Yan *et al. BMC Bioinformatics* (2020) 21:111
https://doi.org/10.1186/s12859-020-3426-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

PESM: predicting the essentiality of miRNAs based on gradient boosting machines and sequences

Cheng Yan[1,2], Fang-Xiang Wu[3], Jianxin Wang[1] and Guihua Duan[1]*

Abstract

Background: MicroRNAs (miRNAs) are a kind of small noncoding RNA molecules that are direct posttranscriptional regulations of mRNA targets. Studies have indicated that miRNAs play key roles in complex diseases by taking part in many biological processes, such as cell growth, cell death and so on. Therefore, in order to improve the effectiveness of disease diagnosis and treatment, it is appealing to develop advanced computational methods for predicting the essentiality of miRNAs.

Result: In this study, we propose a method (PESM) to predict the miRNA essentiality based on gradient boosting machines and miRNA sequences. First, PESM extracts the sequence and structural features of miRNAs. Then it uses gradient boosting machines to predict the essentiality of miRNAs. We conduct the 5-fold cross-validation to assess the prediction performance of our method. The area under the receiver operating characteristic curve (AUC), F-measure and accuracy (ACC) are used as the metrics to evaluate the prediction performance. We also compare PESM with other three competing methods which include miES, Gaussian Naive Bayes and Support Vector Machine.

Conclusion: The results of experiments show that PESM achieves the better prediction performance (AUC: 0.9117, F-measure: 0.8572, ACC: 0.8516) than other three computing methods. In addition, the relative importance of all features also further shows that newly added features can be helpful to improve the prediction performance of methods.

Keywords: MiRNA, Essentiality, Gradient boosting machines

Background

MicroRNAs (miRNAs) are small non-coding RNAs with a length of 22 nucleotides, which are processed from stem-loop regions of longer RNA transcripts [1]. They bind to the 3' untranslated regions (UTRs) of target mRNAs by sequence-specific base pairing to regulate the gene expression at the post-transcriptional level [2, 3]. Studies have shown that miRNAs play crucial roles in many biological processes, such as cell differentiation, growth, immune reaction and death, thereby leading to a variety of diseases [4, 5]. For example, miR-28-5p and miR-28-3p are down-regulated in colorectal cancer (CRC) samples compared with normal colon samples [6]. Members of the let-7 family of microRNAs were significantly downregulated in primary melanomas, and the anchorage-independent growth of melanoma cells are also inhibited by let-7b [7]. The poor clinical features in gastric cancer are associated with the low levels of miR-34b and miR-129 expression [8]. The incidence of lymphoma is regulated by the overexpression of miRNA hsa-mir-451a [9, 10]. Furthermore, after knocking out one or more members of a very broadly conserved miRNA family, some abnormal phenotypes are observed [11]. For example, as paralogous proteins, members of the same seed families often have at least partially redundant functions, with severe loss-of-function phenotypes apparent only after multiple family

*Correspondence: duangh@csu.edu.cn
[1] Hunan Provincial Key Lab on Bioinformtics, School of Computer Science and Engineering, Central South University, 932 South Lushan Rd, 410083 ChangSha, China
Full list of author information is available at the end of the article

members are disrupted, which includes mmu-mir-22 [12], mmu-mir-29 [13].

