Exploring the relationship between fractal features and bacterial essential genes*

Yong-Ming Yu(余永明)[1], Li-Cai Yang(杨立才)[1,†], Qian Zhou(周茜)[2,3], Lu-Lu Zhao(赵璐璐)[1], and Zhi-Ping Liu(刘治平)[1]

[1] *Department of Biomedical Engineering, Shandong University, Jinan* 250061, *China*
[2] *Province-Ministry Joint Key Laboratory of Electromagnetic Field and Electrical Apparatus Reliability, Hebei University of Technology, Tianjin* 300130, *China*
[3] *Department of Biomedical Engineering, Hebei University of Technology, Tianjin* 300130, *China*

(Received 19 December 2015; revised manuscript received 27 February 2016; published online 25 May 2016)

Essential genes are indispensable for the survival of an organism in optimal conditions. Rapid and accurate identifications of new essential genes are of great theoretical and practical significance. Exploring features with predictive power is fundamental for this. Here, we calculate six fractal features from primary gene and protein sequences and then explore their relationship with gene essentiality by statistical analysis and machine learning-based methods. The models are applied to all the currently available identified genes in 27 bacteria from the database of essential genes (DEG). It is found that the fractal features of essential genes generally differ from those of non-essential genes. The fractal features are used to ascertain the parameters of two machine learning classifiers: Naïve Bayes and Random Forest. The area under the curve (AUC) of both classifiers show that each fractal feature is satisfactorily discriminative between essential genes and non-essential genes individually. And, although significant correlations exist among fractal features, gene essentiality can also be reliably predicted by various combinations of them. Thus, the fractal features analyzed in our study can be used not only to construct a good essentiality classifier alone, but also to be significant contributors for computational tools identifying essential genes.

Keywords: fractal features, bacteria, essential gene, machine learning

PACS: 05.45.Df, 87.14.G **DOI:** 10.1088/1674-1056/25/6/060503

1. Introduction

Essential genes are a minimal set of genes that are considered to be indispensable for the survival of an organism under most favorable conditions.[1] Identifying and predicting gene essentiality for the survival of an organism have both theoretical and practical significance in genome biology and medicine. Because essential genes are indispensable for cell viability, their protein products are attractive drug targets for new antimicrobials. A designed drug that disrupts the function of such a protein will kill the pathogens.[2,3] More broadly, essential genes provide an insight into understanding the origin of life and evolutionary relationships among species, and they are the foundation for revealing the smallest requirements for sustaining a life form, and it is critical for synthetic biology.[4,5]

Today, systematic genome-wide interrogations of essential genes have been conducted by ribonucleic acid (RNA) interference,[6] gene knockouts,[7] and transposon mutagenesis.[8] However, there are several limitations to these experimental methods: they need massive investments in time and resource; based on large-scale experiment, the speed for them to identify essential gene falls far behind that for genome sequence; moreover, the identification results may be flexible, due to different experimental conditions and criteria.[9,10] With a fast-increasing availability of genome sequence and other functional genomics data, a highly efficient computational method becomes an appealing option for circumventing those large-scale experimental methods.

Substantial efforts have been made to explore the features associated with gene essentiality, with the aim of helping design feature-based computational methods. Sequence-based features which quantitatively measure the conservative property of genes have been considered as the most powerful features for gene essentiality identification. For instance, phyletic retention,[11] domain enrichment,[12] and number of paralogs[13] which were introduced to weight the evolutionary conservation of an individual gene or local gene sequence, have been found to be very predictive of essentiality. On the other hand, some studies exist that point out that in bacteria, a higher expression level is connected with slower evolutionary speed of gene sequence.[14] Codon usage bias,[15] leading strand,[16] and fluctuation of gene expression,[17] which is generally used to measure the level of gene expression, have been clearly shown to be associated with essentiality. Moreover, Lin *et al.*, also have suggested that the essential genes with important functions such as translation, transcription, replication, etc. are more crucial and prevailing at lead-

*Project supported by the Shandong Provincial Natural Science Foundation, China (Grant No. ZR2014FM022).
†Corresponding author. E-mail: yanglc@sdu.edu.cn
© 2016 Chinese Physical Society and IOP Publishing Ltd

ing strand than those non-essential genes from other function subcategories.[18] Hawoong *et al.* have illustrated that in contrast to essential genes, the fluctuation of non-essential genes expression tend to be more wide.[19] Besides, other sequence-based features such as subcellular location,[20] ORF length,[21] and upstream size[22] each can be used as an index to measure the essentiality of genes as well.

