at orders one to three) results associated with the microRNAs affecting the breast tissue of dairy cattle

No.	KL_H	MicroRNA	Amount‡
H(X)1	Min Distance	bta-mir-145 bta-mir-141	4.8822e.-005
	Max Distance	bta-mir-146b bta-mir-33b	0.32154
H(X)2	Min Distance	bta-mir-125b-2 bta-mir-10a	0.001417
	Max Distance	bta-mir-181a-1 bta-mir-33b	0.95847
H(X)3	Min Distance	bta-mir-10a bta-mir-145	0.0055686
	Max Distance	bta-mir-181a-1 bta-mir-33b	1.6997

‡ For microRNAs pair, a value closer to zero indicate more similarity of microRNAs.

where i indicates the occurrence of the previous nucleotide and pi(j) denotes the occurrence probability of the jth nucleotide upon the occurrence of ith one from the set {A, U, C, G} in the microRNA chain.

The third-order entropy (second-order Markov chain) is calculated using the Formula 3:

$$3.\ H(x)_{iii}=-\sum p_i \sum p_i\,(j)\sum p_i,j(k)\sum p_i,j,k(m)\ log_2\,p_i,j,k(m)$$

Here, i, j, and k indicate the awareness of the three preceding nucleotides occurrence, and pi,j,k(m) stands for the probability of the mth nucleotide upon the occurrence of i, j, and kth ones from the set {A, U, C, G} in the microRNA chain. As mentioned above, the entropy values have been calculated at three orders. The index that appears in H represents the order of entropy [23, 24].

Measurement of the Kullback-Leibler divergence

The Kullback-Leibler divergence can be estimated using the Formula 4 [23, 24]:

$$4.\ DKL(P_{(X)}\|Q_{(X)})=\sum^{n}_{i}=1P^{(X)}log_2\,P_{(X)}/Q_{(X)}$$

, where n is the number of nucleotides in a microRNA strand and DKL is neither a symmetric criterion nor a real distance. This method, referred to as KL(H) for ease of use within the context, is based on the entropy values of microRNAs. The entropy values of the two microRNAs were calculated separately where the numerical values of the first and second sequences were placed as P and Q in the formula, respectively. In this method, an asymmetric matrix was created with the size of the number of microRNAs based on which the sequences with the highest similarity and distance were identified.

Different types of clustering methods

The sequences clustering has been accomplished with four common clustering methods: the single linkage, UPGMA, neighbor-Joining, and maximum likelihood.

Results

Table 1 presents the microRNAs information. Investigating the microRNAs characteristics indicated that the highest and lowest sequence lengths are associated with the bta-mir-486 and bta-mir-484, respectively. The lowest first-order entropy is dedicated to bta-mir-33b, which seems to be due to its short length. In return, bta-mir-146b and bta-mir-181a-1 meet higher entropy values due to their higher lengths. However, this is not always the case, because the shortest microRNA which has been examined is bta-mir-484 with 63 bp, while its entropy is not the lowest. The highest microRNA length corresponds to bta-mir-486, which is 123 bp, while its entropy is not the highest at any of the orders.

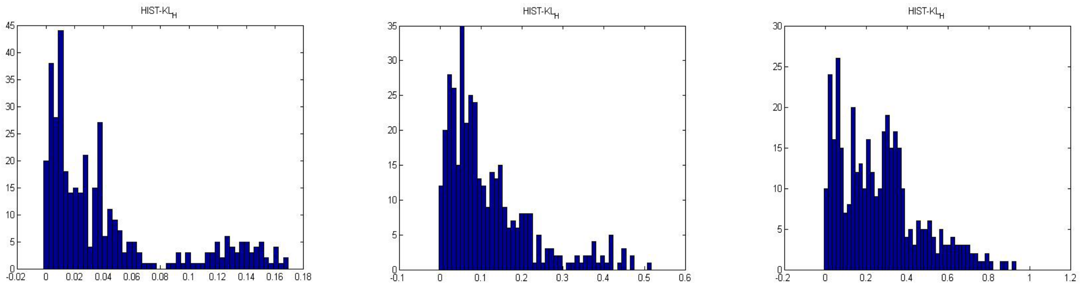

Figure 1. Frequency histogram of the Kullback-Leibler values based on the first- (left side) to third-order (right side) entropy values. The X-axis is the Kullback-Leibler divergence value and Y-axis is the number of microRNA pair comparisons

