INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Bioinformatics Research and Applications

ISSN online: 1744-5493 - ISSN print: 1744-5485
https://www.inderscience.com/ijbra

Prediction of essential genes using single nucleotide compositional features in genomes of bacteria: a machine learning-based analysis

Annushree Kurmi, Piyali Sen, Madhusmita Dash, Aswini Kumar Patra, Suvendra Kumar Ray, Siddhartha Sankar Satapathy

DOI: 10.1504/IJBRA.2023.10054584

Article History:

Received:	30 March 2022
Last revised:	31 March 2022
Accepted:	04 January 2023
Published online:	05 June 2023

Prediction of essential genes using single nucleotide compositional features in genomes of bacteria: a machine learning-based analysis

Annushree Kurmi and Piyali Sen

Department of Computer Science and Engineering,
Tezpur University,
Napaam-784028, Assam, India
Email: annushreekurmi@gmail.com
Email: piyalisen18@gmail.com

Madhusmita Dash

Department of Electronics and Communication Engineering,
NIT,
Jote-791113, Arunachal Pradesh, India
Email: madhusmita.dash81@gmail.com

Aswini Kumar Patra

Department of Computer Science and Engineering,
North Eastern Regional Institute of Science and Technology (NERIST),
Nirjuli (Itanagar)-791109, Arunachal Pradesh, India
Email: aswinipatra@gmail.com

Suvendra Kumar Ray

Department of Molecular Biology and Biotechnology,
Tezpur University,
Napaam-784028, Assam, India
Email: suven@tezu.ernet.in

Siddhartha Sankar Satapathy*

Department of Computer Science and Engineering,
Tezpur University,
Napaam-784028, Assam, India
Email: ssankar@tezu.ernet.in
*Corresponding author

Abstract: Essential genes are crucial for understanding the cellular processes of an organism. In this article, we have done an extensive machine learning-based analysis of single nucleotide composition in 35 bacterial genomes across several phylogenetic groups. With an objective of classifying essential genes from the remaining genes, we have used seven machine

learning-based classifiers – logistic regression, Gaussian Naïve Bayes, k-nearest neighbours, decision tree, random forest, extreme gradient boosting and support vector machine. Random forest classifier was a better performer among the seven classifiers and achieved an AUC score of at least 70% for thirteen organisms. Higher AUC scores were achieved for several organisms such as *Salmonella enterica*, *Sphingomonas wittichii*, *Bacillus thuringiensis*, and *Streptococcus pyogenes*. Prediction result obtained in general from the machine learning-based analysis suggests that the single nucleotide compositional features may be useful in predicting gene essentiality in some bacteria species though not universally.

Keywords: essential genes; single nucleotide composition; bacterial genome; machine learning.

Reference to this paper should be made as follows: Kurmi, A., Sen, P., Dash, M., Patra, A.K., Ray, S.K. and Satapathy, S.S. (2023) 'Prediction of essential genes using single nucleotide compositional features in genomes of bacteria: a machine learning-based analysis', *Int. J. Bioinformatics Research and Applications*, Vol. 19, No. 1, pp.1–18.

Biographical notes: Annushree Kurmi is a PhD student in the Department of Computer Science and Engineering, Tezpur University, Assam, India. She received the MTech in Information Technology from Tezpur University. At present, she is also working as an Assistant Professor in the Department of Computer Science and Enginerring in The Assam Kaziranga University. Her research interests include bioinformatics, machine learning and software engineering.

Piyali Sen is a PhD student in the Department of Computer Science and Engineering, Tezpur University, Assam, India. She received her MSc in Computer Science from Assam University, India. Her research interests include machine learning, bioinformatics and image processing.

Madhusmita Dash is a PhD student in the Department of Electronics and Communication Engineering, National Institute of Technology, Arunachal Pradesh, India. Her research interests include machine learning and bioinformatics.

Aswini Kumar Patra is an Assistant Professor in the Department of Computer Science and Engineering, NERIST, Arunachal Pradesh, India. He is pursuing his PhD in Computer Science and Engineering at the Indian Institute of Technology, Guwahati. His research interests include evolutionary computing, data mining and complex networks.

Suvendra Kumar Ray is a Professor in the Department of Molecular Biology and Biotechnology, Tezpur University. He received his PhD degree from the Centre for Cellular and Molecular Biology, Hyderabad, India in 2002. His research interests include genetics, evolutionary genetics and molecular biology.

Siddhartha Sankar Satapathy is an Associate Professor in the Department of Computer Science and Engineering, Tezpur University, Assam, India. He received his MTech and PhD degrees from Tezpur University. His research interests include bioinformatics and ad hoc networks.

1 Introduction

Each strand of the double-stranded DNA is a sequence of four nucleotide bases denoted as *A*, *T*, *G*, and *C*, whose size varies from a few kilo-bases to the order of mega-bases among different bacterial species. These four bases differ in terms of both chemical and physical properties and are accordingly categorised into different groups. In terms of associated nitrogenous bases, *A* and *G* are classified as purines (*R*) and *C* and *T* as pyrimidines (*Y*); with reference to the associated functional groups, *A* and *C* are grouped as amino (*M*) and *G* and *T* as keto (*K*); and according to complementary base pairing strength, *A* and *T* are grouped as weak (*W*) and *C* and *G* are grouped as strong (*S*). Based on these physio-chemical properties of the nucleotides, the compositional feature of any genome is popularly represented as (*G* + *C*)% or equivalently also as (*A* + *T*)%. In any hypothetical randomly generated large genome sequence, all the bases are likely to occur in equal frequencies, and therefore (*G* + *C*)% is expected to be 50.0%. However, base composition in natural DNA sequences varies among living organisms (Bohlin, 2008); (*G* + *C*)% ranges from less than 15.0% to more than 75.0% in bacterial genomes (Hildebrand, 2010). Though this considerable variation in the inter-genomic base composition among bacteria is not yet fully understood, some of the intrinsic factors attributed towards this variation are as follows: growth rate (Rocha and Danchin, 2002), growth temperature (Zheng and Wu, 2010), the genome size (Satapathy et al., 2010), external gene transfer (Srividhya et al., 2007; Langille at al., 2008), mutation (Muto and Osawa, 1987) and selection (Raghavan et al., 2012).

