Intelligent Classification and Analysis of Essential Genes using Quantitative Methods*

RANJEET KUMAR ROUT, Computer Science & Engineering, National Institute of Technology, Srinagar
SK. SARIF HASSAN, Mathematics, Pingla Thana Mahavidyalaya, Vidyasagar University Maligram
SANCHIT SINDHWANI, National Institute of Technology, Jalandhar-144001, Punjab, India
HARI MOHAN PANDEY, Computer Science, Edge Hill University.
SAIYED UMER, Computer Science & Engineering, Aliah University.

Essential genes are considered to be the necessary genes that are required to sustain life of different organisms. These genes encode proteins which maintain central metabolism, DNA replications, translation of Genes, basic cellular structure and mediate transport process within and out of the cell. The identification of essential genes is one of the essential problems in computational genomics. In this present study, in order to discriminate essential genes from the other genes from the non-biologists perspective, the purine and pyrimidine (Pu-Py) distribution over the essential genes of four exemplary species namely: Homo Sapiens (HS), Arabidopsis Thaliana (AT), Drosophila Melanogaster (DM) and DanioRerio (DR) are thoroughly experimented using some quantitative methods. Moreover, the Indigent classification method has also been deployed for classification on the essential genes of the said species. Based on Shannon entropy (SE), Fractal dimension (FD), Hurst exponent (HE), purine and pyrimidine (Pu-Py) bases distribution the ten different clusters have been generated for the essential genes of the four species. Some proximity results are also reported herewith the clusters of the essential genes.

Additional Key Words and Phrases: Essential genes, Fractal Dimension, Purines, Pyrimidines, Shannon Entropy, Hurst exponent

ACM Reference format:
Ranjeet Kumar Rout, Sk. Sarif Hassan, Sanchit Sindhwani, Hari Mohan Pandey, and Saiyed Umer. 2018. Intelligent Classification and Analysis of Essential Genes using Quantitative Methods*. 1, 1, Article 1 (January 2018), 22 pages.
DOI: 10.1145/1122445.1122456

1 INTRODUCTION

The Central Processing unit of a cell contains the genome. The instructions that describe the genetic functions of the cell are found in the genome sequences which is the blueprint of the whole body. It determines the important characteristics like the cellular chemistry, structure, replication and more. Genomes have been encoded information for the functions often found in all forms of life, and it is possible that the instructions may be species specific. This is contingent on the cell cytoplasm for its expression. Gene sequences are within the genome that functions by giving rise to a discrete product in turn form the basis of heredity, but the functions performed by the genes have been widely reported to be redundant [1, 2]. Identification of such genes is an eye-catching branch in synthetic molecular biology. Some genes are

significant for survival, while others seem to be not necessary. To segregate genes and identify the core sustaining cellular life factors, essential genes have been presented. A cell would vanish if any of these genes is absent. Thus, Essential genes prediction helps to identify minimal set of genes that are required for the indispensable viability of individual cell types. Recognition and analyzing essential genes helps to understand the origin of life. The processes of evolution facilitate the determination of last universal common ancestor [LUCA] [3, 4]. Thus, essential genes are a group of fundamental genes that required for a specific organism to survive in a specific environment [5]. Many of these are essential only under certain conditions, for example, if amino acid lysine is supplied to a cell, the gene responsible for production of lysine is non-essential, but, in the absence of amino acid supply the gene encoding enzyme for lysine biosynthesis become essential as no protein synthesis is possible without it. In almost every species, the elementary cell activities are controlled by essential genes. These core processes tend to be carried out using the same molecular machinery, in every species let it be a fruit fly, roundworm or tropical fish. Infertility is the most common result of lack of such essential genes.

Such essential genes depend on the environment in which an organism survive. A lot of work has been done for recognizing of essential genes when all the nutrients are available in the organism [6]. The various experimentation have been shown that a bacteria requires $250 - 300$ number of genes which encode proteins to maintain a central metabolism, replicate DNA, convert genes into proteins, maintain the structure of cell that helps in facilitating transport process into and out of the cell. Identification of essential genes helps to find the minimal set of genes that form the basis of life and it also finds the human disease genes [7] with new drug targets. The genome-wide identification of essential genes is valuable for rational drug design [8]. Hu et al. [9] and Roemer et al. [10] have identified the drug targets for Aspergillus fumigatus and Candida albicans respectively based on corresponding essential genes (identified from mouse models). Necessary proteins in pathogenic organisms are used for new antibiotics [11], thus, the development of antibacterial drugs is based on the research on these required proteins. There are three strategies for this purpose, namely: gene knockout [12] [13], gene knock down [14–16] and transposon muta-genesis [11, 17]. These strategies are expensive and time consuming but they are capable for generating the accurate groups of essential genes. The approaches used for the identification and differentiation of essential genes are based on experimental and computational methods [18, 19]. Experimental methods are time consuming and expensive while the different experimental methods may yield different results [20]. Thus, computational way of predicting the essential genes is a better alternative. These methods are able to accurately identify the genes, which are then ranked on the basis of how essential they are. This ranking can assist in other drug discovery experiments focused on testing of the top rank genes. The first approach to identify essential genes was given by Mushegian and Kooninin 1996. Till now many methods have been developed which have ultimately resulted in accumulation of sequence data for the number of organisms and groups of experimentally verified essential genes for the number of organisms [DEG [6] and CEG [21]. This data aids the research in identifying the essential genes, the reveal features incorporated with essentially and develop computational methods for the process of identification of essential genes. This data has been utilized by number of researchers in studying the properties of essential genes. There is a relationship between degree in protein-protein interaction networks and essentiality [22–25]. From other studies the relationship between essentiality and the number of transcription factor binding sites upstream of a gene [26]. The analysis of Flux balance [27] have also been used for the understanding and the identification of essential genes. This method generates hypotheses regarding which essential genes are essential under certain hypothetical conditions. Integration of transpose on mutagenesis and sequencing has facilitated many methods for the identification of essential genes. As not only in bacteria, essential genes in archaea [28] and eukaryotes such as saccharomyces [29], fission

yeast schizosaccharomycespombe [30], AT [31], Mus musculus [32] and Homo sapiens [33, 34] are all identified. For classifying essential gene sequences several features are being considered. Mainly they are categorized into sequence information (protein length, codon composition etc.) [35, 36], network topology [37–40], physicochemical property (e.g., molecular weight and number of moles in amino acids) [41] and many other classes.

In brief, the contributions of this work are summarized as follows:

(1) *Formulation of encoding scheme of protein sequences of four different species:-* In this research an encoding method has been introduced to decode the patterns of the DNA sequences of HS, AT, DM and DR species based on Pu and Py bases distributions. There are nitrogenous bases that make up the two different kinds of nucleotide bases in DNA and RNA. The two-carbon nitrogen ring bases (adenine (A) and guanine (G)) are Pu, while the one-carbon nitrogen ring bases (thymine (T) and cytosine (C)) are Py.
(2) *FD from Indicator Matrix:-* Organization of essential genes into clusters that gives some sense of indispensability based on FD which has been explored. We investigate genes bases for the protein sequences of HS, AT, DM and DR through FD from the Indicator Matrix.
(3) *HE as Auto-correlation Parameter:-* The resonance of essential genes over the protein sequences using quantitative parameter HE is determined and found many of protein sequences have identical auto-correlation even if the nucleotides are different. They are basically compared on the basis of their distribution.
(4) *SE and SE with different code words (SED) size:-* These two parameters are employed among all DNA to explore the closeness of the organization of four species. Considering different patterns of frequency distribution for all proteins of four species are determined. Analyses on different quantitative parameters, an analogy over the set of DNA sequences have been reported.

The rest of this paper is organized as follows. In Section 2, definition of different fundamental parameters and our methods have been discussed with the appropriate description. The experimental results and discussions have been demonstrated with the effectiveness of our method in Section 3. Section 4 concludes this paper with emphasizing the key factors of the entire analysis.

2 PROPOSED METHODS

In this section, four different quantitative parameters have been defined to characterize the essential gene sequences of different species. Based on quantitative parameters (SE, FD, HE, and Pu-Py bases distribution) ten different clusters have be generated for the four species. Here *k-means* [42] clustering algorithm has been employed to obtain ten clusters for each species respect to each quantitative parameters. Some proximity results have been observed among the clusters of all the four species reported in Section 3. For convenience, some of the nomenclatures used in the proposed method, have been shown in Table1.