In fact, genes and their protein products in cell are not isolated. They interact with each other and constitute a biological network to accomplish the biological functions of cells. While sequence-base features above extract the information from individual genes, topology-based feature which is introduced to value the properties of protein-protein interaction (PPI) networks is mainly of interest in the information about interacting protein pairs. Based on that, a group of studies have shown that topology-based features such as degree-like features, between-like features, closeness-like features and subgraph-like features, etc. are powerful and prevailing to measure the relationship between essentiality and topological properties of protein in PPI networks.[23] In addition to the features detailed above, some studies also found respectively the biological process annotation,[24] hydrophobicity of protein,[25] and synthetic accessibility,[26] etc. to be important contributors for predicting genes essentiality. Based on the integration of the features detailed above, many computational techniques have been introduced to predict essential genes.[11,19,24,27] However, there are some practical limitations to these computational methods: they are restricted to few specific model organisms, lack easily available powerful features, and perform poorly in classification accuracy. Consequently, exploring features associated with essentiality is still crucial for designing computational tools.

In recent years, some interest has been shown in applying the methods of fractal theory for analyzing gene sequences and protein sequences.[28–30] Our previous studies have also found that gene essentiality tends to be more chaotic and less structured and the fractal features, Hurst exponents and information dimension can quantitatively assess those regular complexity.[31,32] Here, while individually assessing the connection to essentiality, we also focus on measuring the relationship between those fractal features and essentiality at integrative context. All the currently available identified essential genes and non-essential genes from the database of essential genes (DEG) are analyzed. Subsequently, statistic checkout analysis is carried out to investigate the correlation and difference in fractal feature between essential genes and non-essential genes. By combining fractal features in different categories, two machine learning algorithms are introduced to calculate the average area under the curve (AUC) to quantitatively weight the predictive utility of fractal features.

2. Material and methods

2.1. Data resource

DEG (http://www.essentialgene.org/) was established in 2004. With the continuous update, DEG can provide all currently available identified essential and non-essential genes and proteins of bacteria archaea and eukaryotes. A total of 27 bacteria (30 records) with their essential proteins and non-essential proteins are collected from DEG 10. The information about analysis objects is listed in Table 1.

Table 1. Information about the analysis objects. In the table, the superscript "$*$" indicate that these objects are protein data obtained from the same strain of the same bacterium, identified under different conditions or by different methods. the superscript "Δ" denotes that these objects are protein data obtained from the same bacterium, but from different strains, identified under different conditions or by different methods.

Number	Analysis objects	Essential gene /protein number	Non-essential gene /protein number
1	*B. subtilis* 168	271	3955
2	*V. cholera* N16961	779	2943
3	*F. novicida U*112	392	1329
4	*A. baylyi* ADP1	499	2594
5	*M. genitalium* G37	381	94
6	*M. pulmonis* UAB CTIP	310	322
7	*S. sanguinis*	218	2052
8	*P. gingivalis* ATCC 33277	463	1627
9	*B. thetaiotaomicron* VPI-5482	325	4453
10	*B. thailandensis* E264	406	5226
11	*S. wittichii* RW1	535	4315
12	*S. oneidensis* MR-1	402	1103
13	*C. jejuni* NCTC 11168	228	1395
14	*C. crescentus*	480	3224
15	*H. pylori* 26695	323	1135
16	*H. influenzae* Rd KW20	642	512
17	*E. coli* MG1655 I^{*1}	609	2923
18	*E. coli* MG1655 II^{*1}	296	4077
19	*S. serovar Typhimurium* LT2$^{\Delta 1}$	230	4228
20	*S. serovar Typhimurium* 14028S$^{\Delta 1}$	105	5210
21	*S. serovar Typhimurium* SL1344$^{\Delta 1}$	353	4035
22	*S. enterica serovar Typhi* $^{\Delta 2}$	353	4005
23	*S. enterica serovar Typhi* Ty2$^{\Delta 2}$	358	3906
24	*M. tuberculosis* H37Rv^{*2}	614	2552
25	*M. tuberculosis* H37Rv II^{*2}	771	3171
26	*M. tuberculosis* H37Rv III^{*2}	678	3070
27	*S. aureus* N315$^{\Delta 3}$	302	2281
28	*S. aureus* NCTC 8325$^{\Delta 3}$	351	2541
29	*P. aeruginosa* UCBPP-PA14$^{\Delta 4}$	335	960
30	*P. aeruginosa* PAO1$^{\Delta 4}$	117	5454

2.2. Topological entropy of gene sequences

Topological entropy is a measure of complex regulation of a gene sequence. For a given gene sequence $x(s)$, N corresponds to the length of $x(s)$ and n is defined as a unique integer

by the following equation:

$$4^n + n - 1 \leq N < 4^{n+1} + (n+1) - 1. \quad (1)$$

Based on deferent values of n, the complexity function C_x is defined as

$$C_x(n) = |\{m : |m| = n \text{ and } m \text{ appears as a subword of } x\}|, \quad (2)$$

where C_x represents the number of different n-length subwords (overlaps allowed) that appear in $x(s)$.