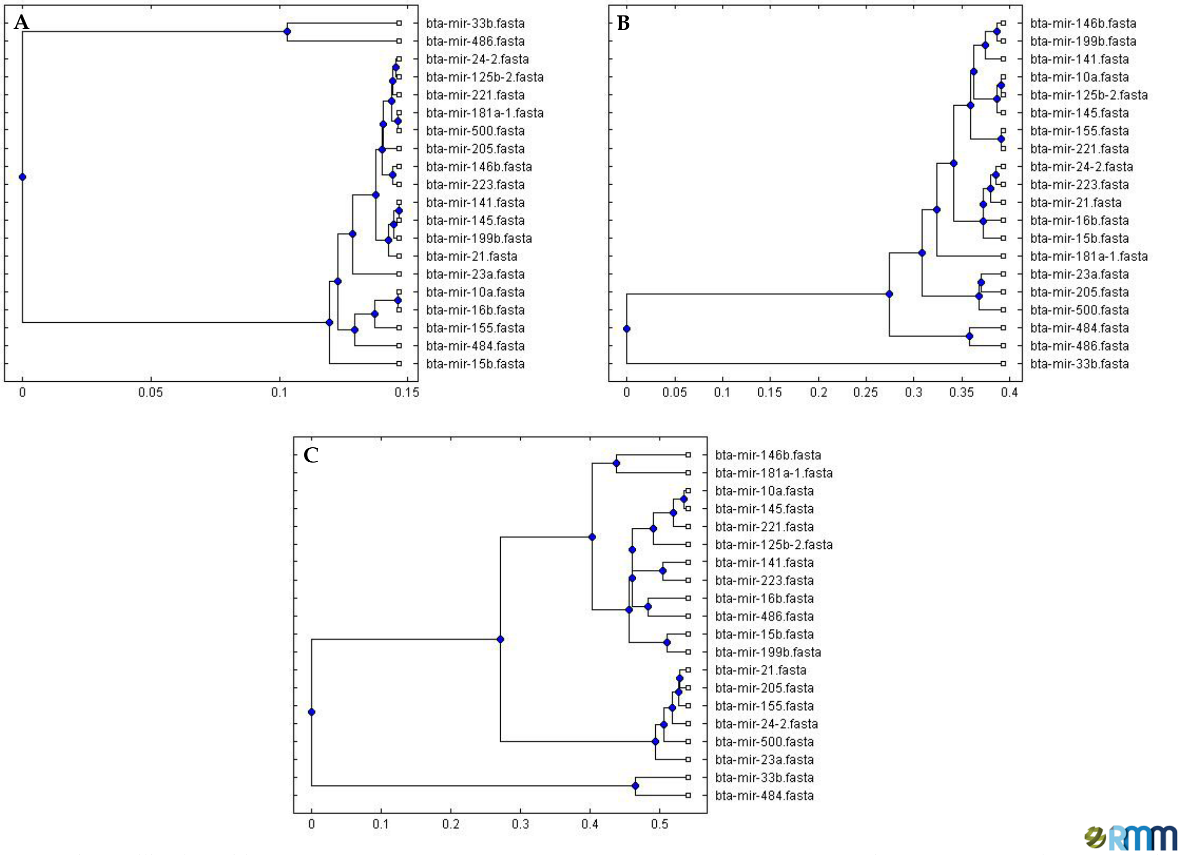

Figure 2. The Kullback-Leibler clustering results based on the first- (A) to third-order (C) relative entropy values using the single linkage approach

Investigating the genes and exons affecting milk production indicated that *NOP2, YWHAH* genes (with lengths of 60167 and 1445, respectively) and exon 1 *of HSP* and *ACTR2* genes which are the largest and smallest investigated genes and exons in this study meet the maximum and minimum entropy values, respectively [23, 24]. Also, a study on the global genome information analyzed the genome's entropy results of 25 different species according to the information theory and plotted a two-dimensional graph based on the chromosome length and entropy values for the genome of each specie. It was observed that the entropy values are higher in some species with longer chromosome lengths, while this was not the case in some others [25]. This result is similar to our finding, in which bta-mir-146b and bta-mir-181a-1 had the highest entropy value while the bta-mir-486 was the longest sequence. On the other hand, bta-mir-484 was the shortest sequence but did not have the least entropy value.