Apart from inter-genomic variation, (*G* + *C*)% varies to a great extent within the genome of an organism. This intra-genomic base composition variation among different functional segments within genome of a species is attributed to several inherent mutation and selection factors (Frank and Lobry, 1999). Chargaff et al. observed approximately equivalent frequencies of complementary nucleotides within individual DNA strands of bacterial chromosomes (Karkas et al., 1968; Rudner et al., 1969). This observation is known as intra-strand parity (ISP) (Forsdyke and Mortimer, 2000). Violation of ISP was also observed in several chromosomes (Nikolaou and Almirantis, 2006). Because of replication associated factors, base composition differs between continuously and dis-continuously synthesised halves of a DNA strand (Francino and Ochman, 1997; Rocha et al., 1999; Lobry and Sueoka, 2002; Rocha, 2004; Nikolaou and Almirantis, 2005). *G* + *C* content also differ between protein coding genic regions (CDS) and non-coding intergenic regions (IRs). Usually (*G* + *C*)% in CDS is more than that of IRs in any bacterial genomes, because of synonymous codon assignments in genetic code table. Amino acid specific selection on codon usage factors have been attributed towards variation in (*G* + *C*)% among the protein coding genes in an organism (Bulmer, 1991; Sueoka, 1995; Sharp et al., 2010; Satapathy et al., 2016). Variation in (*G* + *C*)% also has been attributed to selection for DNA stability as free energy associated between *G* and *C* base pair is more than that between *A* and *T* pair (Yakovchuk et al., 2006). These genomic compositional features have been used by researchers using computational methods towards addressing several biological issues such as estimation of gene expression (Sharp and Li, 1987; Sen et al., 2019) and selection, finding evolutionary relationship among organisms (Roth et al., 2012).

Among all the genes that code for proteins in a genome of an organism, some genes are essential for the survival, growth, or reproduction of the organism compared to other genes. Various resource-intensive biological experiments such as single-gene knockout, conditional knockouts, RNA interference, and transposon mutagenesis have been carried out to identify essential genes (EGs) in microbes. Availability of this gene essentiality information for several microbes in public databases and a huge volume of genome sequences in the public domain have created an avenue for analysing base compositional features using machine learning-based algorithms. Results of this analysis can contribute towards understanding gene essentiality. Previous studies report use of gene and protein sequences (Campos et al., 2019), gene networks, protein-protein interactions (Azhagesan et al., 2018), metabolic networks (Plaimas et al., 2010), gene expression (Fan et al., 2017; Zhong et al., 2013), GO terms (Chen et al., 2017), gene evolutionary data (Wei et al., 2013), etc. for prediction of EGs. All of these works have employed features from a combination of two or more gene intrinsic or extrinsic data, or a combination of both intrinsic and extrinsic features (Acencio and Lemke, 2009; Aromolaran et al., 2020; Deng et al., 2010; Hwang et al., 2009; Lin et al., 2019). In this work we have tried to classify EGs and non-essential genes (NEGs) of a larger set of organisms – 35 bacterial strains using single-nucleotide composition of genes, a feature set which have not been reported so far for the prediction of EGs. Our analysis suggests that the single nucleotide compositional features may be useful in predicting gene essentiality in some bacteria species but not universally.

2 Materials and methods

2.1 Data collection and pre-processing

The database of essential genes (DEG) (http://www.essentialgene.org/, accessed 22 May 2021) (Zhang et al., 2004) hosts records of EGs of 66 bacterial strains. Of the 66 strains, detailed genome annotations are found for 35 strains in the NCBI database. Therefore, we have considered these 35 strains for detailed compositional analysis (Table 1). Of the 35 strains, 31 organisms are unique species; 20 bacteria belong to the phylum proteobacteria, nine bacteria belong to firmicutes, three belong to bacteroidetes, and one each from tenericutes, actinobacteria, and cyanobacteria. The majority of these 35 bacteria are pathogens that cause serious illnesses. For example, *Campylobacter jejuni* is responsible for food poisoning (Altekruse et al., 1999), Mycobacterium tuberculosis is the causative agent of tuberculosis (Smith, 2003), *Vibrio cholerae* the causative agent of cholera (Faruque et al., 1998). Few of them are commercially beneficial bacteria, such as *Bacillus thuringiensis* is used as a biological pesticide (Ibrahim et al., 2010), *Synechococcus elongatus* grows fast using sunlight, having biotechnological applications (Yu et al., 2015), especially for incorporating genetic modification. The genome size of these bacteria ranges from 963,879 base pairs (bp) to 6,723,972 bp, genome ($G + C$)% ranges from 26.60% to 68.40%, and reported EG% ranges from 1.97 % to 39.59%.

Table 1 Genomic features of the bacteria considered in this study

Sl. no.	*Bacteria name*	*Phylum*	*Genome size (bp)*	*Genome G + C%*	*Total no. of genes**	*No. of essential genes***
1	*Acinetobacter baumannii* ATCC 17978	Proteobacteria	4,335,793	38.90	3,810	615
2	*Bacillus thuringiensis* BMB171	Firmicutes	5,314,794	35.20	5,663	516
3	*Bacillus subtilis* 168	Firmicutes	4,215,606	43.51	4,226	271
4	*Bacteroides fragilis* 638R	Bacteroidetes	5,310,990	43.40	4,290	547
5	*Bacteroides thetaiotaomicron* VPI-5482	Bacteroidetes	6,293,399	42.90	4,778	325
6	*Brevundimonas subvibrioides* ATCC 15264	Proteobacteria	3,445,263	68.40	3,379	412
7	*Burkholderia cenocepacia* J2315	Proteobacteria	3,870,082	66.90	7,070	383
8	*Burkholderia pseudomallei* K96243	Proteobacteria	4,074,542	68.10	5,727	505
9	*Burkholderia thailandensis* E264	Proteobacteria	6,723,972	67.60	5,632	406
10	*Campylobacter jejuni jejuni* 81-176	Proteobacteria	1,616,554	30.50	1,653	384
11	*Campylobacter jejuni jejuni* NCTC 11168	Proteobacteria	1,641,481	30.50	1,572	166
12	*Escherichia coli* MG1655	Proteobacteria	4,639,675	50.80	4,146	296
13	*Francisella novicida* U112	Proteobacteria	4,016,947	32.50	1,721	392
14	*Francisella tularensis schu* S4	Proteobacteria	1,892,775	32.30	1,556	453
15	*Haemophilus influenzae* Rd KW20	Proteobacteria	1,830,138	38.20	1,658	642
16	*Mycobacterium tuberculosis* H37Rv	Actinobacteria	4,411,532	65.60	4,030	614
17	*Mycoplasma pulmonis* UAB CTIP	Tenericutes	963,879	26.60	783	310
18	*Porphyromonas gingivalis* ATCC 33277	Bacteroidetes	2,354,886	48.40	2,090	460
19	*Pseudomonas aeruginosa* PAO1	Proteobacteria	6,265,484	66.60	5,515	336
20	*Pseudomonas aeruginosa* UCBPP-PA14	Proteobacteria	6,264,404	66.30	5,892	335
21	*Rhodopseudomonas palustris* CGA009	Proteobacteria	5,459,213	65.00	4,874	522

Notes: *no. of genes reported in the NCBI database of genes.
**no. of EGs reported in the DEG.