2.1 FD from Indicator Matrices.

Let $D = \{A, T, C, G\}$ be the set of finite alphabet nucleotides and $S(l)$ be a DNA sequences with the repetition of four characters from D of length l. Here, we convert each DNA sequences into indicator matrices [43, 44]. In literature [45] there are several methods to find out the self organising structure of DNA sequences through indicator matrix defined in Equation 2. Then the indication function for each sequence is defined in Equation 1:

$$F : S(l) \times S(l) \rightarrow \{0, 1\}, and\ S(l) = \{0, 1\} \tag{1}$$

Table 1. List of Nomenclature used in the proposed method

Name	Symbol used
Homo Sapiens	HS
Arabidopsis Thaliana	AT
Drosophila Melanogaster	DM
Danio Rerio	DR
Naming Convention of Homo Sapiens	$[hs_1 - hs_{2051}]$
Naming Convention of Arabidopsis Thaliana	$[at_1 - at_{356}]$
Naming Convention of Drosophila Melanogaster	$[dm_1 - dm_{339}]$
Naming Convention of Danio Rerio	$[dr_1 - dr_{315}]$
Shannon Entropy	SE
Shannon Entropy with Different word size	SED
Fractal Dimension	FD
Hurst Exponent	HE
Purines	Pu
Pyrimidines	Py
Indicator Matrix	ϑ_{hk}

such that the indicator matrix:

$$\vartheta_{hk} = \vartheta(p_h, q_k) = \begin{cases} 1, if p_h = p_k \\ 0, if p_h \neq p_k \end{cases} \quad where p_h, p_k \in S(l) \tag{2}$$

Here ϑ_{hk} is a matrix with the distribution of 0 and 1. A binary image can be obtained from the matrix through which we can visualise correlation between Pu and Py and auto-correlation for the same sequence. It can be well understood by assigning a black dot to 1 and a white dot to 0. From the indicator matrix we can visualise the fractal like distribution of $0's$ and $1's$ (Pu *and* Py). The FD from the indicator matrix can be calculated as the average number of $\sigma(p)$ of 1, which can be taken from $P \times P$ indicator matrix with $p \times p$ randomly. Using $\sigma(p)$, the FD is defined in Equation 3

$$FD = -\frac{1}{P}\sum_{n=2}^{P}\frac{log(\sigma(p))}{log(p)} \tag{3}$$

2.2 HE of Binary Sequences.

The HE is used for time series analysis to interpret the autocorrelation. The value of HE is in between 0 to 1. The HE value $0 < HE < 0.5$ and $0.5 < HE < 1$ designates negative and positive auto-correlation of a time series respectively and 0.5 denotes a absolute randomness of a time series which indicates the equally likely value from a particular value either by increasing or by decreasing. The HE of a binary sequence s_n is defined in Equation 4.

$$\left(\frac{n}{2}\right)^{HE} = \frac{X(n)}{Y(n)} \tag{4}$$

where

$$Y(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - m)}$$

and $X(n) = \max\{T(i,n)\} - \min\{T(i,n)\}$, where

$$T(i) = \sum_{j=1}^{n}(s_i - t)$$

and

$$t = \sqrt{\frac{1}{n}\sum_{i=1}^{n} s_i}$$

2.3 SE of protein sequences

Proteins are made of various combination of amino acids with its length ranging from 30 to 3000 . Some regions of *DNA* sequences like *TTTTTTTTTTC* and *GGGGGGCCCCCC* can code one or different amino acids or encode proteins with duplicate amino acids considered as Homopolymeric regions. Therefore, we assume that these positions are less likely to encode functional proteins. As sequence data are represented as a system where DNA is transformed into amino acids, which can be used to calculate the amount of information or the uncertainty of a sequence. The uncertainty measurement of a sequence with respect to base pair is from 0 to 2n where n is the length of a word. The distribution of each word is depending of the uncertainty. For example, a sequence *TTTTTT* has less information as it can be read as three letter word of "TT". In this contrast a sequence *ATCG* can be read as *AT*, *TC* and *CG* contains more information and uncertainty. The *SE* measures information-entropy with probability p of the two outcomes (0/1) [46–48]. The *SE* for a binary sequence is defined in Equation 5.

$$SE = -\sum_{i=1}^{2} p_i \log_2(p_i) \tag{5}$$

The SE is used to calculate the uncertainty in a binary string. When the probability $p = 0$, the event is certain never to occur which leads that there is no uncertainty, leading to an entropy of 0. Similarly, if the probability $p = 1$, the result is certain, so the entropy must be 0. When $P = 1/2$, the uncertainty is at a maximum and consequently the *SE* is 1. The SE of different word size (*SED*) is defined in Equation 6.

$$SED = -\sum_{j=1}^{k} w_j \log_2(w_i) \tag{6}$$

Where w_i is the frequency of i^{th} word in a sequence. For example, for word length one w_i is calculated from the frequencies of the Pu or Py 0, 1. Similarly, for word length two w_i is calculated from the frequencies of the two-time repetition of Pu or Py 00, 11 and so on. Here k is the number of words obtained by considering the maximum length of both Pu and Py.

3 EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Database used and Specification

In this work we have employed Database of Essential Genes (DEG) [www.essentialgenes.org] dataset for experimental results and discussion purpose. This dataset contains 2051 essential genes of HS, 356 essential genes of AT, 339 essential genes of DM and 315 essential genes of DR respectively. The essential genes of HS are encoded from hs_1 to hs_{2051}, the essential genes of AT are encoded from at_1 to at_{356}, the essential genes of DM are encoded from dm_1 to dm_{339} and the essential genes of DM are encoded from dr_1 to dr_{315}. We have transformed each DNA sequence to a binary sequence of

$0's$ and $1's$ which is defined in equation 7. Here Pu and Py nucleotide bases are represented as "1" and "0" respectively. We then try to find clusters of the corresponding related resonating nucleotide. Thus, the protein sequences are encoded into binary sequencing of $0's$ and $1's$ as per the convention given below. Here we consider four data sets of resulting sequences for four species HS, AT, DM and DR.

$$\begin{aligned} A/G &\rightarrow 0 \\ T/C &\rightarrow 1 \end{aligned} \tag{7}$$

Equation(7) represents Pu and Py nucleotide bases which are encoded as 1 and 0respectively into the transformed binary sequence of essential genes for the four employed species (HS, AT, DM and DR).

3.2 Results and discussion

In this section, results and discussions have been carried out based on different features (FD, HE, SE, SED) for the four employed species and these are as follow:

3.2.1 Analysis of essential genes based on FD from Indicator Matrix. Here from each essential gene sequences of HS, AT, DM and DR, the FD quantitative parameter has been calculated. Based on this quantitative parameter, the dataset has been classified in 10 different clusters as shown in Table 2. The FD of essential gene from HS, AT, DM and DR are also plotted in Fig 1. From this Table 2 and Fig. 1, it is observed that the values of FD on genes for HS lie in the interval of $[0.72 - 1.88]$ while the largest cluster center at 1.497 contains 642 DNA sequences. For AT, the FD lies in the interval of $[0.4 - 1.6]$ and its largest cluster center at 1.199 contains 63 DNA sequences. Similarly the FD for DM lies in range of $[1.27 - 1.66]$ and the largest cluster center at 1.486 contains 91 essential gene DNA sequences. The FD of DR lies in $[0.6 - 1.76]$ range and the largest cluster of DR (center at 1.245) contains 72 essential genes DNA sequences. From these observation it is derived that FD of AT and DM remain in between $[0.70 - 1.9]$ range of HS and the largest FD cluster center of essential gene sequences of HS, DM are approximately same at 1.48 that reflect they are evolutionary close.

Moreover, the largest cluster center of essential genes for other species are not interrelated and it is derived that there is no evolutionary relationship among them. It is also observed that the maximum number of DNA sequences fall in the range $[1.36 - 1.49]$ of the FD cluster centers in case of HS. Hence there are 27, 32, 29 and 4 number of DNA sequences at largest cluster center for HS, AT, DM and DR respectively as shown in Table 2.

Table 2. FD of Indicator Matrix for four Species

Cluster	HS		AT		DM		DR	
	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center
1	34	0.727	2	0.454	22	1.274	9	0.612
2	47	0.856	8	0.578	25	1.316	10	0.739
3	33	0.984	18	0.702	12	1.359	22	0.865
4	112	1.112	24	0.827	21	1.401	51	0.992
5	246	1.240	57	0.951	36	1.444	55	1.118
6	641	1.369	58	1.075	91	1.486	59	1.245
7	642	1.497	63	1.199	23	1.529	52	1.372
8	237	1.625	55	1.323	31	1.571	37	1.498
9	32	1.753	39	1.448	49	1.614	16	1.625
10	27	1.881	32	1.572	29	1.656	4	1.751