Genes clustering using the Kullback-Leibler divergence

First, the entropy was calculated separately at orders one to three, and the Kullback-Leibler values were calculated for each order of entropy and each microRNA. Because alignment is not necessary for this method, sequences are allowed to be evaluated with their actual content. The results of this section are presented in Table 2. Then, the microRNA clustering steps were performed based on their relative entropy and without applying the alignment. The frequency histogram of the microRNAs Kullback-Leibler results is reported as below based on the first-to-third order entropy values.

As seen in Figure 1, based on the second- and third-order entropies, about 3% (12 observations) and 2.5% (10 observations) of the estimates which are quite similar with values close to zero are placed at the second and third orders, respectively.

Discussion

MicroRNAs are divided into two main groups upon comparing and analyzing all the created clusters (Figure 2). The first group contains 18 microRNA and the second group contains 2 microRNAs at the first- and third-order and their entropies. The second-order entropy contains 19 microRNA in the first group and only 1 microRNA in

Table 3. Targets and biological pathways corresponding to the microRNAs located in a cluster, and microRNAs in comparison with a pairwise manner

microRNA	Target	Pathway
Bta-mir-146[b]	ZNF148, ZNF532	Thermogenesis
Bta-mir-181a-1	ZNF704, ZFAND5	Thermogenesis
Bta-mir-10[a]	TMEM170B	RIG-I-like receptor signaling pathway
Bta-mir-145	TMEM65	Phosphatidylinositol signaling system
Bta-mir-221	CDKN1B	Pathway in cancer
Bta-mir-125b-2	CDK6	Pathway in cancer
Bta-mir-141	SOX17	Signaling pathway
bta-mir-223	SOX6	Signaling pathway
Bta-mir-16[b]	ATP1B4	Mineral absorption
Bta-mir-486	ATP7A	Mineral absorption
bta-mir-15[b]	ATP1B4	Mineral absorption and pancreatic secretion
Bta-mir-199[b]	ATP1B3	Mineral absorption and pancreatic secretion
Bta-mir-21	SLC30A10 ZBTB47	- -
Bta-mir-205	ZBTB47 SLC35B3	- -
Bta-mir-155	NFI/A	
Bta-mir-24-2	NFAT5	-
Bta-mir-500	ZNF-277 SLC12A1	- -
Bta-mir-23[a]	ZNF-299 SLC4A4	- -
Bta-mir-33b	SNAI3	Hippo signaling pathway
Bta-mir-484	SNAI2	Hippo signaling pathway

the second group. The bta-mir-33b and bta-mir-486 microRNAs are located in the same cluster at the first-order entropy, while they are in different clusters at the second- and third-order ones. At the first- and second-order entropies, bta-mir-141 and bta-mir-199b are in the same cluster while they are located in two separate clusters at the third-order one. In general, the clustering topology changes as the entropy order changes from 1 to 3, with the most significant changes being seen in the clustering resulted from the third-order entropy. The clustering was found to be independent of the length and strongly correlated with microRNA content and frequency of their nucleotides in the chain. Investigation of the results indicated that the relative entropy-based clustering is a correct, logical, and fast method.

Ghaderi et al. (2016) implemented a simple form of this criterion to find the distance between the contigs of the Escherichia coli genome which affects mastitis [26]. Also, Dehghanzadeh et al. used it to cluster several genes in dairy cattle [23]. Porto-Diaz et al. (2012), who employed the information theory for clustering 12 microarray data, indicated that this is a fast and reliable way of clustering large volumes of data and considered it better compared to other methods. This is mainly because this method lacks the disadvantages of aligning sequences, it examines their actual contents and forms in practice and does not require high computational memory for the sequences with higher lengths. Sequence data sequencing is increasing. Also, as the information content of the sequence is not lost, the accuracy of the sequence data clustering increases.