Table 1 Genomic features of the bacteria considered in this study (continued)

Sl. no.	*Bacteria name*	*Phylum*	*Genome size (bp)*	*Genome G + C%*	*Total no. of genes**	*No. of essential genes***
22	*Salmonella enterica serovar* Typhi Ty2	Proteobacteria	4,791,961	52.10	4,714	358
23	*Salmonella enterica serovar* Typhimurium 14028S	Proteobacteria	4,870,265	52.20	5,315	105
24	*Salmonella typhimurium* LT2	Proteobacteria	4,857,432	52.20	4,458	230
25	*Shewanella oneidensis* MR-1	Proteobacteria	4,969,803	45.90	4,069	403
26	*Sphingomonas wittichii* RW1	Proteobacteria	5,382,261	67.90	4,850	535
27	Staphylococcus *aureus* NCTC 8325	Firmicutes	2,821,361	32.90	2,892	351
28	*Streptococcus agalactiae* A909	Firmicutes	2,127,839	35.60	1,906	317
29	*Streptococcus mutans* UA159	Firmicutes	2,030,921	36.80	1,960	197
30	*Streptococcus pneumoniae*	Firmicutes	2,038,615	38.57	2,238	244
31	*Streptococcus pyogenes* NZ131	Firmicutes	1,815,785	38.60	1,418	241
32	*Streptococcus sanguinis*	Firmicutes	2,388,435	43.40	2,270	218
33	*Streptococcus suis*	Firmicutes	2,096,309	41.30	2,041	361
34	*Synechococcus elongatus* PCC 7942	Cyanobacteria	2,695,903	55.40	2,422	682
35	*Vibrio cholerae* N16961	Proteobacteria	4,033,464	47.50	3,722	779

Notes: *no. of genes reported in the NCBI database of genes.
**no. of EGs reported in the DEG.

2.2 Determining the feature set

In our analysis, we have considered single-nucleotide compositional features of genes to predict EGs. Considering the physio-chemical properties discussed in the introduction section, we have calculated the following compositional features in this study: percentage of the occurrence of the nucleotides *A*, *T*, *G*, and *C*, length of the genes, (*A* + *T*)%, (*G* + *C*)%, *AT-skew*, *GC-skew*, *AG-skew*, *CT-skew*, *RY-skew*, *AC-skew*, *GT-skew*, and *KM-skew*. Mathematically, these features are defined as follows. For a nucleotide n with a count x_n in a sequence, nucleotide frequency f_n is defined as

$$f_n = \frac{x_n}{\sum_{n=\{A,T,C,G\}} x_n} \times 100 \tag{1}$$

For a genome sequence, (*A* + *T*)% and (*G* + *C*)%, are defined as

$$(A+T)\% = \frac{x_A + x_T}{\sum_{n=\{A,T,C,G\}} x_n} \times 100 \tag{2}$$

$$(G+C)\% = \frac{x_G + x_C}{\sum_{n=\{A,T,C,G\}} x_n} \times 100 \tag{3}$$

For a genome sequence, different skew values are defined as

$$AT_{skew} = \frac{x_A - x_T}{x_A + x_T} \tag{4}$$

$$GC_{skew} = \frac{x_G - x_C}{x_G + x_G} \tag{5}$$

$$AG_{skew} = \frac{x_A - x_G}{x_A + x_G} \tag{6}$$

$$CT_{skew} = \frac{x_C - x_T}{x_C + x_T} \tag{7}$$

$$AC_{skew} = \frac{x_A - x_C}{x_A + x_C} \tag{8}$$

$$GT_{skew} = \frac{x_G - x_T}{x_G + x_T} \tag{9}$$

$$RY_{skew} = \frac{(x_A + x_G) - (x_C + x_T)}{\sum_{n=\{A,T,C,G\}} x_n} \tag{10}$$

$$KM_{skew} = \frac{(x_G + x_T) - (x_A - x_C)}{\sum_{n=\{A,T,C,G\}} x_n} \tag{11}$$

2.3 Classification

We used seven machine learning-based classifiers, available in the sci-kit-learn library of Python – logistic regression (LR) (Peng et al., 2002), Gaussian Naïve Bayes (GNB) (Pérez et al., 2006), k-nearest neighbours (kNNs) (Baek and Sung, 2000), decision tree (DT) (Quinlan, 1986), random forest (RF) (Breiman, 2001), extreme gradient boosting (XGB) (Sheridan et al., 2016) and support vector machine (SVM) for the prediction of EGs.

2.4 Metrics for performance estimation

To estimate the performance of the classifiers, we considered *k*-fold cross-validation (5-fold in our case) score, precision, recall, F1-score, and AUC-ROC measures. Cross-validation is used to evaluate machine learning models which involves resampling

of data. Given a data sample, it is split into k equally distributed sub-samples. Of the k sub-samples, one sub-sample is kept for validating the model, and the remaining $k - 1$ sub-samples are used to train the model. Each sub-sample is used exactly once for validating the model for the k number of iterations (k-folds). Results from each of the folds are then averaged to give a single result. Precision is a performance metric that gives the ratio of the observations that are correctly predicted as positive to the total number of observations that are predicted as positive by the model. Recall is another performance metric that gives the ratio of observations correctly predicted as positive to all the observations in the actual class. The metric, F1-score gives a weighted average of precision and recall.

For a machine learning model to perform efficiently, it is desirable to reduce the number of features in the dataset (Cai et al., 2018). Redundant variables reduce the generalisation capability of a model and may also reduce the overall accuracy of the classifier. Moreover, large numbers of features increase the complexity of the model (Kotsiantis, 2011). So an efficient feature selection method is necessary to select the best features in the dataset that helps increase the prediction accuracy of the classifier. Input variables that have the strongest relationship with the target variables are selected. We have used the Boruta feature selection algorithm (Kursa and Rudnicki, 2010) after initial analysis, which was done without using any statistical feature selection algorithm.

2.5 Software used

All the nucleotide compositional features are computed using our program written in Python (version 3.9.4). We have used sci-kit-learn library (version 0.24.2) of Python for the classification that features various classification and regression algorithms. R (version 4.0.5) is used for plotting graphs and different statistical analyses.

2.6 Overview of the methodology used

We considered all the genes of an organism into two sets for compositional feature-based machine learning analysis. Genes reported as EGs in DEG were considered in the first set, and the remaining genes were grouped in the second set. As there was no information about the NEGs in DEG, we performed a one-class classification among genes in the second set to remove outliers (if any). We applied *one-class SVM* for outlier detection and discarded the outlier genes. The number of genes in the first set in all 35 organisms was comparatively less than the number of genes in the second set. So, to avoid any effect of skewed dataset on classification results, we performed a randomised sub-sampling among genes in the second set to extract the same number of genes as that of the first set and prepare two balanced sets of genes. First, the set of EGs and second, the set of remaining genes was considered the proxy of NEGs. Considering the base compositional features and seven machine learning-based classifier algorithms, we did a detailed analysis to understand to what extent essentiality can be predicted. For determining the classification accuracy, we applied a 5-fold cross-validation. The mean of the accuracy of the 5-folds cross-validation was considered as the final accuracy score; also, AUC score was used to estimate the performance of the classifiers. Other metrics such as precision, recall, F1-score, and support were also used to assess the performance of the classifiers. We have summarised the methodology in Figure 1.