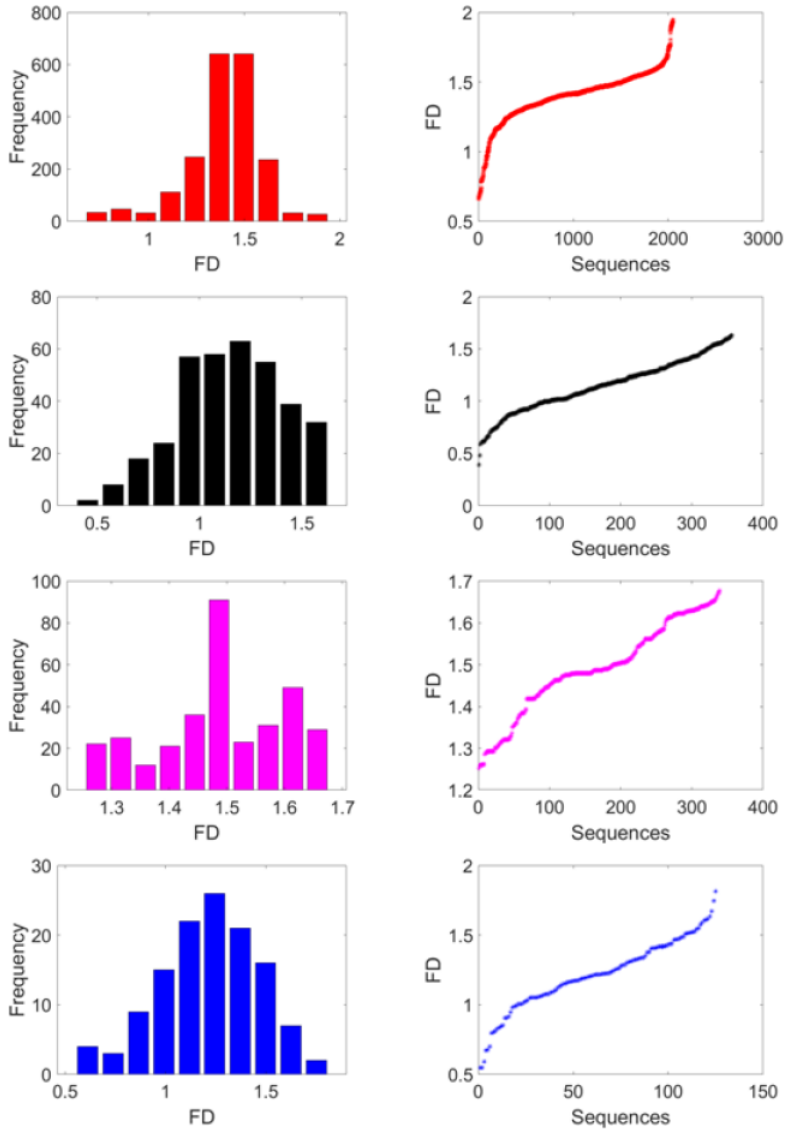

Fig. 1. Histogram of FD of Essential Genes of HS, AT, DM and DR.

3.2.2 Analysis of essential genes based on HE. Here the performance is obtained for HE applied on each essential gene sequence of HS, AT, DM and DR and the dataset of each species has been classified into 10 cluster centers as shown in Table 3. Moreover, the HE of essential genes of HS, AT, DM and DR are also plotted in Fig 2. From Table 3 and Fig. 2 it is observed that HE applied on genes of HS, lie in the range of $[0.517 - 0.719]$ the largest cluster center at 0.608 contains 477 essential gene DNA sequences whereas the HE for AT lie in $[0.519 - 0.785]$ range with the largest cluster center at 0.667 containing 77 number of gene sequences. Similarly, for DM spices the HE lies is in the range of

[0.485 − 0.710] and largest cluster center at 0.585 contains 75 essential gene sequences. The HE for DR lie in the range [0.531 − 0.773] and the largest cluster center at 0.612 contains 67 DNA sequences. Hence the maximum number of DNA sequences for HS, AT , DM and DR at 0.790, 0.785, 0710 and 0.773 cluster center are 7, 1, 7, 2 respectively for HE values as shown in Table 3. Considering the HE values for each cluster for all four species HS, AT, DM and DR have been found closely related as per Table 3.

Table 3. HE for four Species

Cluster	HS		AT		DM		DR	
	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center
1	18	0.517	6	0.519	2	0.485	4	0.531
2	99	0.547	12	0.549	7	0.510	13	0.558
3	291	0.577	34	0.578	19	0.535	33	0.585
4	477	0.608	69	0.608	36	0.560	67	0.612
5	460	0.638	76	0.637	75	0.585	46	0.639
6	385	0.668	77	0.667	71	0.610	51	0.666
7	191	0.699	46	0.696	64	0.635	51	0.693
8	91	0.729	23	0.726	38	0.660	31	0.719
9	32	0.759	12	0.755	20	0.685	17	0.746
10	7	0.790	1	0.785	7	0.710	2	0.773

3.2.3 Analysis of essential genes based on SE. Considering each essential gene sequence of HS, AT, DM and DR, the SE is determined on the dataset to find 10 cluster centers for each species (Table 4). The SE for essential genes of HS, AT, DM and DR are plotted as shown in Fig 3. Here from table and figure it is observed that SE for genes of HS lie in the interval [0.790, 0.988] and the largest cluster center at 0.988 contains 1822 DNA sequences, the SE of AT lie in [0.936, 0.996] range and the largest cluster center at 0.996, contains 257 DNA sequences. Similarly, the SE of DM lie in the range [0.95, 0.997] and largest cluster center at 0.997 contains 271 essential gene sequences. The SE of DR lie in range [0.951, 0.997] and the largest cluster center at 0.997 contains 198 DNA sequences. Here for SE there are no essential gene sequences found for HS in the range [0.791, 0.877], for AT in [0.937, 0.955] range, for DM in [0.951, 0.975] range and for DR in [0.952, 0.960] range. On the intersection of these ranges, it is derived that there are no essential gene sequences in [0.952, 0.955] for AT, DM and DR, the largest cluster centers for SE are equal in the case of DM and DR.

3.2.4 Analysis of essential genes based on modified SE (SED). Here the dynamic code words have been computed using the SE applied on each essential gene sequences of HS, AT, DM and DR respectively and based on it 10 different cluster centers have been computed from each spices. This dynamic code words based on SE is the modified version of SE. The performance evaluation for this experimentation has been shown in Table 5 and the results are plotted in Fig 4. From the Table 5, it has been observed that modified SE of HS genes lie in the interval [0.98 − 1.56] and the largest cluster center at 1.243 contains 527 DNA sequences whereas for AT, it lie in [1.03 − 1.53] range and the largest cluster center at 1.312 contains 90 DNA sequences. Similarly the modified SE of DM lie in the range [1.04 − 1.50] and largest cluster center at 1.297, contains 84 essential gene sequences while for DR the range is [1.05 − 1.61] and the largest cluster center for DR at 1.363 contains 82 essential genes sequences. It has been also observed that FD for AT and DM lie in the modified SE interval [0.98 − 1.56] of HS. The largest FD cluster center on the essential gene sequences for

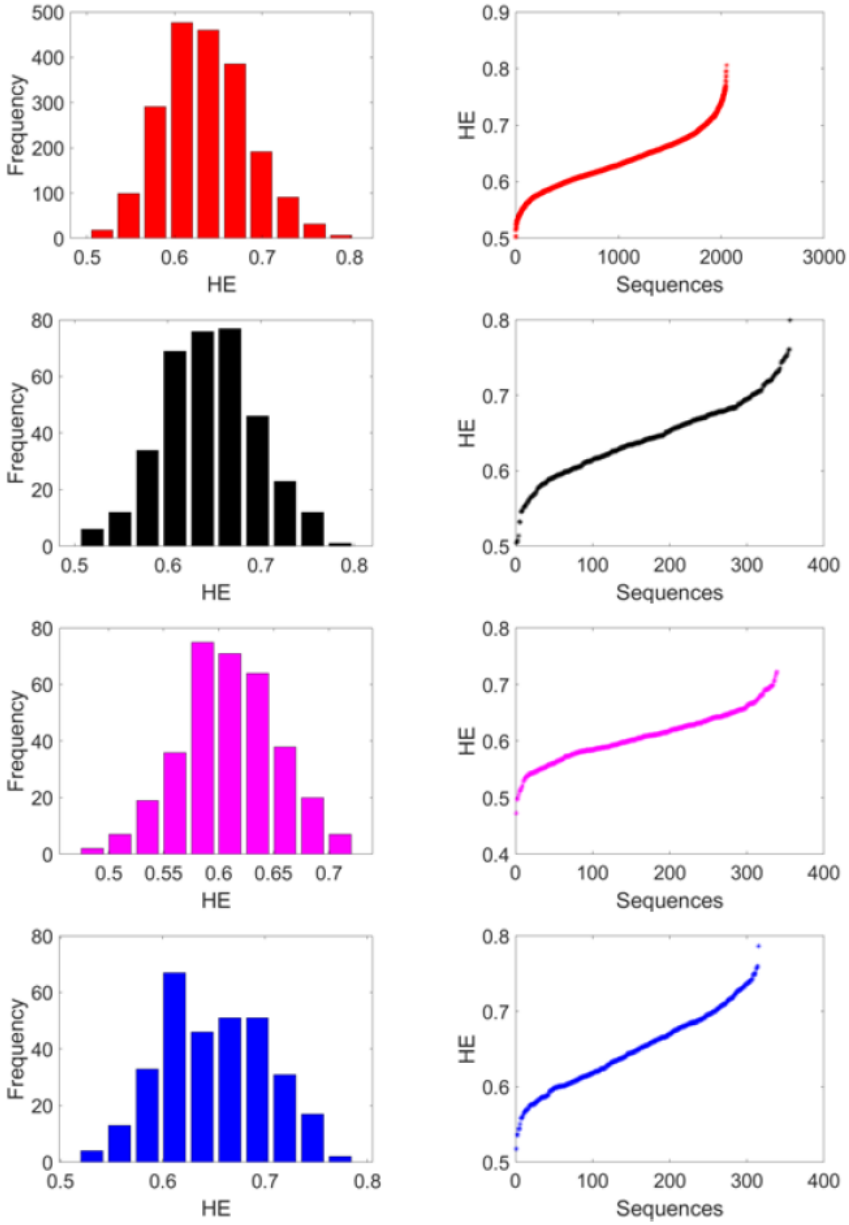

Fig. 2. Histogram of HE of Essential Genes of HS, AT, DM and DR.