Then for x^{4^n+n-1} the first $4^n + n - 1$ letters of x, the definition of topological entropy of the finite sequence is

$$H_{\text{top}}(x) = \frac{\log_4 \left(C_{x^{4^n+n-1}}(n)\right)}{n}, \quad (3)$$

where $H_{\text{top}}(x)$ is the topological entropy of $x(s)$.

2.3. Chaos game representation of gene sequences

Chaos game representation (CGR) is a scale-independent representation method for genomic sequences proposed by Jeffrey in 1990.[33] In the CGR representation of a gene sequence $x(s)$, the four bases A, T, C, and G correspond to the four vertices of a square $[0,1] \times [0,1]$. The first point is plotted halfway between the center of the square and the vertex corresponding to the first base of the sequence $x(s)$, and the i-th point of the plot is placed halfway between the $(i-1)$-th point and the vertex corresponding to the i-th base. In this way, the CGR points are generated by an iterated function system defined by the following equation:

$$P_i = 0.5\,(P_{i-1} + S_i), \quad 1 \leq i \leq N, \quad (4)$$

where P_i is the coordinates of one point in the CGR square, corresponds to the i-th element of transformed sequences $x(s)$, and $P_0 = (0,0)$; S_i corresponds to the vertex coordinates of this transformed element.

2.4. Information dimension of gene CGR representation

Information dimension is an index for characterizing fractal patterns or sets by quantifying their complexity as a ratio of the change in detail to the change in scale. We divide the CGR pattern into ε^2 square grids of equal size, and then the side length of each grid is $1/\varepsilon$. The numbers of non-empty grids are denoted by $Z(\varepsilon)$. Dividing the number of points falling in the i-th grid by the total point number in the CGR square yields a probability p_i for the i-th grid. Information function and information dimension for the CGR are respectively defined as

$$I(\varepsilon) = -\sum_{i=1}^{Z(\varepsilon)} p_i \log p_i, \quad (5)$$

$$D_I = \lim_{1/\varepsilon \to 0} \frac{I(\varepsilon)}{\log(1/\varepsilon)}. \quad (6)$$

The information function $I(\varepsilon)$ in a range of $\log(1/\varepsilon)$ has a scaling region, where ε is an integer from 2 to 10. The information dimension D_I can be obtained from the slope in the scaling region.

2.5. Hurst exponent of gene time series

The Hurst exponent is the measure of the smoothness of fractal time series based on the asymptotic behavior of the rescaled range of the process. Here, we use the method proposed by Milan Randic to convert the CGR gene map into a time series and then, estimate the Hurst exponent of the times series of gene sequence based on rescaled range (R/S) analysis method.[34] It involved the following basic steps.

For a given gene sequence $x(s)$, F is a transformed times series over a total duration N. For a deterministic integer τ, the cumulative total at each point at different times is defined as

$$\Gamma_{\tau,k} = \sum_{i=1}^{k} (F_i - \mu_\tau), \quad 0 < k \leq \tau, \quad (7)$$

where F_i is the value of the time series at time i, μ_τ is the mean over the whole data set given by

$$\mu_\tau = (1/\tau) \sum_{i=1}^{\tau} F_i. \quad (8)$$

The range R of Γ is given by

$$R_\tau = \text{Max}\left(\Gamma_{\tau,k}\right) - \text{Min}\left(\Gamma_{\tau,k}\right). \quad (9)$$

The standard deviation (SD) of the values over the whole data set is given as

$$S_\tau = \sqrt{(1/\tau) \sum_{i=1}^{\tau} (F_i - \mu_\tau)^2}. \quad (10)$$

The rescaled range is given by R/S. The Hurst exponent is estimated by plotting the values of $\log(R/S)$ versus $\log\ \tau$, $(4 \leq \tau \leq 0.5N)$. The slope of the best fitting line gives the estimate of the Hurst exponent.

For a protein sequence, the detailed HP model[35] is used to divide a protein sequence into four sets: non-polar set, negative polar set, uncharged polar set, and positive polar set. Then, each protein sequence could be transformed into a new sequence composed of only four types of characters, like a gene sequence. The way we compute the fractal features of protein sequences is just the same as that we compute the fractal features of gene sequences. In such a way, gene sequence is represented by information dimension of DNA (ID), Hurst exponent of DNA (HD), topological entropy of DNA (TD), and the feature vector of protein sequence contains the information dimension of protein (IP), Hurst exponent of protein (HP) and topological entropy of protein (TP).