Figure 1 Block diagram of the methodology used for machine learning-based analysis

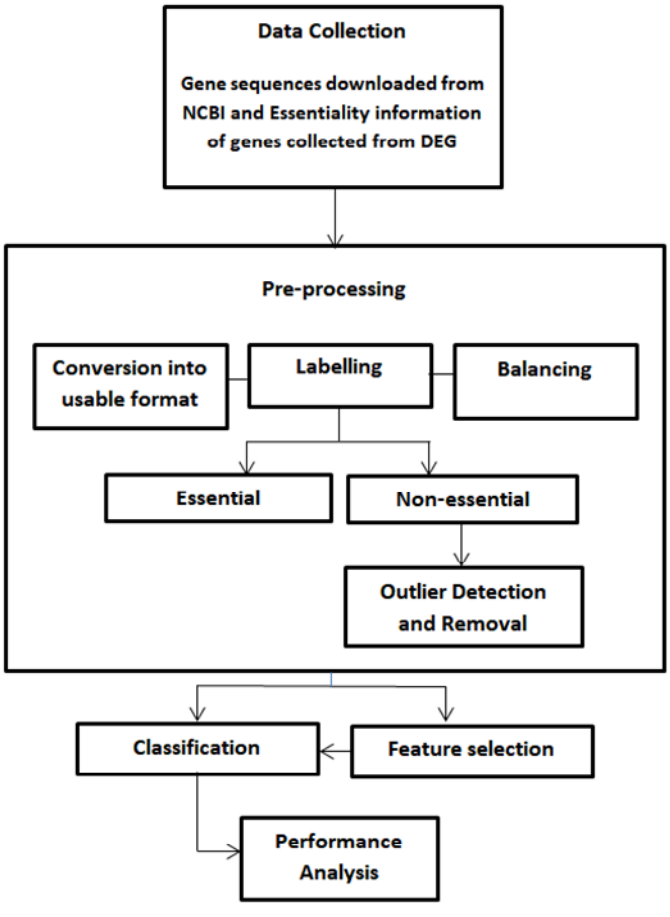


Figure 1 represents the step by step methodology we have applied, starting from data collection to the final analysis.

3 Results and discussion

3.1 Variable performance of single nucleotide compositional feature-based machine learning classifiers towards essentiality prediction

Assuming differential values single-nucleotide compositional features between the essential and NEG sets, we used several advanced machine learning-based classifiers and measured performance metrics such as 5-fold CV score, AUC, precision, recall and F1-score. The performance scores achieved by each of the classifiers vary to a great extent. It can be observed that most of the classifiers performed poorly in classifying essential and NEGs. Classification scores are less than 70% for the majority of the datasets. AUC scores greater than 70% obtained by LR achieved for 15 organisms, by GNB for 11 organisms, by kNNs for 3 organisms, by DT for 1 organisms, by RF for 13

organisms, by XGB for 12 organisms and by SVM for 14 organisms. Though LR achieved AUC greater than or equal to 70% for 14 organisms, it achieved a 5-fold cross-validation score greater than or equal to 70% for only two organisms, while RF achieved a 5-fold CV score of at least 70% for three organisms. The range of scores achieved for the performance metrics for each of the classifiers is shown in Table 2. Table 3 shows the number of organisms for which each classifier achieved AUC score and a 5-fold CV score greater than or equal to 70%. Results obtained for the performance metrics – 5-fold CV score, AUC, precision, recall, and F1-score are in Appendices 1, 2 and 3. Box plots of a range of 5-fold CV scores and AUC scores are presented in Figures 2 and 3, respectively. Figure 4 shows the ROC curves of the seven classifiers for four randomly picked organisms.

Table 2 Range of performance scores for the seven classifiers

	5-fold CV score	*AUC*	*Precision*	*Recall*	*F1-score*
LR	0.52–0.76	0.54–0.95	0.46–0.81	0.49–0.86	0.49–0.73
GNB	0.49–0.70	0.53–0.82	0.52–0.78	0.29–0.86	0.38–0.77
KNN	0.49–0.76	0.51–0.76	0.47–0.73	0.43–0.75	0.48–0.74
DT	0.50–0.68	0.49–0.76	0.42–0.77	0.48–0.72	0.47–0.73
RF	0.52–0.77	0.53–0.84	0.49–0.74	0.53–0.83	0.52–0.75
XGB	0.50–0.75	0.55–0.80	0.52–0.78	0.49–0.94	0.53–0.85
SVM	0.49–0.75	0.54–0.82	0.54–0.79	0.35–0.80	0.44–0.74

Table 2 shows the range of accuracy achieved by each of the seven classifiers for the performance metrics – 5-fold CV score, AUC, precision, recall, and F1-score.

Table 3 Number of organisms for which AUC scores and 5-fold CV scores greater than or equal to 70% achieved by the seven classifiers

Classifier	*AUC score >=70%*	*5-fold CV score >=70%*
LR	15	2
GNB	11	1
KNN	3	1
DT	1	0
RF	13	3
XGB	12	2
SVM	14	1

Table 3 gives the number of organisms for which AUC score and 5-fold CV score are greater than or equal to 70%. For each of the classifiers, we have a different score.

Figure 2 represents the range of 5-fold CV scores achieved for the 35 organisms by the seven classifiers. Each box in the box-plot represents the CV scores achieved by each of the seven classifiers.

Figure 3 represents the range of AUC scores achieved for the datasets of the 35 organisms by each of the seven classifiers. Each box in the box-plot represents the AUC scores achieved by each of the seven classifiers.

Figure 4 represents ROC curve and AUC scores for four randomly selected organisms.

Figure 2 Box-plot of 5-fold CV scores of the seven classifiers for the 35 organisms

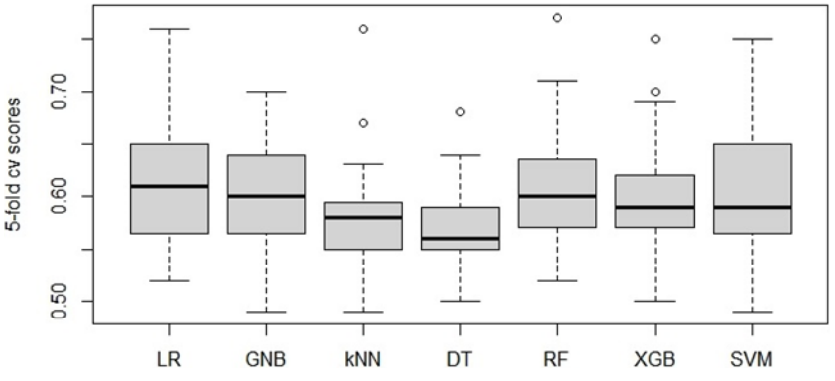
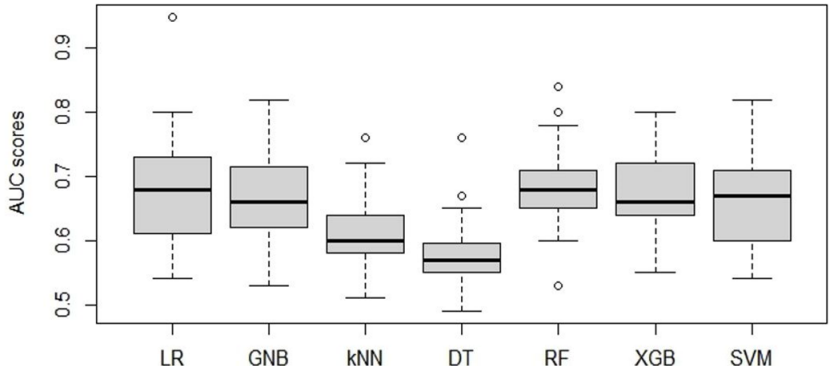