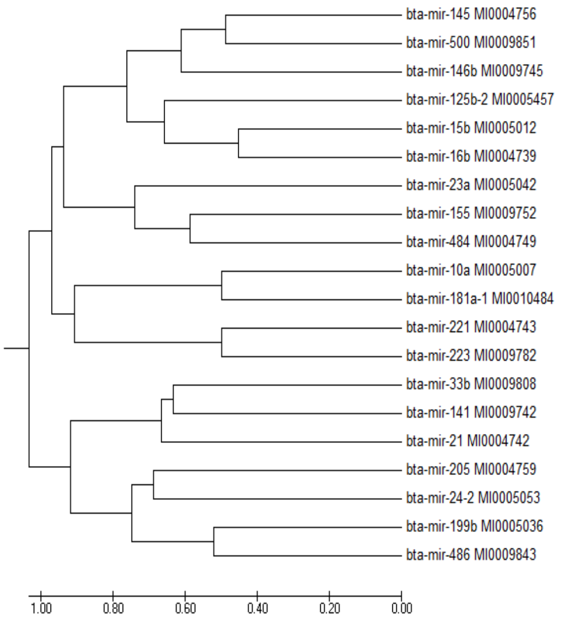

Figure 3. MicroRNA clustering with UPGMA method

One of the most important pre-processing in this clustering method is to reduce the feature space dimensions to improve the performance of the classification system. Reducing the dimensions of the feature space leads to a reduction in the complexity of the classification process and thus reduces the occurrence of errors. This clustering method is based on information theory. The measurement criteria based on it such as Kullback-Leibler and mutual information provide various features of the sequence relationships. These criteria possess the maximum amount of information about the output which enhances the classification accuracy.