AT and DM are approximately same at 1.40 and it reflects that both are evolutionary close while in case of HS the maximum number of DNA sequences fall in the range [1.11 – 1.43] clusters where FD lies in. The number of DNA sequences for HS, AT, DM and DR at the largest cluster center, are 13, 4, 4 and 6 members respectively.

Table 4. SE for four Species

Cluster	HS		AT		DM		DR	
	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center
1	1	0.790	1	0.936	1	0.950	2	0.951
2	0	0.812	0	0.943	0	0.955	0	0.956
3	0	0.834	0	0.950	0	0.960	1	0.961
4	0	0.856	4	0.956	0	0.965	4	0.966
5	1	0.878	3	0.963	0	0.971	1	0.971
6	3	0.907	6	0.970	5	0.976	7	0.977
7	8	0.922	14	0.976	9	0.981	17	0.982
8	41	0.944	19	0.983	11	0.986	35	0.987
9	175	0.966	52	0.990	42	0.992	50	0.992
10	1822	0.988	257	0.996	271	0.997	198	0.997

Table 5. The modified SE for four species.

Cluster	HS		AT		DM		DR	
	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center
1	8	0.988	1	1.035	5	1.043	1	1.053
2	39	1.052	3	1.092	9	1.094	3	1.115
3	109	1.115	13	1.467	20	1.449	11	1.177
4	251	1.179	46	1.202	40	1.195	36	1.239
5	527	1.243	79	1.258	58	1.246	62	1.302
6	478	1.306	90	1.312	84	1.297	82	1.363
7	403	1.370	65	1.368	57	1.348	61	1.425
8	181	1.433	44	1.423	44	1.399	35	1.487
9	42	1.497	11	1.478	18	1.450	18	1.549
10	13	1.561	4	1.534	4	1.501	6	1.611

3.2.5 Analysis of essential genes based on Pu-Py base density. From the percentage of purine and pyrimidine present in each essential gene sequence for the four species under this consideration, it is clearly visible that the distribution of the bases over each gene sequence is nearly the same. It ranges from $[40 : 60 - 75 : 25]$ for each sequences. Accordingly the sequences have been categorized into ten clusters as shown in Table 6. The Purine density of HS, AT, DM and DR has been plotted in Fig.5. Accordingly, the purine density of HS lie in the range $[0.42 - 0.75]$ and the largest cluster center at 0.530, contains 847 gene sequences and that of AT lie in range $[0.36 - 0.61]$ with its largest cluster centered at 0.529 containing 130 essential gene sequences. Similarly in case of DM the largest cluster is centered at 0.532 having 117 gene sequences and for DR the density lie in the range of $[0.47 - 0.64]$ with largest cluster center at 0.538 having 69 gene sequences. This is quite observable that all the species have their largest cluster centers at around 0.530 indicates the percentage of purine is nearly same as that of pyrimidine and the range of HS purine density covers all of the other species. The number of sequences for all the four species present in the 10^{th} cluster as well as in the first cluster is too less such as for HS it is 1, for AT, it is 11, for DM it is 1 and for DR, it is 3 (as shown in Table 6). Considering these it is inferred that density of bases is nearly same for all the four species essential gene sequences (Fig. 5).

3.2.6 Analysis of essential genes based on the distribution of Pu-Py bases. The distribution of purine and pyrimidine is not even in all species. In some of the sequences, Pu are dominant whereas in some sequences Py are dominant.

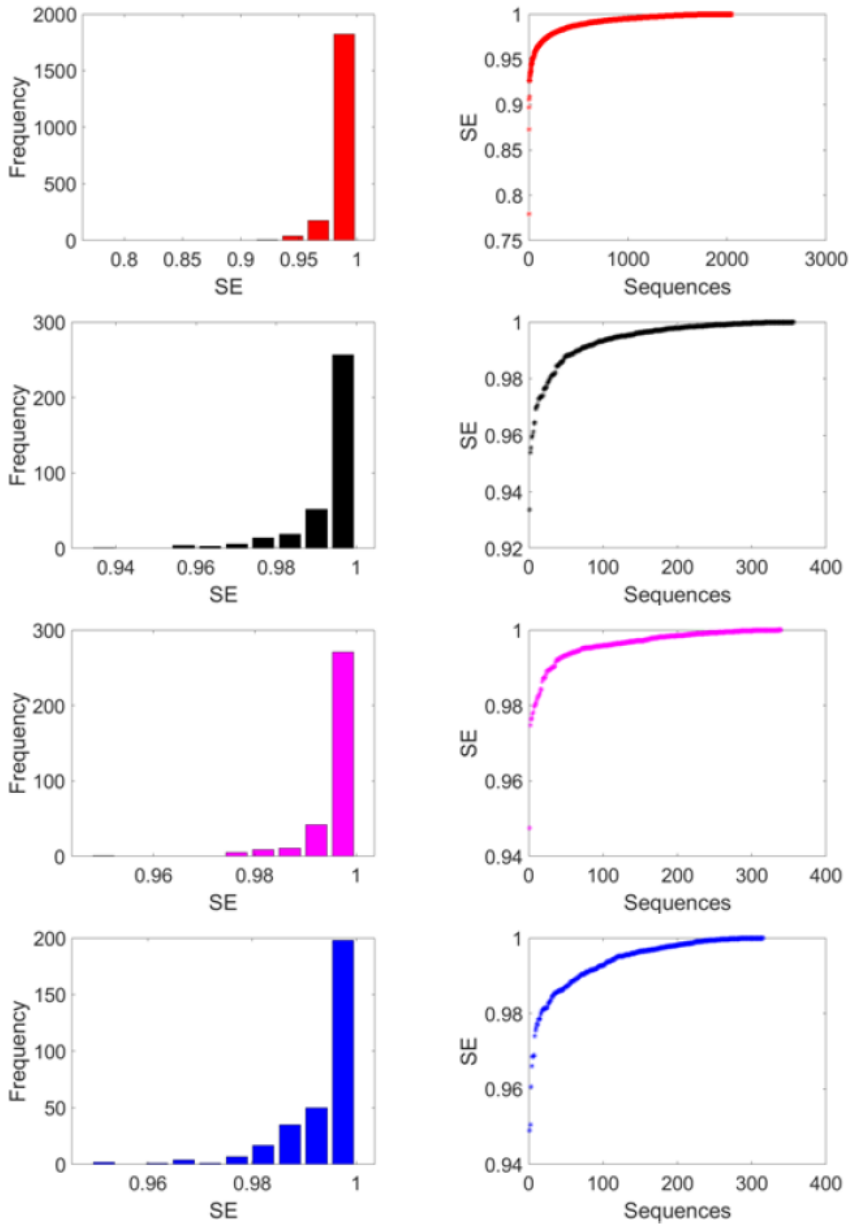

Fig. 3. Histogram of SE of Essential Genes of HS, AT, DM and DR.