Figure 3 Box-plot of AUC scores achieved by the seven classifiers for the 35 organisms



3.2 Variable performance of classifiers across organisms

High variability of the classification scores across organisms can be seen. While the highest 5-fold CV score achieved by the RF classifier is 0.77 for the organism *Salmonella enterica serovar* Typhimurium 14028S, the lowest is 0.52 for *Haemophilus influenzae* Rd KW20. In case of the RF classifier, 5-fold CV scores greater than or at least equal to 0.70 was achieved for three organisms, *Bacteroides fragilis* 638R, *Sphingomonas wittichii* RW1, and *Salmonella enterica serovar* Typhimurium 14028S. For *Salmonella enterica serovar* Typhimurium 14028S, a 5-fold CV score is at least 70% for all classifiers. For Escherichia coli MG1655, 5-fold CV score is 0.65 for LR, 0.60 for GNB, 0.59 for kNNs, 0.56 for DT, 0.55 for RF, 0.59 for XGB, and 0.64 for SVM.

3.3 Feature selection is not able to improve the performance of the classifiers

After the initial analysis, we applied a feature selection algorithm called Boruta algorithm that is used as a wrapper around the RF classifier. The list of features selected by Boruta algorithm varies from organism to organism. For many of the microorganisms, all the 15 features are preferred, but the feature rankings vary across organisms, while for a few organisms like *Haemophilus influenzae* Rd KW20 and *Salmonella typhimurium* LT2 very

few features were selected. Classification scores after feature selection did not improve. A feature ranking for two organisms is shown in Figure 5. We have also plotted the mean feature rank values given by the algorithm across the 35 organisms (Figure 6), from which it can be seen that medians of all the features lie within a similar range. List of selected and rejected features for the 35 organisms, mean of feature importance for each dataset, and graphical representation of feature rankings are given in Appendices 4, 5 and 6, respectively.

Figure 4 ROC curves of four organisms based on the nucleotide compositional feature set (see online version for colours)

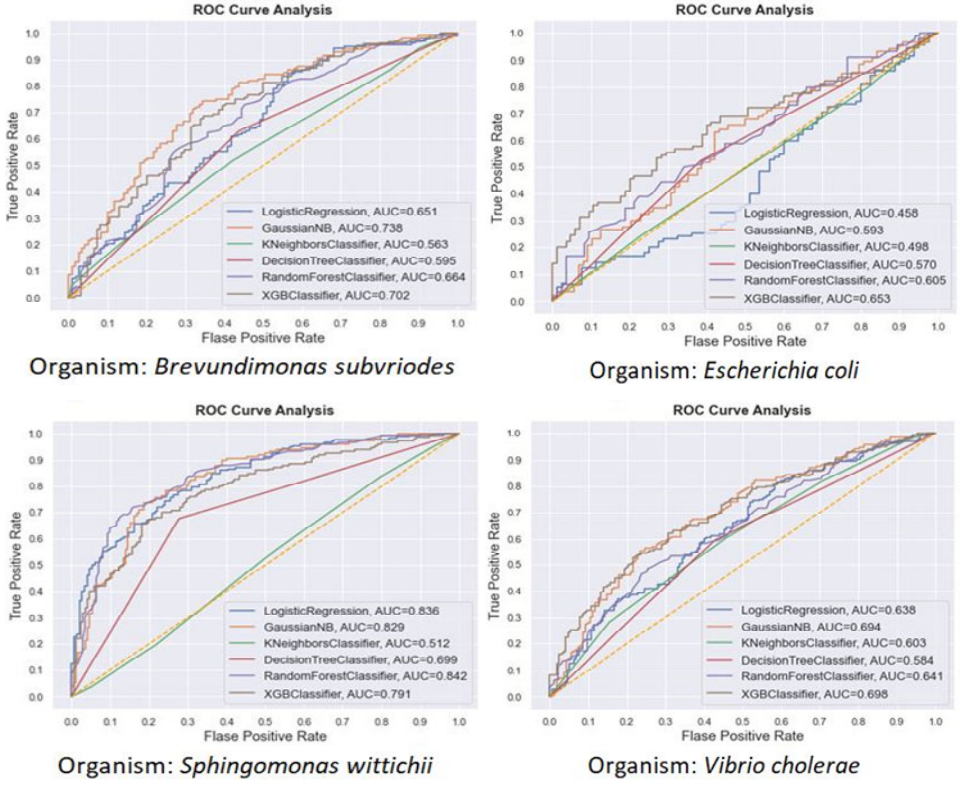
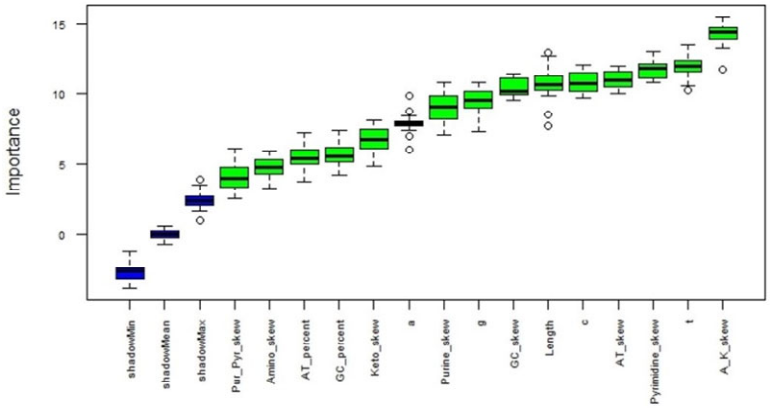