2.6. Training and testing set construction and classifier design

For most bacteria, there are much more non-essential genes than essential genes and they are classifications of highly imbalanced datasets. We solve the class imbalance problem through re-sampling, where we first randomly select semi-essential and non-essential genes as two sets, keep all genes of smaller set, and then from other sets randomly pick up the same number of genes as that of the set with fewer genes. These extracted essential and non-essential genes constitute the training sets and accordingly, all the remaining gene data are part of testing set. Naïve Bayes and Random Forest classifier are used to construct classifier model. For each bacterium, we train our classifier by using selected training set and predict the probability of essentiality for each gene in testing set by using the trained classifiers. On the basis of the true gene label and the probability of essentiality, we calculate the AUC of prediction to evaluate the classification performance. Each time, we repeat the random process 100 times, and then obtain the average AUC and its associated SD to assess the accuracy and robustness of our model, which is appropriate for the class unbalance problem.

3. Results

3.1. Assessing predictive utility for essentiality in individual fractal feature

For all 30 analysis objects, these fractal features of all essential and non-essential genes are calculated and then, Student's T-test analysis is utilized to explore the difference of each fractal feature between essentiality and non-essentiality. Figure 1 shows the results of the mean and significant level of each fractal feature for all essential and non-essential genes in S. sanguinis. For gene and protein sequences, the average information dimensions and topological entropies of essentialities are higher than those of non-essentialities and the results of Hurst exponents shift towards the opposite direction. There are significant differences in mean of fractal feature between essentiality and non-essentiality. It is suitable for most analysis objects. It suggests that essential genes are less edited and have weaker long-range correlation than nonessential genes.

In addition to exploring the difference between two kinds of genes, two classifier models are developed to further access the predictive power of individual fractal feature. Figure 2 shows the average AUC score of essentiality prediction in 27 bacteria based on Naïve Bayes and Random Forest classifier respectively. As shown in Fig. 2, the average AUC scores of six fractal features of all 27 bacteria exceed 0.624 and 0.625 for Naïve Bayes and Random Forest respectively, and some bacteria are above 0.7. It is also found that there are above 75% analysis objects whose prediction performance using information dimension outperforms using Hurst exponent and topological entropy, and nevertheless there is no obvious change between using Hurst exponent and using topological entropy.

On the other hand, the trend exists among all fractal features that the average classification accuracies using 100 balanced dataset from *Mycobacterium tuberculosis* H37Rv, *Vibrio cholera* N16961, and *Salmonella enterica serovar Typhimurium* 14028S are generally higher than those of others, improving 20%. Consequently, we propose that these fractal features are highly powerful annotation to make essentiality prediction and a better predictive performance is obtained by information dimension.

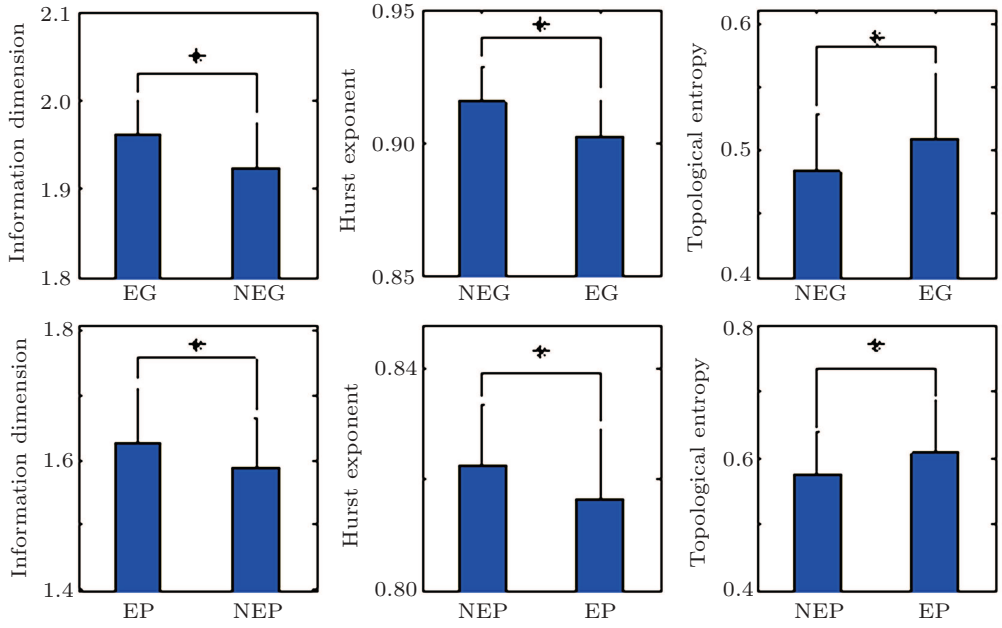


Fig. 1. (color online) Mean values and SDs of fractal features of essentiality and non-essentialirty in *S. sanguinis*. $*$ refers to significant difference existing at the 5% level. NEG: non-essential genes, EG: essential genes, NEP: non-essential proteins, and EP: essential proteins.