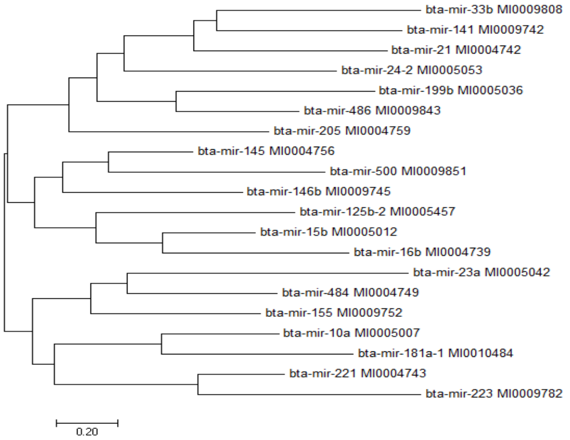

Figure 5. MicroRNA clustering with the neighbor-joining method

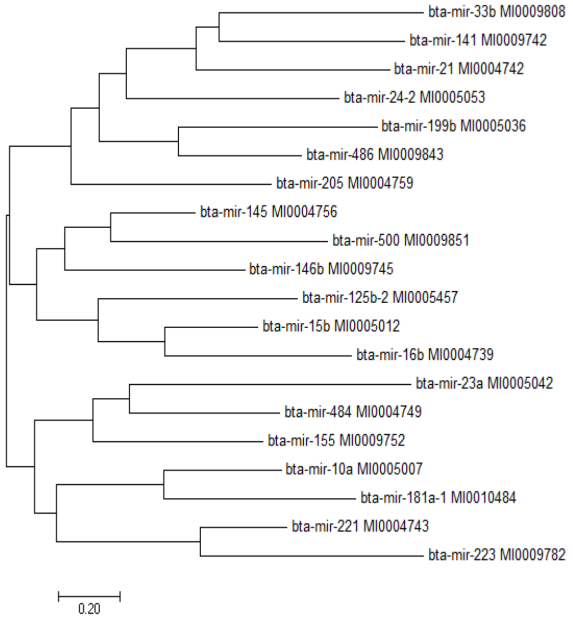

Figure 4. MicroRNA clustering with the maximum-likelihood method

A comparison of the sequences' entropy results with the entropy-based Kullback-Leibler criterion points out to the direct relationship between the Kullback-Leibler calculation formula and the entropy values. Concerning this relationship, microRNAs with the lowest and highest entropy values at the orders one to three are the same as those with the highest gene spacing in the entropy-based Kullback-Leibler. Also, if the entropies of the two sequences are similar, then their Kullback-Leibler distance will be zero, and those sequences which possess such property or their Kullback-Leibler distance is close to zero (especially at higher-order entropy), their corresponding microRNAs are reportedly expected to have common biological roles as they are located in the same cluster during the clustering process [18, 21, 27].

The distance matrix created in the above-mentioned methods can be the input of unsupervised algorithms such as hierarchical clustering. In this case, fragments showing themselves as clusters will be easily recognizable. Accordingly, fragments located within a cluster are likely to act in a common biological pathway. Various studies have been carried out in this regard. Several research studies have been performed to take advantage of metabolic data to understand better the evolutionary relationship of the different species [28] with respect to the accumulation of metabolic data and using graph theory. These studies indicated that the created phylogenetic tree is consistent with the laboratory results [29]. To confirm the accuracy of this clustering method, the biological pathway of each target was predicted using the KEGG server after determining the targets corresponding to each of the microRNAs using the mirBASE one. The microRNAs located in a cluster with the shortest distance were compared in a pairwise manner and then the biological pathway of the targets was predicted by the KEGG server. The microRNAs located in a cluster with the shortest distance were compared in a pairwise manner and then the biological pathway of the targets was predicted by the KEGG server (Table 3). This method was true for microRNAs in a cluster and these microRNAs had the same biological pathways, indicating the accuracy of this method.

These results confirm the capability of the proposed method for clustering. To further investigate the accuracy of this technique, microRNAs have been clustered using the UPGMA, neighbor-Joining, and maximum-likelihood methods by MEGA-7 (Figures 3-5). Similar to the third-order entropy, bta-mir-24 and bta-mir-205 are located in close clusters in the UPGMA method. Also, similar to the third-order entropy, bta-mir-199, bta-mir-486, bta-mir-10a, and bta-mir-181a-1 are close to each other according to the maximum-likelihood method. Furthermore, in the neighbor-joining method, bta-mir-181a-1 and bta-mir-10a, and also bta-mir-21 and bta-mir-24-2 are located close to each other. Besides, the second-order entropy has been compared with each of the above four methods. A comparison of the second-order entropy with the UPGMA method indicated that bta-mir-16b and bta-mir-15b are in the same cluster.

Similar to the case seen at the second-order entropy, bta-mir-24-2 and bta-mir-21are in the same cluster according to the maximum-likelihood approach. Further to these, bta-mir-16b and bta-mir-15b, as well as bta-mir-21 and bta-mir-24-2 are located in the same clusters both in the neighbor-joining method and second-order entropy. Also, bta-mir-21 and bta-mir-141 are located in the same cluster both in the UPGMA method and first-order entropy. It is also seen that bta-mir-155 and bta-mir-484, as well as bta-mir-500 and bta-mir-146 are located in the same clusters both in the maximum-likelihood method and first-order entropy. The results illustrate that none of these three methods fully comply with the entropy clustering. Comparing the three methods points out that the maximum-likelihood and UPGMA methods have clustered the microRNAs in the same way. The similarity of the maximum-likelihood and neighbor-joining methods is greater than that of the two previous methods and even the cluster positions are similar. Given that the entropy method performs clustering based on the sequence contents while the three mentioned methods perform it based on the sequence length, different results are expected.

As a conclusion, in the proposed method of clustering, we obtained a biological grouping of genes. This new and innovative method can be used in clustering other genes due to the extraction of features obtained from clustering results. The algorithm is applicable for clustering a range of genes and even genomes based on their DNA sequences' entropy. The proposed algorithm takes an alignment-free approach, and it is a Kullback-Leibler divergence approach based on gene entropy that can cluster the genes.