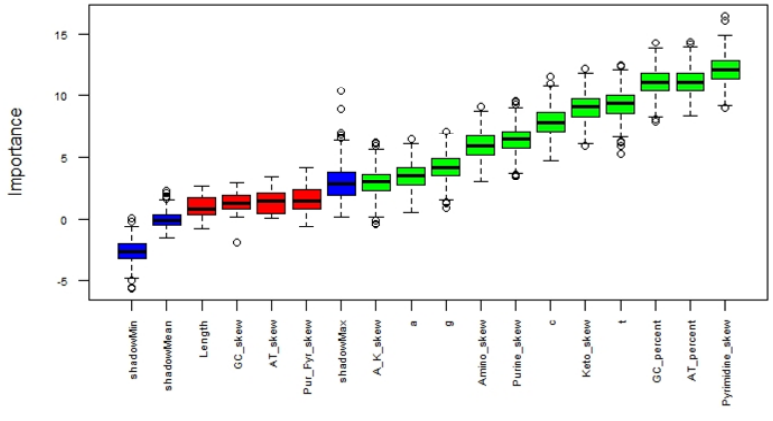
Figure 5 represents feature importance given by Boruta algorithm in graphical form using box plots. Each feature represents a box in the box plot. Green boxes represent the selected features, blue boxes represent the shadow attributes – ShadowMax, ShadowMean, and ShadowMin generated by the algorithm and red boxes represent the rejected features.

Figure 6 represents a box-plot of the mean values of the feature ranking given by Boruta algorithm for all the features across the 35 organisms. The X-axis denotes the features, and the Y-axis represents the mean values.

Figure 5 Feature ranking by Boruta algorithm for datasets of two organisms, (a) organism: *Brevundimonas subvibriodes* (b) organism: *Escherichia coli* (see online version for colours)

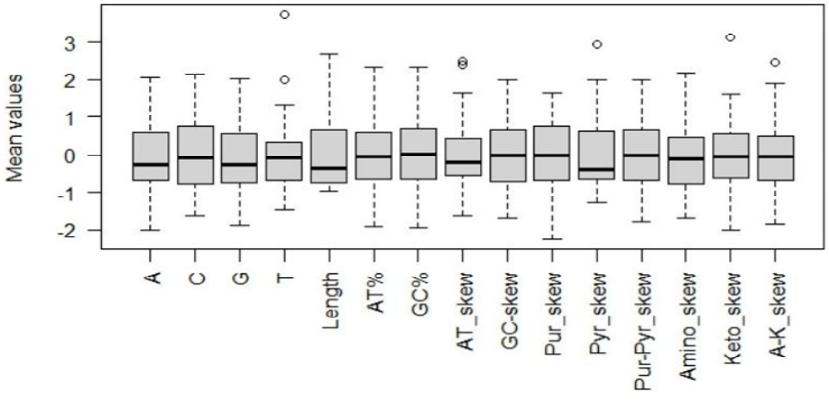


(a)



(b)

Figure 6 Mean of feature importance given by Boruta algorithm across 35 organisms



3.4 Similar compositional features distribution between essential and NEGs

To investigate the reason behind the low classification accuracy of our machine learning models we have created box plots of the features we considered for the classification. GC percentage values of essential and NEGs lied in an almost similar range (Figure 7). Moreover, medians of the box-plots of the other features, percentage of *A*, *T*, *G* and *C*, length of the genes, (*A* + *T*)%, AT-skew, GC-skew, AG-skew, CT-skew, RY-skew, AC-skew, GT-skew, KM-skew too lied in an almost similar range across essential and NEGs (Figures 8 and 9). Since there is no difference in the range of values across essential and NEGs for the feature set, we had considered the accuracy of the classifiers was very low.

Figure 7 Box plot of range of GC percentage in EG and NEG across the 35 bacteria (see online version for colours)

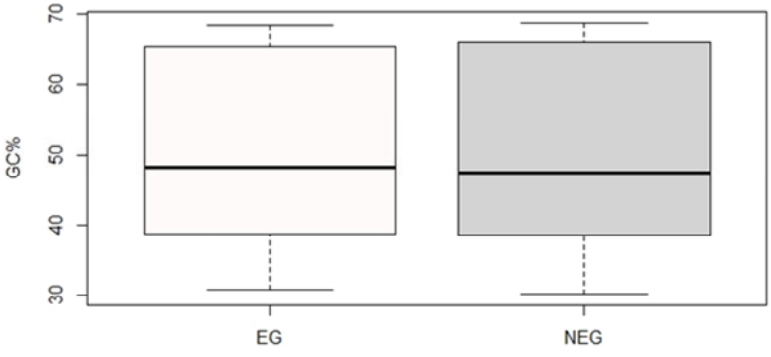


Figure 7 shows the range of GC% in EGs and NEGs of the 35 organisms. The range and median values are similar in EG and NEG.

Figure 8 represents a box plot of EG and NEG nucleotide composition in *Escherichia coli*. The X-axis represents the nucleotide composition in EG and NEG, and the Y-axis represents the percentage of the nucleotides. The range of nucleotide composition in NEGs is larger compared to EGs, but median values for the nucleotides in EGs and NEGs in each of the four features are almost equal.

Figure 8 Box plot of *A*, *C*, *G* and *T* percentage in EG and NEG in *Escherichia coli*

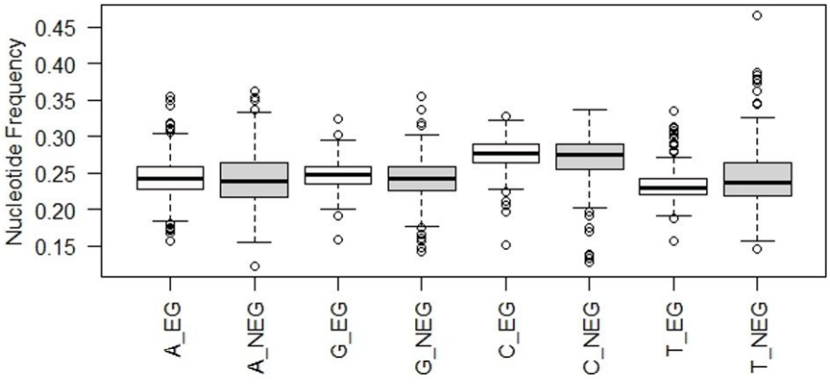
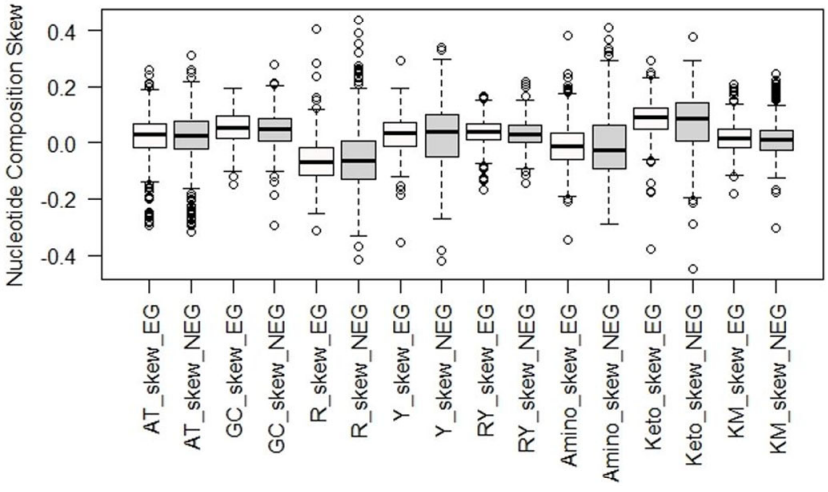


Figure 9 represents a box plot of nucleotide composition skew of EG and NEG in *E. coli*. X-axis represents six types of skews – AT-skew, GC-skew, Purine-skew, Pyrimidine-skew, RY-skew, Amino-skew, Keto-skew, KM-skew in EGs and NEGs, Y-axis represents the skew values. Observation: The median value for EG and NEG in each of the six skew measures is almost equal.