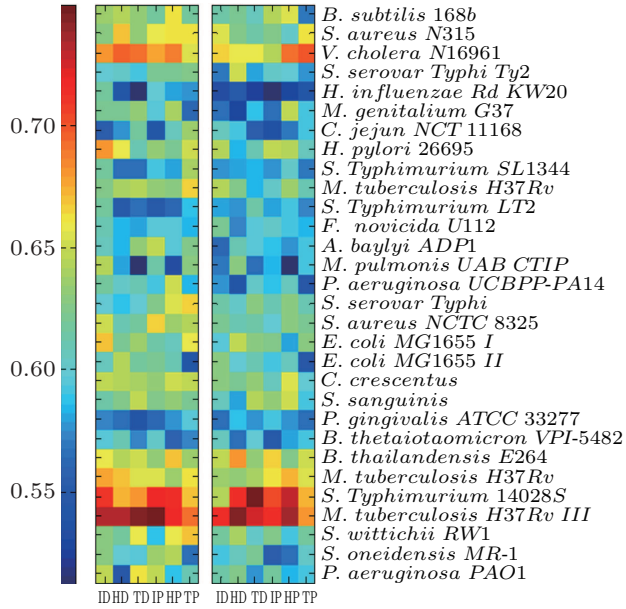


Fig. 2. (color online) Average AUC scores of essentiality prediction in 27 bacteria using Naïve Bayes (left) and Random Forest (right) classifier respectively. The exact values of this figure are provided in Supplement 1.

3.2. Assessing predictive utility for essentiality in integrative strategies

3.2.1. Correlation among fractal features

Since these fractal features are confirmed to be predictive to essentiality, we are interested in how many fractal features are independent among these fractal features. The Spearman correlation coefficients between pairs of those fractal features of all 27 bacteria are calculated and shown in Fig. 3. There are weak negative correlations between Hurst exponent and information dimension, and between Hurst exponent and topological entropy. Instead, much stronger positive correlation exists between information dimension and topological entropy. It is suitable for gene sequences and protein sequence. It can also be found that the features of gene sequence weakly relate to those of protein sequence and their average correlation coefficients range from -0.25 to 0.25. These results for 27 bacteria remain stable, which probably provides facilitated access for exploring genetic structure.

3.2.2. Prediction of essential gene base on diverse combination of fractal features

For 27 bacteria, Naïve Bayes and Random Forest classifiers are employed to make prediction of essentiality by integrating fractal features in three strategies: fractal features of gene sequences, fractal features of protein sequences, and all fractal features. The average AUC scores of using Naïve Bayes classifier are shown in Fig. 4 and Random Forest classifier is displayed in Fig. 5. From both figures, we find that the average AUC of all bacteria is above 0.62 in either of features integrating strategies and those of some analysis objects are all around 0.7. Specifically, of all the 30 analysis objects, appropriately 83%, 86%, and 93% obtain average AUC scores above 0.6 for the combination of fractal features of gene sequences, of protein sequences, and all of them, respectively.

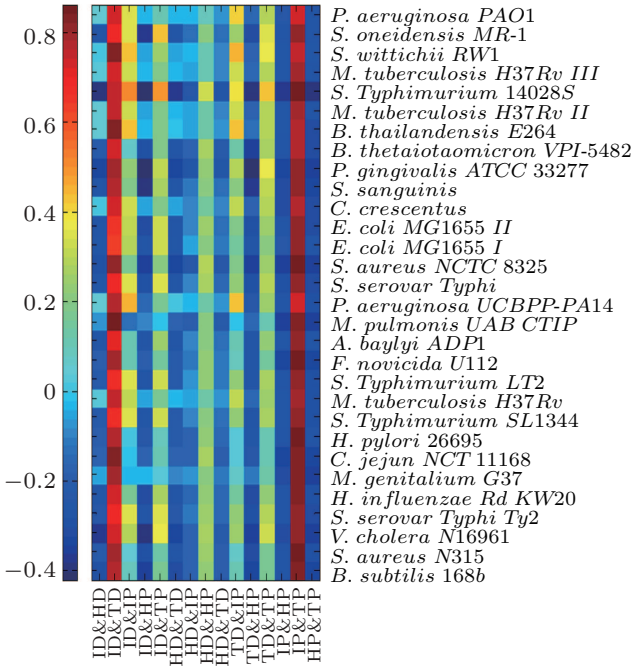


Fig. 3. (color online) Spearman correlation coefficients between pairs of fractal features of all 27 bacteria. Significant correlations generally exist among these fractals at 5% level. The exact values of this figure are provided in Supplement 2.

On the other hand, there are 15 (55.56%) and 20 (75.07%) bacteria obtaining better predictive performance with fractal features of gene sequences than that of protein sequences, based on Naïve Bayes and Random Forest classifiers respectively. So, we deduce that gene fractal features maybe generally outperform protein sequences for gene essential prediction.