The novelty of this algorithm is its ability to support and cluster sequences of different lengths using information theory and relative entropy. The method presented in this paper can help assign and predict the biological activity of those genes that lack robust annotations because it relies only on the DNA sequence of the genes and the size and length of the genes' effect on nature.

Ethical Considerations

Compliance with ethical guidelines

All ethical principles are considered in this article. The participants were informed of the purpose of the research and its implementation stages.

Funding

This study was financially supported by Agricultural Sciences and Natural Resources University of Khuzestan.

Authors' contributions

Conceptualization, Methodology, Writing – original draft, and Writing – review & editing: All authors; Investigation: Arezo Askari, Houshang Dehghanzade; Data collection and Data analysis: Arezo Askari, Jamal Fayazi; Supervision: Jamal Fayazi.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgments

The authors thank all the reviewers who elevated this study with their comments and the Research Deputy of Agricultural Sciences and Natural Resources University of Khuzestan for their support.

Refrences

[1] Alinaghizadeh R, Mohammadabadi M, Zakizadeh S. Exon 2 of BMP15 gene polymorphism in Jabal Barez Red Goat. Agric Biotechnol J. 2010; 2(1):69-80. https://jab.uk.ac.ir/m/article_362.html?lang=en

[2] Mohammadabadi MR, Nikbakhti M, Mirzaee HR, Shandi A, Saghi DA, Romanov MN, et al. Genetic variability in three native Iranian chicken populations of the Khorasan province based on microsatellite markers. Russ J Genet. 2010; 46(4):505-9.[DOI:10.1134/S1022795410040198]

[3] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116(2):281-97. [DOI:10.1016/S0092-8674(04)00045-5]

[4] Gholizade M, Fayazi J. Prediction of MicroRNAs bind to Toll-like Receptors Pathway in Chicken based on Bioinformatics Method. Res Mol Med. 2019; 7(4):33-42. [DOI:10.32598/rmm.7.4.33]

[5] Gigli I, Maizon DO. microRNAs and the mammary gland: A new understanding of gene expression. Genet Mol Biol. 2013; 36(4):465-74.[DOI:10.1590/S1415-47572013005000040] [PMID] [PMCID]

[6] Li C, Wang J. Relative entropy of DNA and its application. Phys A: Stat Mech Appl. 2005; 347:465-71. [DOI:10.1016/j.physa.2004.08.041]

[7] Li Z, Liu H, Jin X, Lo L, Liu J. Expression profiles of microRNAs from lactating and non-lactating bovine mammary glands and identification of miRNA related to lactation. BMC Genomics. 2012; 13(1):731. [DOI:10.1186/1471-2164-13-731] [PMID] [PMCID]

[8] Javdani H, Parsamanesh N. MicroRNA based Novel Strategies for Cancer Treatment. Res Mol Med (RMM). 2018:5-15. [DOI:10.18502/rmm.v6i1.3933]

[9] Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948; 27:379-423. [DOI:10.1002/j.1538-7305.1948.tb01338.x]

[10] Gray RM. Entropy and Information Theory. First Edition. New York: Springer-Verlag New York Publisher; 2013.

[11] Vinga S, Almeida J. Alignment-free sequence comparison—a review. Bioinformatics. 2003; 19(4):513-23. [DOI:10.1093/bioinformatics/btg005] [PMID]

[12] Kim J, Kim S, Lee K, Kwon Y. Entropy analysis in yeast DNA. Chaos, Solitons & Fractals. 2009; 39(4):1565-71. [DOI:10.1016/j.chaos.2007.06.036]

[13] Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A, Fontenla-Romero O. A study of performance on microarray data sets for a classifier based on information theoretic learning. Neural Netw. 2011; 24(8):888-96. [DOI:10.1016/j.neunet.2011.05.010] [PMID]

[14] Pham TD, Crane DI, Tannock D, Beck D. Kullback-leibler dissimilarity of markov models for phylogenetic tree reconstruction. In proceedings of 2004 international symposium on intelligent multimedia, video and speech processing, 2004. 2004 Oct 20. (pp. 157-160). IEEE. [DOI: 10.1109/ISIMP.2004.1434024]