Figure 9 Box plot of AT-skew, GC-skew, Purine-skew, Pyrimidine-skew, RY-skew, Amino-skew, Keto-skew, KM-skew variation across EGs and NEGs in *Escherichia coli*



4 Conclusions

In this machine learning-based compositional feature analysis of bacterial genomes, the classifiers could achieve low to moderated classification accuracy. The accuracy scores were also found to vary from organism to organism. It was observed that the values of the features considered for the work lied in a similar range across essential and NEGs across organisms and therefore the classifiers could not distinguish between the two classes of genes. So, from this analysis, the single nucleotide composition-based feature set we had considered is not sufficient for classifying EGs and NEGs even though nucleotide composition has some contribution in gene essentiality prediction in some of the organisms. Future analysis considering higher-order compositional features might be helpful in this regard. Codon composition in the bacterial genome is known to be influenced by the expression level of the genes. It will be interesting to explore the relationship between gene essentiality and codon usage in the future.

Appendices/Supplementary materials are available on request by emailing the corresponding author.

References

Acencio, M.L. and Lemke, N. (2009) 'Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information', *BMC Bioinformatics*, Vol. 10, No. 1 [online] https://doi.org/10.1186/1471-2105-10-290.

Altekruse, S.F., Stern, N.J., Fields, P.I. et al. (1999) 'Campylobacter jejuni – an emerging foodborne pathogen', *Emerging Infectious Diseases*, Vol. 5, No. 1, pp.28–35, DOI: 10.3201/eid0501.990104.

Aromolaran, O., Beder, T., Oswald, M., Oyelade, J., Adebiyi, E. and Koenig, R. (2020) 'Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features', *Computational and Structural Biotechnology Journal*, Vol. 18, pp.612–621.

Azhagesan, K., Ravindran, B. and Raman, K. (2018) 'Network-based features enable prediction of essential genes across diverse organisms', *Plos One*, Vol. 13, No. 12, p.e0208722 [online] https://doi.org/10.1371/journal.pone.0208722.

Baek, S.J. and Sung, K.M. (2000) 'Fast K-nearest-neighbour search algorithm for nonparametric classification', *Electronics Letters*, Vol. 36, No. 21, pp.1821–1822.

Bohlin, J. (2008) 'Investigations of oligonucleotide usage variance within and between prokaryotes', *PLoS Computational Biology*, Vol. 4, No. 4, DOI: 10.1371/journal.pcbi.1000057.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Bulmer, M. (1991) 'The selection-mutation-drift theory of synonymous codon usage', *Genetics*, Vol. 129, No. 1, pp.897–907.

Cai, J., Luo, J., Wang, S. et al. (2018) 'Feature selection in machine learning: a new perspective', *Neurocomputing*, Vol. 300, No. 1, pp.70–79.

Campos, T.L., Korhonen, P.K., Gasser, R.B. and Young, N.D. (2019) 'An evaluation of machine learning approaches for the prediction of essential genes in eukaryotes using protein sequence-derived features', *Computational and Structural Biotechnology Journal*, Vol. 17, No. 1, pp.785–796.

Chen, L., Zhang, Y.H., Wang, S., Zhang, Y., Huang, T. and Cai, Y.D. (2017) 'Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways', *Plos One*, Vol. 12, No. 9, p.e0184129 [online] https://doi.org/10.1371/journal.pone.0184129.

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A.A., Hassett, D.J. and Lu, L.J. (2010) 'Investigating the predictability of essential genes across distantly related organisms using an integrative approach', *Nucleic Acids Research*, Vol. 39, No. 3, pp.795–807 [online] https://doi.org/10.1093/nar/gkq784.

Fan, Y., Tang, X., Hu, X., Wu, W. and Ping, Q. (2017) 'Prediction of essential proteins based on subcellular localization and gene expression correlation', *BMC Bioinformatics*, Vol. 18, No. S13 [online] https://doi.org/10.1186/s12859-017-1876-5.

Faruque, S.M., Albert, M.J. and Mekalanos, J.J. (1998) 'Epidemiology, genetics, and ecology of toxigenic *Vibrio cholerae*', *Microbiology and Molecular Biology Reviews*, Vol. 62, No. 4, DOI: 10.1128/MMBR.62.4.1301-1314.1998.

Forsdyke, D.R. and Mortimer, J.R. (2000) 'Chargaff's legacy', *Gene*, Vol. 261, No. 1, pp.127–137.

Francino, M.P. and Ochman, H. (1997) 'Strand asymmetries in DNA evolution', *Trends Genet.*, Vol. 13, No. 1, pp.240–245.

Frank, A.C. and Lobry, J.R. (1999) 'Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms', *Gene*., Vol. 238, No. 1, pp.65–77.

Hildebrand, F. (2010) 'Evidence of selection upon genomic GC-content in bacteria', *PLoS Genetics*, Vol. 6, No. 9, DOI: 10.1371/journal.pgen.1001107.

Hwang, Y.C., Lin, C.C., Chang, J.Y., Mori, H., Juan, H.F. and Huang, H.C. (2009) 'Predicting essential genes based on network and sequence analysis', *Molecular BioSystems*, Vol. 5, No. 12, p.1672 [online] https://doi.org/10.1039/b900611g.

Ibrahim, M.A., Griko, N., Junker, M. et al. (2010) '*Bacillus thuringiensis* a genomics and proteomics perspective', *Bioengineered Bugs*, Vol. 1, No. 1, pp.31–50.

Karkas, J.D., Rudner, R. and Chargaff, E. (1968) 'Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerases', *Proc. Natl Acad. Sci.*, USA, Vol. 60, pp.915–20.

Kotsiantis, S. (2011) 'Feature selection for machine learning classification problems: a recent overview', *Artificial Intelligence Review*, Vol. 42, No. 1, p.157.

Kursa, M.B. and Rudnicki, W.R. (2010) 'Feature selection with the Boruta package', *Journal of Statistical Software*, Vol. 36, No. 11, pp.1–13.

Langille, M.G.I., Hsiao, W.W.L. and Brinkman, F.S.L. (2008) 'Evaluation of genomic island predictors using a comparative genomics approach', *BMC Bioinformatics*, Vol. 9, DOI: 10.1186/1471-2105-9-329.