Throughout the analysis of the distribution of bases, it is found that the clusters are normally distributed over the space. Considering the *FD* of the different gene sequences, it is inferred that the distribution of bases in the maximum gene sequence, is evenly distributed. It is also clearly visible from Table 2 that there are 27 essential genes of HS in

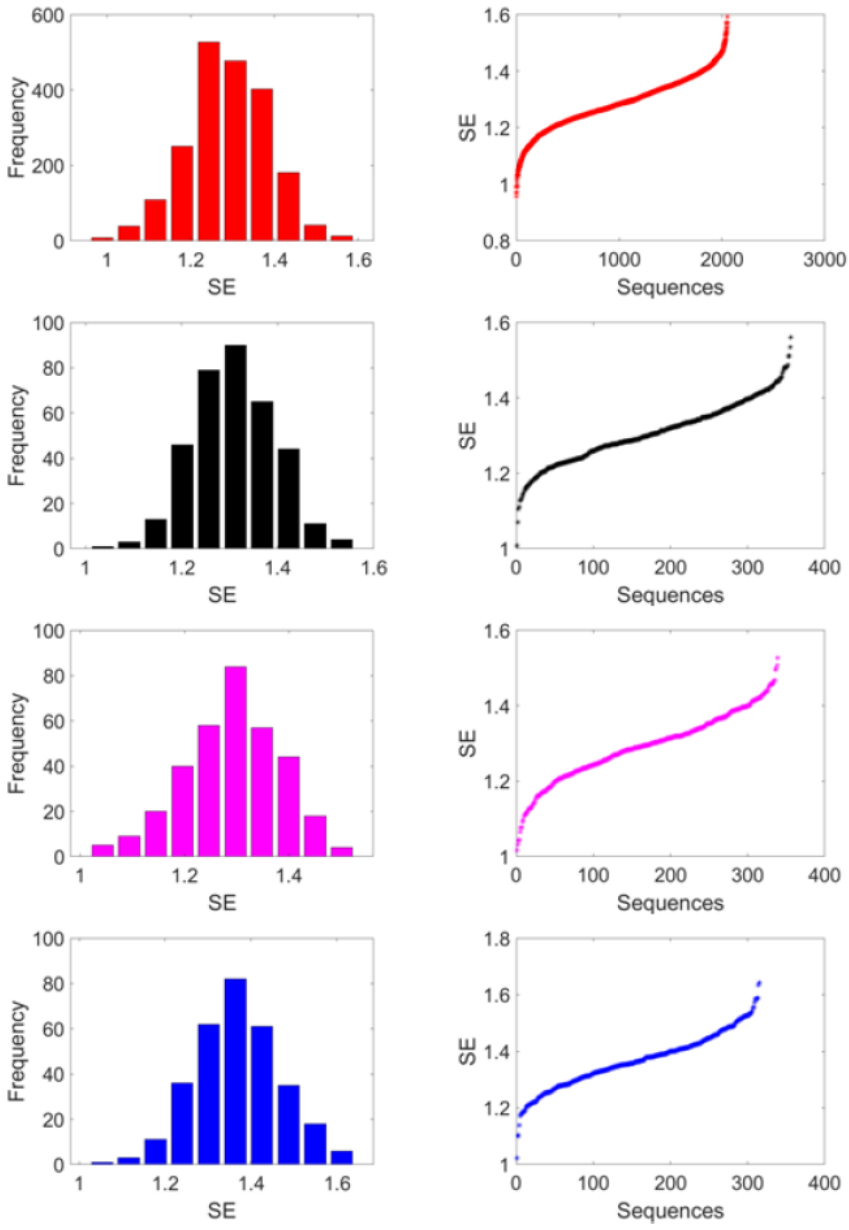

Fig. 4. Histogram of Modified SE of Essential Genes of HS, AT, DM and DR.

Fig. 6 having maximum *FD*. Similarly, there are 32 essential genes of AT, 29 of DM, 4 gene sequences of DR with maximum *FD*. Similar reflection are found for other clusters also. The ratio of Pu to Py is found to be in the range of $[30 : 70 - 70 : 30]$, but in maximum sequences the ratio is found to be in the range of $[45 : 55 - 55 : 45]$. On this

Table 6. Performance based on Purine Density

Cluster	HS		AT		DM		DR	
	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center	No. of Sequences	Center
1	9	0.420	1	0.363	5	0.458	8	0.470
2	50	0.457	2	0.391	14	0.477	21	0.487
3	372	0.494	1	0.419	37	0.495	47	0.504
4	847	0.530	8	0.446	99	0.514	65	0.521
5	531	0.567	29	0.474	117	0.532	69	0.538
6	188	0.604	78	0.501	38	0.551	50	0.555
7	49	0.641	130	0.529	18	0.569	36	0.572
8	3	0.677	70	0.557	10	0.588	12	0.589
9	1	0714	26	0.584	0	0.606	4	0.607
10	1	0.751	11	0.612	1	0.625	3	0.624

consideration of each cluster for all four species, it is inferred that at 9^{th} cluster in Table 2 the DR and DM have nearly same FD value which has been shown in Fig 7.

From Table 3, it is observed that there are the sequences of HS and AT have nearly equal *HE* value when considering any cluster. DR spices have some sequences which have *HE* nearly equal to HS and AT in their respective clusters. From Table 3 it is observed that for HS there are 7 sequences at *HE* value (0.790), 2 for AT (0.785), 2 for DR (0.773) which have nearly same *HE* value as shown in Fig 8. There is no such sequences which has uncorrelated in the Pu and Py spatial ordering. These results derive that there is no randomness in the distribution of bases in the genes sequences for the four species. In some cases, the percentage of Pu is nearly around 40%, for the Py is 60% and for other cases it is vice-versa. Similarly, considering at 5^{th} and 1^{st} cluster this observation is verified from Fig.9 and Fig.10.

Analyzing the distribution of bases using SE it is observed that number of gene sequences having maximum SE are most in number for all the four species. The gene sequences having maximum *SE* are 1822 in case of HS, 257 in case of AT, 271 for DM and 198 for DR (see Table 4). As on contrary the number of sequences having minimum SE are also very less in number for all the four species and they are 1 for HS, AT, DM and 2 for DR (Table 4). The Pu-Py density for sequences having minimum *SE* is around 35 : 65 for DR, DM and AT whereas it is around 76 : 24 for HS. The SE value for HS is around 0.77 whereas it is around 0.94 for other three species as shown in Table 7. For all the four species, no sequence is found in second cluster respectively. These points indicate that the probability of finding the bases with one type either Pu or Py is very less. But there is randomness in the distribution of the bases over the sequences.

Table 7. SE of Essential Genes of HS, AT,Drosophila Melanogaster and DR having minimum SE with their Pu and Py percentage

Name of gene Sequences	SE	Pu Density	Py Density
hs_{1461}	0.779	0.769	0.231
at_{35}	0.948	0.35	0.65
dr_{149}	0.951	0.63	0.37
dr_{177}	0.949	0.632	0.368
dm_{147}	0.948	0.634	0.366

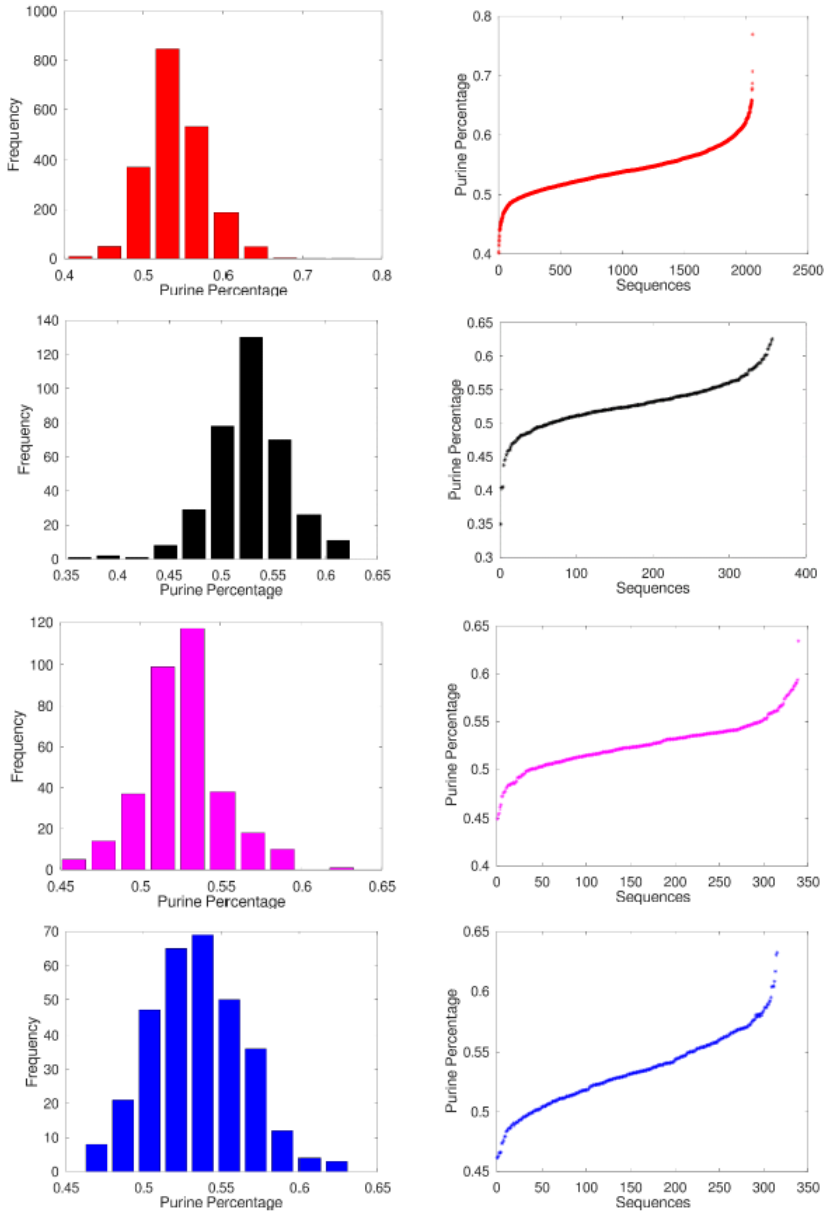

Fig. 5. Histogram of Purine Density of Essential Genes of HS, AT, DM and DR

Considering the value of *SE* for each cluster it is observed that DM and DR are closely related and AT is nearly related to these species. Considering cluster at 1^{st}, 5^{th} and 10^{th} in Table 4, the results for SE and the corresponding Pu and Py densities have been shown in Table 9 and in Fig. 11.