In order to systematically understand the associated mechanisms between miRNAs and diseases, some databases have been constructed, such as HMDD [14], miR2Disease [15], dbDEMC [16], Oncomirdb [17]. With these databases, some computational methods have been proposed to identify potential miRNA-disease associations. Based on a kernelized Bayesian matrix factorization model, Lan et al. proposed a computational method (KBMF-MDI) to predict miRNA-diseases associations based on known miRNA-disease associations, miRNA sequence and disease sematic information [18]. By integrating the miRNA-disease association network, miRNA similarity network and disease similarity network, You et al. [19] developed PBMDA to prioritize the underlying miRNA-disease associations, which used a special depth-first search algorithm in a heterogeneous network. Luo et al. also proposed a network-based method for drug repositioning based on similarities among drugs and diseases [20]. DNRLMF-MDA was proposed to discover hidden miRNA-disease associations based on known miRNA-disease associations, miRNA similarity and disease similarity, the main feature of DNRLMF-MDA was that it assigned higher importance levels to the observed interacting miRNA-disease pairs than unknown pairs [21]. Based on the inductive matrix completion model, IMCMDA was also proposed to predict miRNA-disease associations by integrating miRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity [22]. Chen et al. [23] proposed a computational model named Laplacian regularized sparse subspace learning for miRNA-disease association prediction (LRSSLMDA), which projected miRNA/disease' statistical feature profiles and graph theoretical feature profiles to a common subspace. MDHGI was a computational model to discover new miRNA-disease associations based on the matrix decomposition and heterogeneous graph inference, which integrated the predicted association probability obtained from matrix decomposition through a sparse learning method [24]. DLRMC was a computational method to predict miRNA-disease associations, based on matrix completion model with dual Laplacian regularization (DLRMC) [25]. EDTMDA was a computational method based on the ensemble of decision trees, which built a computational framework by integrating ensemble learning and dimensionality reduction [26]. Based on the logistic model tree, Wang et al. proposed a method for predicting miRNA-disease associations (LMTRDA) [27]. Pasquier et al. proposed a method to calculate the associations of miRNA disease pairs according to the vector similarity of miRNAs and diseases based on the distributional information of miRNAs and diseases in a high-dimensional vector space [28]. RKNNMDA was a type of instance-based learning to predict potential miRNA-disease associations based on the k-nearest neighbor algorithm and support vector machine (SVM) [29]. BNPMDA was a novel computational model of bipartite network projection for miRNA-disease association prediction, and its main feature was that bias ratings were constructed for miRNAs and diseases by using agglomerative hierarchical clustering [30]. VAEMDA was a novel miRNA-disease association prediction method based on an unsupervised deep learning framework with variational autoencoder [31]. Yan et al. proposed ABMDA to predict potential miRNA-disease associations, which balanced the positive and negative samples by performing random sampling based on k-means clustering on negative samples [32]. Based on the k-mer sparse matrix to extract miRNA sequence information and deep auto-encoder neural network (AE), MLMDA was developed to predict miRNA-disease associations [33]. Cheng et al. also proposed a miRNA-disease association prediction method based on adaptive multi-view multi-label learning(AMVML) [34]. By combined the weighted profile and collaborative matrix factorization (CMF), a new computation model logistic weighted profile-based collaborative matrix factorization (LWPCMF) was developed to predict miRNA-disease associations [35]. DBMDA was a novel computational model for miRNA-disease association prediction, the notable feature of this method was inferring the global similarity from region distances based on the miRNA sequences [36]. By combing the kernel-based nonlinear dimensionality reduction, matrix factorization and binary classification, a neoteric Bayesian model (KBMFMDA) was proposed to predict miRNA-disease associations [37]. Chen et al. also proposed a miRNA-disease association prediction method (NCMCMDA) based on a neighborhood constraint matrix completion model [38]. Based on the neural inductive matrix completion with graph convolutional networks, Li et al. also proposed a method to predict miRNA-disease associations [39]. In addition, the matrix completion model was also used in drug repositioning [40–43], predicting lncRNA-disease associations [44, 45] and microbe-disease associations [46].

Furthermore, the miRNA-target interaction was also predicted by miRTRS based on known miRNA-target interactions, miRNA sequences and gene sequences [47]. Bartel et al. [11] described the important biological functions identified for most of the broadly conserved miRNAs of mammals, and they also reviewed how metazoan miRNAs recognized and caused the repression of their targets. Studies demonstrated that some miRNA molecules were essential to the disease development [48]. Therefore, inspired by the bioinformatics development of the protein essentiality prediction [49, 50], Gao et al.

first proposed a computational method (miES) based on machine learning and sequence features to identify the miRNA essentiality [51]. MiES used the miRNA sequences and a logistic regression model for performing miRNAome-wide search for essential miRNAs. In addition, miES further analyzed the miRNA conservation [52], miRNA expression dataset and miRNA disease spectrum width (DSW) [53] to understand the important basis for predicting the essentiality of miRNAs [54]. In addition, the sequence features also used in study of genome [55]. The frequencies of k-mers were also used in ARP to classify the reads into three categories [56]. In MultiMotifMaker, the position weight matrix (PWM) was a used representation of motifs, and its 4 columns (A,C,T,G) described the frequency of occurrence of each base at each position [57].