Lin, Y., Zhang, F.Z., Xue, K., Gao, Y.Z. and Guo, F.B. (2019) 'Identifying bacterial essential genes based on a feature-integrated method', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 16, No. 4, pp.1274–1279 [online] https://doi.org/10.1109/tcbb.2017.2669968.

Lobry, J.R. and Sueoka, N. (2002) 'Asymmetric directional mutation pressures in bacteria', *Genome Biol.*, Vol. 3, No. 1, pp.0058.1–0058.14.

Muto, A. and Osawa, S. (1987) 'The guanine and cytosine content of genomic DNA and bacterial evolution (biased mutation pressure/codon usage/neutral theory)', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 84, pp.166–169.

Nikolaou, C. and Almirantis, Y. (2005) 'A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species', *Nucleic Acid Res.*, Vol. 33, No. 1, pp.6816–6822.

Nikolaou, C. and Almirantis, Y. (2006) 'Deviation from Chargaff's second parity rule in organellar DNA insights into the evolution of organellar genomes', *Gene*., Vol. 381, No. 1, pp.34–41.

Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M. (2002) 'An introduction to logistic regression analysis and reporting', *Journal of Educational Research*, Vol. 96, No. 1, pp.3–14, DOI: 10.1080/00220670209598786.

Pérez, A., Larrañaga, P. and Inza, I. (2006) 'Supervised classification with conditional Gaussian networks: increasing the structure complexity from Naive Bayes', *International Journal of Approximate Reasoning*, Vol. 43, No. 1, pp.1–25.

Plaimas, K., Eils, R. and König, R. (2010) 'Identifying essential genes in bacterial metabolic networks with machine learning methods', *BMC Systems Biology*, Vol. 4, No. 1 [online] https://doi.org/10.1186/1752-0509-4-56.

Quinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, Vol. 1, No. 1, pp.81–106.

Raghavan, R., Kelkar, Y.D. and Ochman, H. (2012) 'A selective force favoring increased G + C content in bacterial genes', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109, No. 36, pp.14504–14507, DOI: 10.1073/pnas.1205683109.

Rocha, E.P. (2004) 'Codon usage bias from tRNA's point of view, redundancy, specialization, and efficient decoding for translation optimization', *Genome Res.*, Vol. 14, No. 1, pp.2279–2286.

Rocha, E.P.C. and Danchin, A. (2002) 'Base composition bias might result from competition for metabolic resources', *Trends in Genetics*, Vol. 18, No. 6, pp.291–294.

Rocha, E.P.C., Danchin, A. and Viari, A. (1999) 'Universal replication biases in bacteria', *Mol. Microbiol.*, Vol. 32, No. 1, pp.11–16.

Roth, A., Anisimova, M. and Cannarozzi, G.M. (2012) 'Measuring codon usage bias, 189–217', in Cannarozzi, G. and Schneider, A. (Eds.): *Codon Evolution: Mechanisms and Models*, Oxford University Press, Inc., New York, NY.

Rudner, R., Karkas, J.D. and Chargaff, E. (1969) 'Separation of microbial deoxyribonucleic acids into complementary strands', *Proc. Natl. Acad. Sci.*, USA, Vol. 63, pp.152–159.

Satapathy, S.S., Dutta, M. and Ray, S.K. (2010) 'Variable correlation of genome GC% with transfer RNA number as well as with transfer RNA diversity among bacterial groups: α-*Proteobacteria* and *Tenericutes* exhibit strong positive correlation', *Microbiological Research*, Vol. 165, No. 3, pp.232–242, DOI: 10.1016/j.micres.2009.05.005.

Satapathy, S.S., Powdel, B.R., Buragohain, A.K. et al. (2016) 'Discrepancy among the synonymous codons with respect to their selection as optimal codon in bacteria', *DNA Research*, Vol. 23, No. 5, pp.441–449.

Sen, P., Waris, A. and Ray, S.K. (2019) 'A web portal to calculate codon adaptation index (CAI) with organism specific reference set of high expression genes for diverse bacteria species', in Tomar, G.S. et al. (Eds.): *International Conference on Intelligent Computing and Smart Communication*, Springer Nature, Singapore.

Sharp, P.M. and Li, W.H. (1987) 'The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Res.*, Vol. 15, No. 3, pp.1281–1295.

Sharp, P.M., Emery, L.R. and Zeng, K. (2010) 'Forces that influence the evolution of codon bias', *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, Vol. 365, No. 1, pp.1203–1212.

Sheridan, R.P., Wang, W.M., Liaw, A. et al. (2016) 'Extreme gradient boosting as a method for quantitative structure-activity relationships', *Journal of Chemical Information and Modeling*, Vol. 56, No. 12, pp.2353–2360.

Smith, I. (2003) 'Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence', *Clinical Microbiology Reviews*, Vol. 16, No. 3, pp.463–496, DOI: 10.1128/CMR.16.3.463-496.2003.

Srividhya, K.V., Alaguraj, V., Poornima, G. et al. (2007) 'Identification of prophages in bacterial genomes by dinucleotide relative abundance difference', *Plos One*, Vol. 2, No. 11, DOI: 10.1371/journal.pone.0001193.

Sueoka, N. (1995) 'Intra-strand parity rules of DNA base composition and usage biases of synonymous codons', *J. Mol. Evol.*, Vol. 40, No. 1, pp.318–325.

Wei, W., Ning, L. W., Ye, Y.N. and Guo, F.B. (2013) 'Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny', *Plos One*, Vol. 8, No. 8, p.e72343 [online] https://doi.org/10.1371/journal.pone.0072343.

Yakovchuk, P., Protozanova, E. and Kamenetskii, M.D.F. (2006) 'Base-stacking and base-pairing contributions into thermal stability of the DNA double helix', *Nucleic Acids Research*, Vol. 34, No. 2, pp.564–574, DOI: 10.1093/nar/gkj454.

Yu, J., Liberton, M., Cliften, P.F. et al. (2015) 'Synechococcus elongatus UTEX 2973, a fast growing cyanobacterial chassis for biosynthesis using light and CO_2', *Scientific Reports*, Vol. 5, No. 1, pp.1–10.

Zhang, R., Ou, H.Y. and Zhang, C.T. (2004) 'DEG: a database of essential genes', *Nucleic Acids Research*, Vol. 32, Database Issue, DOI: 10.1093/nar/gkh024.

Zheng, H. and Wu, H. (2010) 'Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species', *BMC Bioinformatics*, Vol. 11, DOI: 10.1186/1471-2105-11-S11-S7.

Zhong, J., Wang, J., Peng, W., Zhang, Z. and Pan, Y. (2013) 'Prediction of essential proteins based on gene expression programming', *BMC Genomics*, Vol. 14, No. S4 [online] https://doi.org/10.1186/1471-2164-14-s4-s7.