Additionally, for Naïve Bayes classifier, there are 26 (96.26%) bacteria having better predictive performance of all fractal features than those of two other integrating strategies, where the improvement of average AUC score of those bacteria is 0.034 (5.26%). For Random Forest, there are 27 (100%) bacteria of such a kind, and the average improvement of AUC score is 0.021 (3.87%). It suggests that we can construct a classifier with better predictive performance by integrating features of gene and protein sequence.

3.3. Comparative studies

In recent years, several computational methods have been proposed to predict essential genes. In 2009, Acencio and Lemke constructed a decision tree-based meta-classifier based on network topological features, cellular compartments and biological processes to generate various predictors of essential genes.[24] Similarly, Palaniappan and Mukherjee used 52 input genomics, i.e., gene paralogs, sub-cellular location, Codon W features, etc., to learn the parameters of SVM classifier to

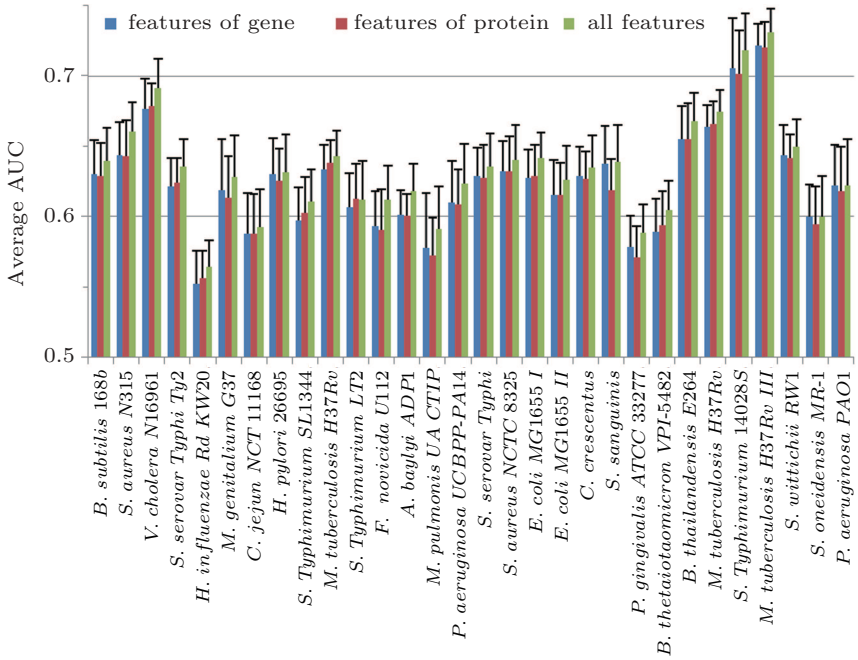


Fig. 4. (color online) Mean values and SDs of AUC scores of genes essentiality prediction in 27 bacteria based on Naïve Bayes classifier trained by different integrations of fractal features. The average AUC scores of 27 bacteria are above 0.62. There are 15 (55.56%) bacteria obtaining better predictive performance with fractal features of gene sequences than that of protein sequences and 26 (96.26%) bacteria having better predictive performance by using all fractal features.

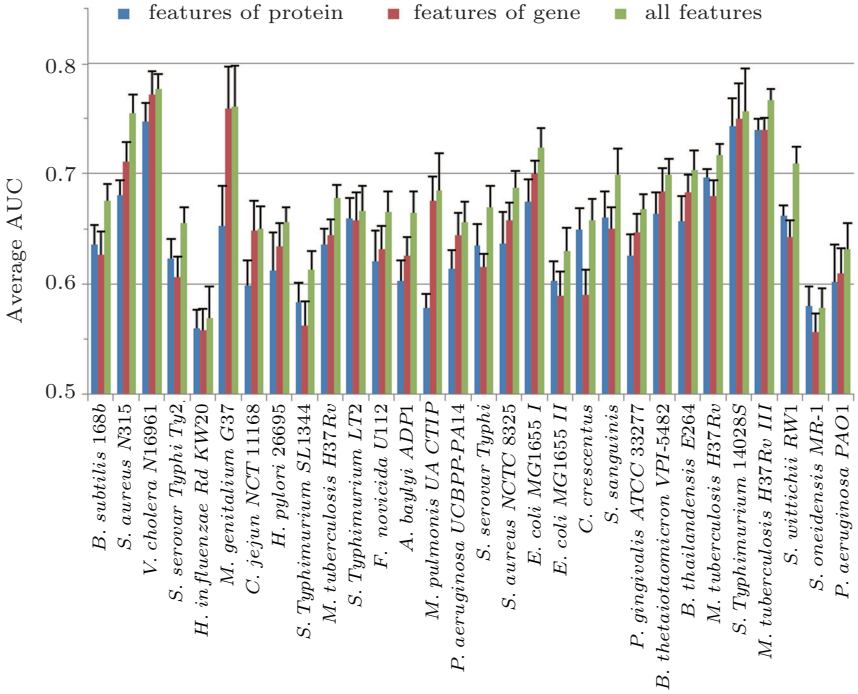


Fig. 5. (color online) Mean values and SDs of AUC scores of genes essentiality prediction in 27 bacteria based on Random Forest classifier trained by different integrations of fractal features. The average AUC scores of 27 bacteria are above 0.64. There are 20 (75.07%) bacteria obtaining better predictive performance with fractal features of gene sequences than that of protein sequences and the better predictive performance is obtained by using all fractal features.