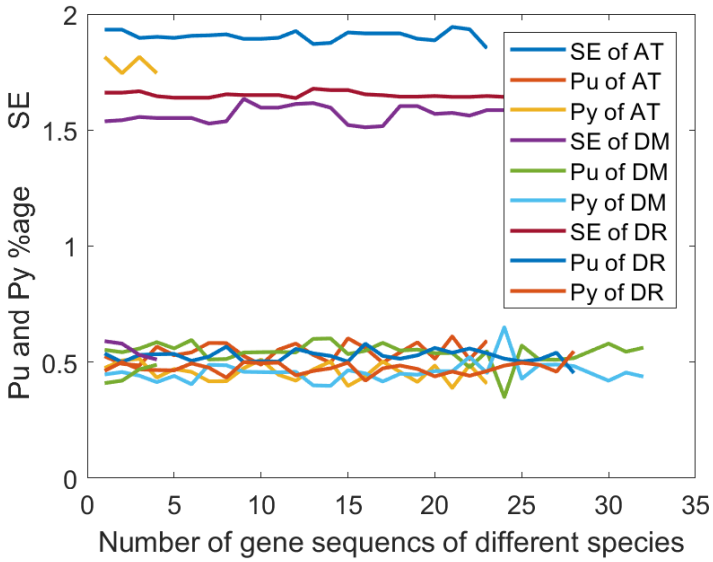

Fig. 6. Line Plot for FD of Essential Genes of HS, AT, DM and DR having maximum FD with their Pu and Py percentage

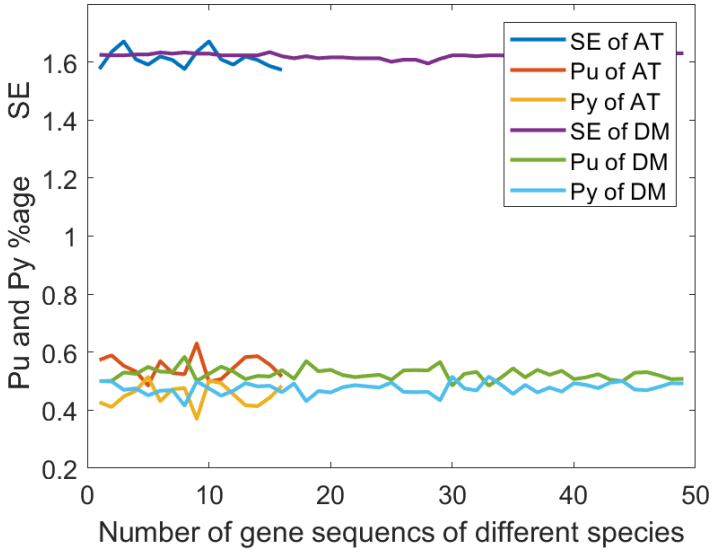

Fig. 7. Line Plot for FD of Essential Genes of DM and DR belonging to 9^{th} cluster with their Pu and Py percentage

Hence it is found that at 10^{th} cluster that is having maximum SE, contains 213 essential gene sequences for all four species that have SE value equal to 1 indicating that they have nearly equal Pu and Py density. Out of these 213 sequences, DM have 76 gene sequences with SE value 1, DR has 64, AT has 70 and for HS 3 gene sequences are there.

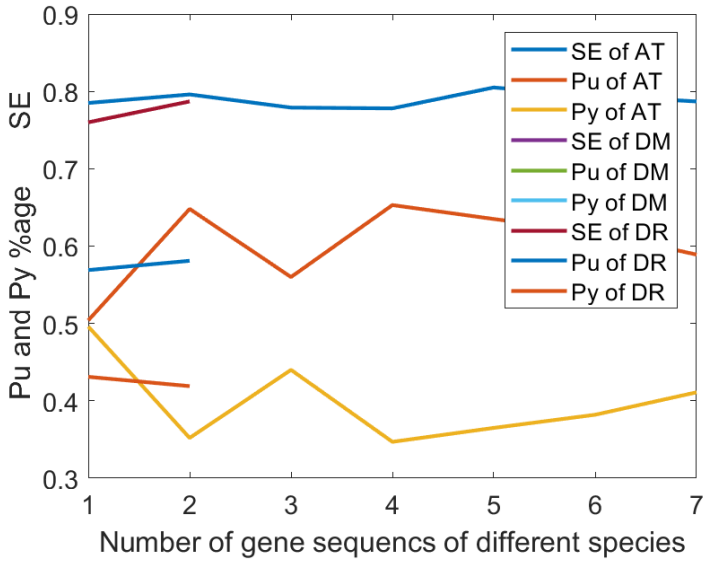

Fig. 8. Line Plot for HE of Essential Genes of HS, AT and DR belonging to cluster 10 with Pu and Py percentage

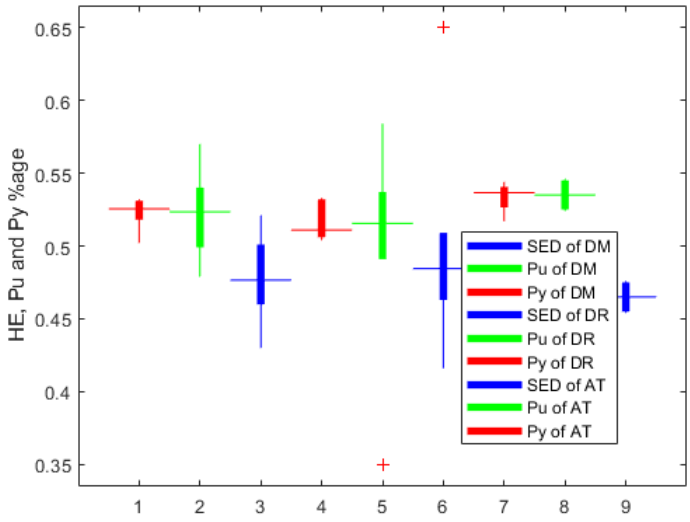

Fig. 9. Line Plot for HE of Essential Genes of HS, AT and DR of cluster 10^{th} having maximum FD with their Pu and Py percentage

By analyzing the distribution of bases using modified SE, it is observed that the number of gene sequences having maximum SE are very few in number for all the four species. The gene sequences having maximum SE are 13 in case of HS, 4 in case of AT, 4 for DM and 6 for DR as shown in Table 5. Similarly the number of sequences having minimum SE

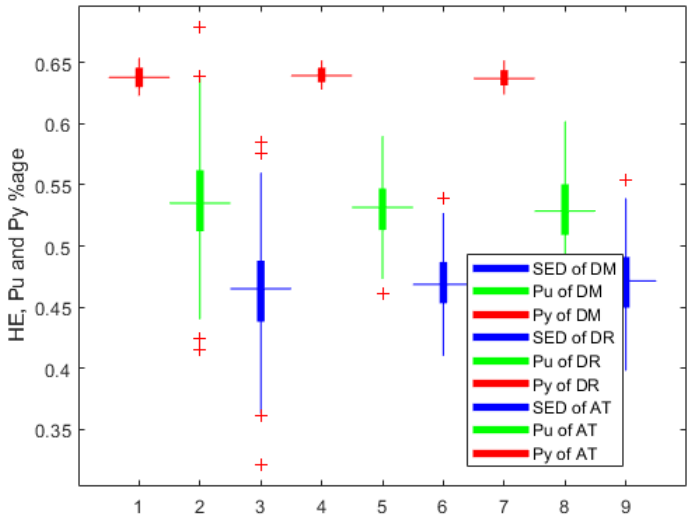

Fig. 10. Box Plot for HE of Essential Genes of HS, AT, and DR belonging to 5th cluster with their Pu and Py percentage