However, the current development of miRNA essentiality prediction method is still not good enough. Complex and deeper features related to miRNAs should be considered to improve the prediction quality of current methods. The more effective and advanced computational methods should also be developed to identify essential miRNAs. Therefore, in this study we propose a computational method (PESM) to predict potential essential miRNA based on the essential miRNA and non-essential miRNAs benchmark dataset. PESM first integrates more miRNA sequence features (such as 18 dinucleotide features : UC%, UG% and so on) as in miES. Then PESM uses gradient boosting machines to predict the essentiality of miRNAs. In order to assess the prediction performance of PESM and compare it with other computational methods, we also conduct the 5-fold cross validation (5CV). In addition, the area under of receiver operating characteristic (ROC) curve (AUC), accuracy (ACC) and F-measure are used as the metrics of all prediction methods. The competing methods include miES, Gaussian Naive Bayes (GaussianNB) and SVM. The experiment results of 5CV show that PESM can obtain better prediction performance in terms of AUC, ACC and F-measure (AUC: 0.9117, ACC: 0.8516 and F-mearsure: 0.8572) than other competing methods: miES (AUC: 0.8837, ACC: 0.8263 and F-mearsure: 0.8326), GaussianNB (AUC: 0. 8720, ACC: 0.8000 and F-mearsure: 0.8093) and SVM (AUC: 0.8571, ACC: 0.8206 and F-mearsure: 0.8271). Comparing with miES, PESM integrates more sequence and structural features of miRNAs. In addition, the gradient boosting machine model is used to compute the predicted scores of essential miRNAs. By analyzing the relative importance of the features, we can also conclude that the added new features can represent the intrinsical characteristics of miRNAs. Finally, the experiment results also prove that the prediction ability of our method is superior to other competing methods.

Methods

Materials

In this study, we use the benchmark dataset of essential miRNAs and non-essential miRNAs, which consists of the pre-miRNA sequences and mature-miRNA sequences of human, rat and mouse from miRbase [52]. The benchmark dataset includes 77 essential mice miRNAs and the same number of non-essential miRNAs [11]. The known essential mice miRNAs (positive samples) and non-essential miRNAs (negative samples) were obtained from the review paper [11]. In miES, the negative samples were generated with two strategies: (1) the random selection; (2) the selection according to the maximum mean AUC.

Feature set

The miRNAs are transcribed as long primary miRNAs, which produce miRNA precursors (pre-miRNAs) by nuclear RNase III Drosha [58]. Then the pre-miRNAs are cleaved into mature miRNAs [1]. All pre-miRNAs have stem-loop hairpin structures [59]. Therefore, by considering the production process of miRNAs and the structure of pre-miRNAs, PESM uses the features of not only mature-miRNAs but also pre-miRNAs. The selected feature set of pre-miRNA sequences and mature-miRNA sequences has important influence on predicting the essentiality of miRNAs. In this study, we first extract the 14 pre-miRNA and mature-miRNA features which include information about sequences and structures. In addition, up to now various feature sets have been proposed to study pre-miRNA and other relative prediction problems. Inspired by the successful application of dinucleotide frequency information in predicting pre-miRNAs, we add the 18 dinucleotide frequency features of pre-miRNAs and mature-miRNAs in this study [60]. In addition, we further add other 6 structure features of pre-miRNAs, includes normalized base-pairing propensity (P(s)), normalized base-pairing propensity divided by its length (nP(s)), normalized Shannon entropy (Q(s)), normalized Shannon entropy divided by its length (nQ(s)), normalized base-pair distance(D(s)) and normalized base-pair distance divided by its length(nD(s)) [61]. We use the module RNAlib of Vienna RNA Package to intrinsic folding quantitative measures P(S), nP(S), Q(s), nQ(s), D(s) and nD(s) [62]. These structure features and Vienna RNA Package have been broadly used in both miRNA prediction and pre-miRNA prediction [63–65]. As a result, our method consists of 38 features. Note that these features also include the 14 features which are used in miES. The more detail about the feature set is described in Table 1.

Gradient boosting regression trees

After computing the above sequence and structure features, we take a supervised learning method named gradient

Table 1 The feature set description

Category	Description	Number of features
Base content in pre-miRNAs	The content of base *S* in pre-miRNAs, $S \in \{U, C, G\}$	3
mature-miRNAs length	The sequence length of mature-miRNAs	1
Base content in mature-miRNAs	The content of base *S* in mature-miRNAs, $S \in \{U, C, G\}$	3
non-mature-miRNAs length	The sequence length of non-mature-miRNAs	1
Base content in non-mature-miRNAs	The content of base *S* in non-mature-miRNAs, $S \in \{U, C, G\}$	3
MFE and nMFE	The minimum free energy of pre-miRNA secondary structures and it is divided by its length	2
Cleavage site base class	The cleavage sites are assigned into 3 classes, 1: all cleavage sites of mature-miRNAs from the same pre-miRNAs are *U*; 0: not all cleavage sites are *U*; -1: all are non-*U*.	1
Dinucleotide pairs frequency in pre-miRNAs	The Dinucleotide pairs *SZ* frequency in pre-miRNAs, $S, Z \in \{U, C, G\}$	9
Dinucleotide pairs frequency in mature-miRNAs	The Dinucleotide pairs *SZ* frequency in mature-miRNAs, $S, Z \in \{U, C, G\}$	9
The structure feature of pre-miRNAs	Normalized base-pairing propensity ($P(s)$), Normalized base-pairing propensity divided by its length ($nP(s)$), Normalized Shannon entropy ($Q(s)$), Normalized Shannon entropy divided by its length ($nQ(s)$), Normalized base-pair distance ($D(s)$), Normalized base-pair distance divided by its length ($nD(s)$)	6