predict essential genes.[36] In 2014 Ning *et al.* built an SVM classifier of gene essentiality by using all 158 features including Codon usage, amino acid usage, Codon W features, etc.[37] Tang *et al.* sought to measure essentiality by using PPI and gene expression data.[38] Here, we use those methods to make essentiality prediction in *E. coli* and then compare them with our classifier model through AUC score. The comparative result is shown in Fig. 5. As shown in Fig. 5, the AUC score of our model is larger than those of the methods of Tang, Palaniappan and Acencio, and competitive with that of the approach of Ning. Besides the limitation of predictive performance, those methods need a number of features, some of which are not easily obtainable from primary sequence. The fractal features in this paper are intrinsic features and

easily extracted from primary gene and protein sequence by computational methods. They are of powerful prediction and our classifier is satisfactorily discriminative between essential genes and non-essential genes.

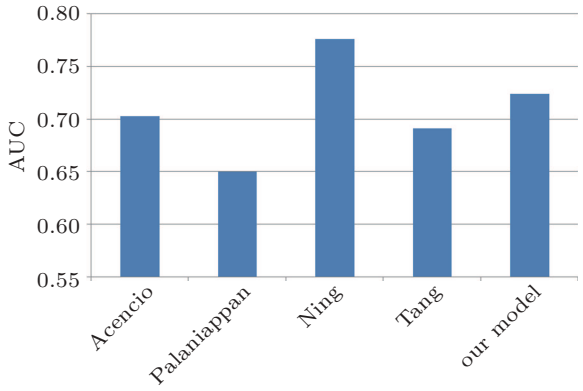


Fig. 6. (color online) Comparative results between our classifier model and other four methods on predicting essential genes in *E. coli* by AUC score.

4. Discussion and summary

The ability to consistently identify gene essentiality is of great significance to biological knowledge discovery and drug discovery operation, since it is basic requirement for sustaining organism alive. In this study, we assess the relationship between fractal features and essential genes by statistical analysis and machine learning-based methods, in hope of exploring the actual relationship between essentiality and these features, thus helping design computational tools to predict and screen essential genes from the bacterial genome. All currently available identified essential and non-essential genes and proteins in 27 bacteria from DEG database are studied in this paper. We find that essential genes significantly differ from non-essential genes in their fractal features in most bacteria. Because of the class imbalanced problem, we also design balanced training datasets to train Naïve Bayes and Random Forest classifiers with the aim of obtaining reliable results. The good performances of two classifiers imply that these fractal features are of powerful predictions of genes essentiality, either in integrative context or individually, although there are weak positive correlations among some of them.

Some studies have observed that genes essentiality powerfully relate to regulation complexity where the more targets a regulator has, the more important it is.[39] Our study provides facilitated access to quantitatively assess the regulation complexity of gene and protein sequences by fractal features. It is found that essential genes and proteins show higher regulatory complexity and less edited than nonessential genes and proteins. Essential genes are more evolutionarily conserved than non-essential gens and compared with non-essential proteins, essential proteins tend to enrich some function categories, such as transcription, translation and replication. However, they all have complex structures between deterministic and random, like chaotic attractors.

On the other hand, rapid and accurate identification of new essential gens requires learning classifier systems to integrate a subset of gene features and the work of Deng *et al.*[40] have shown integrations of both intrinsic and context-dependent features of genomic sequences help significantly improve the robustness and accuracy of prediction system. These fractal features in this study are intrinsic features which do not depend on external conditions. More importantly, they can be easily derived from primary genomic sequences, not relying on experimental environment. An additional advantage of our features over other attributes which only make sense to few specific model organisms is that their predictive utility keep reliable for 27 bacteria, which provides significant assist in the gene essentiality prediction in under-studied organisms.

In summary, through exploring the relationship between fractal feature and genes essentiality in different strategies, we come to the conclusion that these fractal features are effective and predictive quantitative indicators for checking bacterial essential genes by themselves and they are also promising contributors for designing computational tools identifying essential genes, either in combination strategy or individually. More broadly, the result obtained from our study will provide valuable information about synthetic biology and facilitate the annotations of microbial genomes and detection of potential drug target.