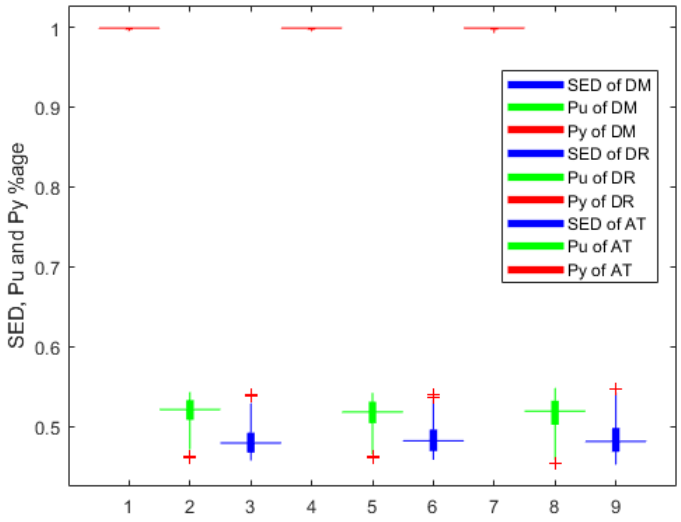

Fig. 11. Box Plot for SE of Essential Genes of AT, DM and DR belonging to 10^{th} cluster with their Pu and Py percentage

are also very less in number for all four species and they are 8 for HS, 1 for AT, 5 for DM and 1 for DR. Considering each Cluster for four species, it is found that on basis of dynamic code words for SE, AT and DM are closely related whereas DR is nearly related respect to these two species. Considering at 1^{st}, 5^{th} and 10^{th} cluster, the SE value and

Table 8. SE of Essential Genes of AT,DM and DR having minimum SE with their Pu and Py percentage

Name of gene sequence	SE	Pu Density	Py Density
at_{35}	0.948	0.35	0.65
dr_{149}	0.951	0.63	0.37
dr_{177}	0.949	0.632	0.368
dm_{147}	0.948	0.634	0.366

the corresponding Pu and Py density is illustrated in corresponding Table 10 and in Fig 12 and 13. It is found that the clusters having maximum number of sequences have also changed on the changing of criteria for SE. On the basis of dynamic code words, the clusters have been shifted from 10^{th} to 5^{th} for HS, 6^{th} for DM, DR and AT.

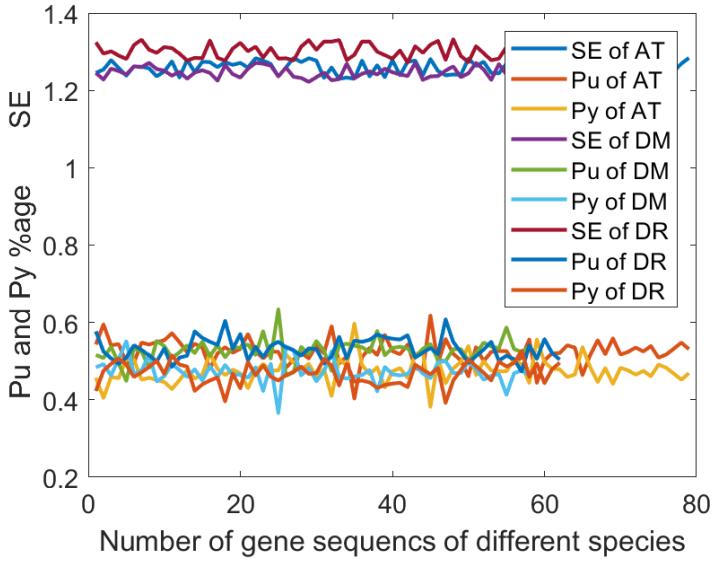

Fig. 12. Line Plot for SE of Essential Genes of AT, DM and DR belonging to 5^{th} cluster with their Pu and Py percentage

Table 9. SE of Essential Genes of AT ,Drosophila Melanogaster and DR belonging to 5th Cluster with their Pu and Py percentage

Name of gene sequence	SE	Pu Density	Py Density
at_{166}	0.965	0.61	0.39
at_{317}	0.961	0.615	0.385
at_{330}	0.964	0.611	0.389
dr_{247}	0.974	0.595	0.405

Here out of 2051 gene sequences for HS, three sequence have equal density as shown in Table 11. Similarly for AT only one out of 356, in case of DM, there are 4 gene sequences and for DR it is only one as shown in Table 11. In case of HS less than 20 sequences are having Pu density less than 50%. Similarly for all other species the same fashion has been

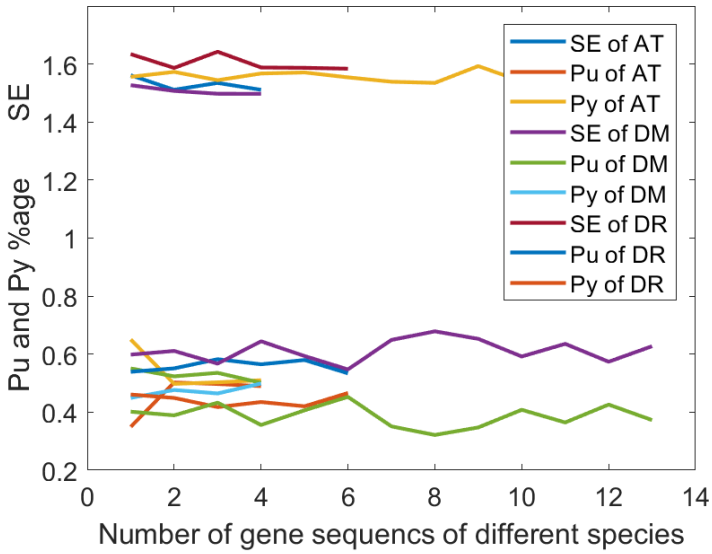

Fig. 13. Line Plot for SE with dynamic code word of Essential Genes of AT, DM and DR belonging to 10^{th} cluster with their Pu and Py percentage

Table 10. SE on basis of dynamic code words of Essential Genes of AT, DM and DR belonging to 1st Cluster with their Pu and Py percentage

Name of gene sequence	SE	Pu Density	Py Density
at_{53}	1.008	0.47	0.53
dr_{243}	1.023	0.53	0.47
dm_{24}	1.032	0.498	0.502
dm_{54}	1.017	0.521	0.479
dm_{91}	1.043	0.5	0.5
dm_{153}	1.044	0.504	0.496
dm_{186}	1.065	0.515	0.485

observed. In case of Pu dominant species have also less in number respect to four species. These all indicate that both bases are equally required for the survival of the species.

Finally, the evolutionary closeness among the species utilizing five parameters are shown in Table 12.

4 CONCLUSIONS

One of the most important and typical constituents of genes are purine and pyrimidine. An attempt has been made to analyze the distribution of the purine and pyrimidine distributions over the essential genes of the four different species and consequently a quantitative classifications and correlations have been fetched out. Based on the quantitative investigations made, some important observations have been reported. This study steps ahead of finding the evolutionary closeness among inter and intra families of different clusters of essential genes. In near future, other available essential genes can be studied and strengthen the observation reported here.

Table 11. Essential Genes of HS, AT, Drosophila Melanogaster and DR having equal Pu and Py percentage

Name of gene sequence	No. of Pu Base	Pu Density	No. of Py Base	Py Density
at_{69}	534	0.5	534	0.5
dm_3	240	0.5	240	0.5
dm_{27}	606	0.5	606	0.5
dm_{92}	192	0.5	192	0.5
dm_{67}	720	0.5	720	0.5
dr_{37}	1186	0.5	1186	0.5
hs_{1823}	387	0.5	387	0.5
hs_{1740}	1629	0.5	1629	0.5
hs_{1671}	108	0.5	108	0.5

Table 12. Closeness of species based on the each parameters

Parameter	Species Name
FD	$\{HS\}, \{AT\}, \{DM\}, \{DR\}$
HE	$\{HS, \{AT, DR\}\}$
SE	$\{\{DM, DR\}, AT\}$
SED	$\{\{AT, DM\}, DR\}$

Competing Interests: The authors declare no competing interests.

REFERENCES

[1] Ryan S O'Neill and Denise V Clark. The drosophila melanogaster septin gene sep2 has a redundant function with the retrogene sep5 in imaginal cell proliferation but is essential for oogenesis. *Genome*, 56(12):753–758, 2013.

[2] Yipin Wu, Michel Baum, Chou-Long Huang, and Aylin R Rodan. Two inwardly rectifying potassium channels, irk1 and irk2, play redundant roles in drosophila renal tubule function. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 309(7):R747–R756, 2015.

[3] Eugene V Koonin. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2):127, 2003.

[4] Eugene V Koonin. How many genes can make a cell: the minimal-gene-set concept. *Annual review of genomics and human genetics*, 1(1):99–116, 2000.

[5] Fabian M Commichau, Nico Pietack, and Jörg Stülke. Essential genes in bacillus subtilis: a re-evaluation after ten years. *Molecular BioSystems*, 9(6): 1068–1075, 2013.

[6] Mitsuhiro Itaya. An estimation of minimal genome size required for life. *FEBS letters*, 362(3):257–260, 1995.

[7] Ren Zhang and Yan Lin. Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*, 37(suppl_1):D455–D458, 2008.