[15] Erill I. Information theory and biological sequences: insights from an evolutionary perspective. Information theory: New research. New York: Nova Science Publishers. 2012:1-28. http://digitalmeasures.umbc.edu/dmeasures/ye55084/intellcont/978-1-62100-325-0_ch1_noPW-1.pdf

[16] Sherwin WB. Entropy and information approaches to genetic diversity and its expression: Genomic geography. Entropy. 2010; 12(7):1765-98. [DOI:10.3390/e12071765]

[17] Zamani P, Akhondi M, Mohammadabadi MR, Saki AA, Ershadi A, Banabazi MH, Abdolmohammadi AR. Genetic variation of Mehraban sheep using two intersimple sequence repeat (ISSR) markers. African J Biotechnol. 2011; 10(10):1812-7. https://www.ajol.info/index.php/ajb/article/view/93089

[18] Barazandeh A, Mohammadabadi MR, Ghaderi-Zefrehei M, Nezamabadi-Pour H. Genome-wide analysis of CpG islands in some livestock genomes and their relationship with genomic features. Czech Journal of Animal Science. 2016; 61(11):487-95. [DOI:10.17221/78/2015-CJAS]

[19] Kullback S, Leibler R. On information and sufficiency. the annals of mathematical statistics. Annals Math Stat. 2006:79-86. [DOI:10.1214/aoms/1177729694]

[20] Zambelli F, Mastropasqua F, Picardi E, D'Erchia AM, Pesole G, Pavesi G. RNentropy: An entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. Nucleic Acids Res. 2018; 46(8):e46-. [DOI:10.1093/nar/gky055] [PMID] [PMCID]

[21] Kasahara VA, do Carmo Nicoletti M. Graph-based clustering of miRNA sequences. MicroRNA. 2017; 6(3):166-86. [DOI:10.2174/2211536606666170724154752] [PMID]

[22] Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ. miRBase: Tools for microRNA genomics. Nucleic Acids Res . 2007; 36(suppl_1):D154-8. [DOI:10.1093/nar/gkm952] [PMID] [PMCID]

[23] Dehghanzadeh H, Ghaderi-Zefrehei M, Mirhoseini SZ, Esmaeilkhaniyan S, Haruna IL, Najafabadi HA. A new DNA sequence entropy-based Kullback-Leibler algorithm for gene clustering.J Appl Genet. 2020:1-8. https://link.springer.com/article/10.1007/s13353-020-00543-x

[24] Neagoe IM, Popescu D, Niculescu VI. Applications of entropic divergence measures for DNA segmentation into high variable regions of cryptosporidium spp. gp60 gene. Rom Rep Phys. 2014; 66(4):1078-87. http://www.rrp.infim.ro/2014_66_4/A15.pdf

[25] Tenreiro Machado JA. Shannon entropy analysis of the genome code. Math Probl Eng. 2012; 2012. [DOI:10.1155/2012/132625]

[26] Ghaderi-Zefrehei MA, Bandi Dastjerdi A, Bahreini Behzadi MR, Samadian F, Meamar M. [Investigation of information accumulation in *Escherichia Coli's* DNA sequence affecting mastitis in dairy cow using information theory (Persian)]. J Rumin Res. 2016; 4:16-29. [10.22069/EJRR.2016.3225]

[27] Yang Y, Huang N, Hao L, Kong W. A clustering-based approach for efficient identification of microRNA combinatorial biomarkers. BMC Genomics. 2017; 18(2):210. [DOI:10.1186/s12864-017-3498-8] [PMID] [PMCID]

[28] Clemente JC, Satou K, Valiente G. Phylogenetic reconstruction from non-genomic data. Bioinformatics. 2007; 23(2):e110-5. [DOI:10.1093/bioinformatics/btl307] [PMID]

[29] Heymans M, Singh AK. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics. 2003; 19(suppl_1):i138-46. [DOI:10.1093/bioinformatics/btg1018] [PMID]