References

[1] Winzeler E A, Shoemaker D D and Astromoff A, *et al.* 1999 *Science* **285** 901
[2] Furney S J, Alb M M and Lopez-Bigas N 2006 *BMC Genomics* **7** 165
[3] Clatworthy A E, Pierson E and Hung D T 2007 *Nat. Chem. Biol.* **3** 541
[4] Glass J I, Hutchison C A, Smith H O and Venter J C 2009 *Mol. Syst. Biol.* **5** 330
[5] Gibson D G, Glass J I and Lartigue C, *et al.* 2010 *Science* **329** 52
[6] Kamath R S, Fraser A G, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman D P, Zipperlen P and Ahringer J 2003 *Nature* **421** 231
[7] Giaever G, Chu A M, Ni L *et al.* 2002 *Nature* **418** 387
[8] Gallagher L A, Ramage E, Jacobst M A, Kaul R, Brittnacher M and Manoil C 2007 *P. Natl. Acad. Sci. USA* **104** 1009
[9] Tong X, Campbell J W, Bal A Zsi G A B, Kay K A, Wanner B L, Gerdes S Y and Oltvai Z N 2004 *Biochem. Bioph. Res. Co.* **322** 347
[10] Antonia Molina Henares M, de la Torre J, Garcia Salamanca A, Jesus Molina Henares A, Carmen Herrera M, Ramos J L and Duque E 2010 *Environ. Microbiol.* 12 1468
[11] Gustafson A M, Snitkin E S, Parker S C, Delisi C and Kasif S 2006 *BMC Genomics* **7** 265
[12] Jehl M A, Arnold R and Rattei T 2011 *Nucleic Acids Res.* 391 D591
[13] Saha S, Heber S and Others 2006 *Genet. Mol. Res.* **5** 224
[14] Sharp P M 1991 *J. Mol. Evol.* **33** 23
[15] Li M, Zheng R, Zhang H, Wang J and Pan Y 2014 *Methods* **67** 325
[16] Rocha E P and Danchin A 2003 *Nat. Genet.* **34** 377
[17] P A L C, Papp B A Z and Hurst L D 2003 *Nature* **421** 496
[18] Lin Y, Gao F and Zhang C T 2010 *Biochem. Bioph. Res. Co.* **396** 472
[19] Hawoong J, Zoltan N O and Albert-Laszlo B 2003 *ComPlexus* **1** 19
[20] Han G S, Yu Z G and Vo A 2011 *Chin. Phys. B* **20** 100504

[21] Hwang Y C, Lin C C, Chang J Y, Mori H, Juan H F and Huang H C 2009 *Mol. Biosyst.* **5** 1672
[22] Wang J X, Peng W and Wu F X 2013 *Proteom. Clin. Appl.* **7** 181
[23] Ning K, Ng H K, Srihari S, Leong H W and Nesvizhskii A I 2010 *BMC Bioinformatics* **11** 505
[24] Acencio M L and Lemke N 2009 *BMC Bioinformatics* **10** 290
[25] Seringhaus M, Paccanaro A, Borneman A, Snyder M and Gerstein M 2006 *Genome Res.* **16** 1126
[26] Wunderlich Z and Mirny L A 2006 *Biophys. J.* **91** 2304
[27] Estrada E and Rodriguez-Velazquez J A 2005 *Phys. Rev. E* **71** 56103
[28] Yu Z G, Xiao Q J, Shi L, Wu Y J and Anh V 2010 *Chin. Phys. B* **19** 068701
[29] Yu Z G, Anh V, Min G Z and Long S C 2002 *Chin. Phys.* **11** 1313
[30] Zhu S M, Yu Z G and Vo A 2011 *Chin. Phys. B* **20** 100505
[31] Zhou Q and Yu Y M 2014 *Chaos, Solitons and Fractals* **69** 209
[32] Zhou Q and Yu Y M 2014 *J. Phys. D: Appl. Phys.* **47** 465401
[33] Jeffrey H J 1990 *Nucleic Acids Res.* **18** 2163
[34] Randic M 2008 *Chem. Phys. Lett.* **456** 84
[35] Yu Z G, Anh V and Lau K S 2004 *J. Theor. Biol.* **226** 341
[36] Palaniappan K and Mukherjee S 2011 *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, Honolulu, Hawaii, USA, p. 189
[37] Ning L W, Lin H, Ding H, Huang J, Rao N and Guo F B 2014 *Genet. Mol. Res* **13** 4564
[38] Tang X, Wang J, Zhong J and Pan Y 2014 *IEEEACM Trans. Comput. Biol. Bioinform.* **11** 407
[39] Yu H Y, Greenbaum D, Lu H X, Zhu X W and Gerstein M 2004 *Trends Genet.* **20** 227
[40] Deng J, Deng L, Su S, Zhang M, Lin X, Wei L, Minai A A, Hassett D J and Lu L J 2011 *Nucleic Acids Res.* **39** 795