[8] Lars M Steinmetz, Curt Scharfe, Adam M Deutschbauer, Dejana Mokranjac, Zelek S Herman, Ted Jones, Angela M Chu, Guri Giaever, Holger Prokisch, Peter J Oefner, et al. Systematic screen for human disease genes in yeast. *Nature genetics*, 31(4):400, 2002.

[9] Gyanu Lamichhane, Matteo Zignol, Natalie J Blades, Deborah E Geiman, Annette Dougherty, Jacques Grosset, Karl W Broman, and William R Bishai. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 100(12):7213–7218, 2003.

[10] Wenqi Hu, Susan Sillaots, Sebastien Lemieux, John Davison, Sarah Kauffman, Anouk Breton, Annie Linteau, Chunlin Xin, Joel Bowman, Jeff Becker, et al. Essential gene identification and drug target prioritization in aspergillus fumigatus. *PLoS pathogens*, 3(3):e24, 2007.

[11] Terry Roemer, Bo Jiang, John Davison, Troy Ketela, Karynn Veillette, Anouk Breton, Fatou Tandia, Annie Linteau, Susan Sillaots, Catarina Marta, et al. Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery. *Molecular microbiology*, 50(1): 167–181, 2003.

[12] Scott A Becker and Bernhard Ø Palsson. Genome-scale reconstruction of the metabolic network in staphylococcus aureus n315: an initial draft to the two-dimensional annotation. *BMC microbiology*, 5(1):8, 2005.

[13] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Veronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno Andre, et al. Functional profiling of the saccharomyces cerevisiae genome. *nature*, 418(6896):387, 2002.

[14] Yu Chen and Dong Xu. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics*, 21(5): 575–581, 2004.

[15] Jens Harborth, Sayda M Elbashir, Kim Bechert, Thomas Tuschl, and Klaus Weber. Identification of essential genes in cultured mammalian cells using small interfering rnas. *Journal of cell science*, 114(24):4557–4565, 2001.

[16] Yinduo Ji, Barbara Zhang, Stephanie F Van, Patrick Warren, Gary Woodnutt, Martin KR Burnham, Martin Rosenberg, et al. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense rna. *Science*, 293(5538):2266–2269, 2001.

[17] Larry A Gallagher, Elizabeth Ramage, Michael A Jacobs, Rajinder Kaul, Mitchell Brittnacher, and Colin Manoil. A comprehensive transposon mutant library of francisella novicida, a bioweapon surrogate. *Proceedings of the National Academy of Sciences*, 104(3):1009–1014, 2007.

[18] Gemma C Langridge, Minh-Duy Phan, Daniel J Turner, Timothy T Perkins, Leopold Parts, Jana Haase, Ian Charles, Duncan J Maskell, Sarah E Peters, Gordon Dougan, et al. Simultaneous assay of every salmonella typhi gene using one million transposon mutants. *Genome research*, 19(12): 2308–2316, 2009.

[19] Ranjeet Kumar Rout, Pabitra Pal Choudhury, Santi Prasad Maity, BS Daya Sagar, and Sk Sarif Hassan. Fractal and mathematical morphology in intricate comparison between tertiary protein structures. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6 (2):192–203, 2018.

[20] Kang Ning, Hoong Kee Ng, Sriganesh Srihari, Hon Wai Leong, and Alexey I Nesvizhskii. Examination of the relationship between essential genes in ppi network and hub proteins in reverse nearest neighbor topology. *BMC bioinformatics*, 11(1):505, 2010.

[21] Yuan-Nong Ye, Zhi-Gang Hua, Jian Huang, Nini Rao, and Feng-Biao Guo. Ceg: a database of essential gene clusters. *BMC genomics*, 14(1):769, 2013.

[22] Yao Lu, Jingyuan Deng, Matthew B Carson, Hui Lu, and Long J Lu. Computational methods for the prediction of microbial essential genes. *Current Bioinformatics*, 9(2):89–101, 2014.

[23] Ping Xu, Xiuchun Ge, Lei Chen, Xiaojing Wang, Yuetan Dou, Jerry Z Xu, Jenishkumar R Patel, Victoria Stone, My Trinh, Karra Evans, et al. Genome-wide essential gene identification in streptococcus sanguinis. *Scientific reports*, 1:125, 2011.

[24] Sergei Maslov and Kim Sneppen. Protein interaction networks beyond artifacts. *FEBS letters*, 530(1-3):255–256, 2002.

[25] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[26] Haiyuan Yu, Dov Greenbaum, Hao Xin Lu, Xiaowei Zhu, and Mark Gerstein. Genomic analysis of essentiality within protein networks. *TRENDS in Genetics*, 20(6):227–231, 2004.

[27] Balázs Papp, Csaba Pál, and Laurence D Hurst. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, 429(6992):661, 2004.

[28] Felipe Sarmiento, Jan Mrázek, and William B Whitman. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon methanococcus maripaludis. *Proceedings of the National Academy of Sciences*, 110(12):4726–4731, 2013.

[29] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Veronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno Andre, et al. Functional profiling of the saccharomyces cerevisiae genome. *nature*, 418(6896):387, 2002.

[30] Dong-Uk Kim, Jacqueline Hayles, Dongsup Kim, Valerie Wood, Han-Oh Park, Misun Won, Hyang-Sook Yoo, Trevor Duhig, Miyoung Nam, Georgia Palmer, et al. Analysis of a genome-wide set of gene deletions in the fission yeast schizosaccharomyces pombe. *Nature biotechnology*, 28(6):617, 2010.

[31] David Meinke, Rosanna Muralla, Colleen Sweeney, and Allan Dickerman. Identifying essential genes in arabidopsis thaliana. *Trends in plant science*, 13(9):483–491, 2008.

[32] Ben-Yang Liao and Jianzhi Zhang. Mouse duplicate genes are as essential as singletons. *Trends in Genetics*, 23(8):378–381, 2007.

[33] Vincent A Blomen, Peter Májek, Lucas T Jae, Johannes W Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco, Ferdy R van Diemen, Nadine Olk, Alexey Stukalov, et al. Gene essentiality and synthetic lethality in haploid human cells. *Science*, 350(6264):1092–1096, 2015.

[34] Tim Wang, Kıvanç Birsoy, Nicholas W Hughes, Kevin M Krupczak, Yorick Post, Jenny J Wei, Eric S Lander, and David M Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, 2015.

[35] LW Ning, H Lin, H Ding, J Huang, N Rao, and FB Guo. Predicting bacterial essential genes using only sequence composition information. *Genet. Mol. Res*, 13:4564–4572, 2014.

[36] Yongming Yu, Licai Yang, Zhiping Liu, and Chuansheng Zhu. Gene essentiality prediction based on fractal features and machine learning. *Molecular BioSystems*, 13(3):577–584, 2017.

[37] Kitiporn Plaimas, Roland Eils, and Rainer König. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC systems biology*, 4(1):56, 2010.

[38] Marcio L Acencio and Ney Lemke. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC bioinformatics*, 10(1):290, 2009.

[39] Yao Lu, Jingyuan Deng, Judith C Rhodes, Hui Lu, and Long Jason Lu. Predicting essential genes for identifying potential drug targets in aspergillus fumigatus. *Computational biology and chemistry*, 50:29–40, 2014.

[40] Jian Cheng, Zhao Xu, Wenwu Wu, Li Zhao, Xiangchen Li, Yanlin Liu, and Shiheng Tao. Training set selection for the prediction of essential genes. *PloS one*, 9(1):e86805, 2014.

[41] Xiao Liu, Bao-Jin Wang, Luo Xu, Hong-Ling Tang, and Guo-Qing Xu. Selection of key sequence-based features for prediction of essential genes in 31 diverse bacterial species. *PloS one*, 12(3):e0174638, 2017.

[42] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[43] Carlo Cattani. Fractals and hidden symmetries in dna. *Mathematical problems in engineering*, 2010, 2010.

[44] Jayanta Kumar Das, Pabitra Pal Choudhury, Adwitiya Chaudhuri, Sk Sarif Hassan, and Pallab Basu. Analysis of purines and pyrimidines distribution over mirnas of human, gorilla, chimpanzee, mouse and rat. *Scientific reports*, 8(1):9974, 2018.

[45] Cheryl L Berthelsen, James A Glazier, and Mark H Skolnick. Global fractal dimension of human dna sequences treated as pseudorandom walks. *Physical Review A*, 45(12):8902, 1992.

[46] Konstantin Makarychev, Yury Makarychev, Andrei Romashchenko, and Nikolai Vereshchagin. A new class of non-shannon-type inequalities for entropies. *Communications in Information and Systems*, 2(2):147–166, 2002.

[47] Wojciech H Zurek. Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731, 1989.

[48] Ty Roach, James Nulton, Paolo Sibani, Forest Rohwer, and Peter Salamon. Entropy in the tangled nature model of evolution. *Entropy*, 19(5):192, 2017.