boosting regression trees derived from the gradient boosting machine model to predict essential miRNAs [66, 67]. This method has been successfully used in other classification issues [68, 69]. In the common supervised learning scenario, the sample data set can be represented by a set containing feature vectors and labels: $D = \{(x_i, y_i)\}(i = 1, ..., N)$, where N is the number of samples [70]. In this study, $x_i \in R^d$ is the feature vector of the $i - th$ miRNA, while y_i is its essentiality score. d is the dimensionality of features. According to the gradient boosting regression tree model, the predicted essentiality score $\hat{y}_i$ of miRNA i from its input feature vector can be calculated as follows:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

where K is the maximum depth of regression trees and F is a set of functions containing the partition of the region and score [70]. In order to learn the set of trees $\{f_i\}$, the regularized objective function is defined as follows [70]:

$$L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \tag{2}$$

where l is a differentiable convex loss function that is used to calculate the difference between the prediction $\hat{y}_i$ and target y_i. To avoid the overfitting, the second term Ω is used to control the complexity of the model. This regularized function can penalize the complicated models. Finally, the model with simple and predictive functions can be selected.

Since this model includes functions as parameters, it can not use traditional optimization methods in the Euclidean space to establish it. Instead, a new tree f_t is added to the ensemble, which optimizes the objective function and is searched from the functional space F at each iteration t. The process is defined as follows:

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l\left(\left(y_i, \hat{y}_i^{(t-1)}\right) + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(f_i) \end{aligned} \tag{3}$$

where $\hat{y}_i^{(t)}$ is the prediction of the $i - th$ instance at the $t - th$ iteration. The model finds f_t to optimize the above objective function.

Equation (3) is still hard to optimize in the general setting, so the second order Taylor expansion is used to approximate the objective function as follows:

$$\begin{aligned} L^{(t)} \simeq \sum_{i=1}^{n} \Big[l\Big(\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) \\ + \frac{1}{2} h_i {f_t}^2(x_i)\Big)\Big] + \sum_{i=1}^{t} \Omega(f_i) \end{aligned} \tag{4}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$. By removing the terms independent of $f_t(x_i)$, the following approximate objective function at step t can be obtained:

$$\overline{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)) \right] + \sum_{i=1}^{t} \Omega(f_i) \tag{5}$$

A gradient boosting algorithm iteratively adds functions that optimizes $\overline{L}^{(t)}$ for a number of user-specified iterations.

In order to learn the function f_t in each step, the mapping $q : R^d \rightarrow \{1, 2, ..., T\}$ is defined to map the input to the index of the region. The function is defined as follow:

$$f_t(X) = w_{q(X)} \tag{6}$$

where w is a vector of scores in each region and q represents the decision tree structure. Furthermore, the function complexity was defined as follow:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

where T is the number of trees. The parameters γ and λ are used to make a balance. w_j^2 is the prediction score for data corresponding to the $j - th$ leaf from f_t.

Then Eq(5) can be rewritten as follow:

$$\overline{L}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{8}$$

where $I_j = \{i | q(x_i) = j\}$ is defined as the instance set of region j. When $q(x)$ is fixed, the optimal weight w_j^* of region j can be calculated as follows:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{9}$$

The optimal objective value is calculated as follow:

$$\overline{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{10}$$

Equation (10) is used to score the region partition specified by q. It also can find a good structure according to the previous reference [70]. Since there can be infinitely many possible candidates of the tree structure, it applied a greedy algorithm in practice [70]. The one step of the algorithm was that splitting a leaf into two leaves. In each round, it greedily enumerated the features and split the feature that gives the maximum reduction calculated by Eq. (10). The main feature of this model is the explicit regularization term which prevents the model from overfitting. The detail of this model can be found in Chen et al. [67].

Results

Performance evaluation

In order to assess the prediction performance of our method and other computing methods, we conduct the 5CV based on the same benchmark dataset. The competing methods include miES [51], GaussianNB [71] and SVM [72, 73]. The benchmark dataset is downloaded from miES. In each round of the 5CV, we divide the essential miRNAs and non-essential miRNAs into the 5 sets, 4 of which are used to train the model while the left one is used as the testing set. We repeat the 5CV 50 times in this study.

In addition, the AUC value is used to measure the prediction performance of computational methods. The ROCs are drawn with TPR (true positive rate) with respect to FPR (false positive rate) values. TPR is the fraction of essential miRNAs that are correctly predicted, while FPR is the fraction of non-essential miRNAs that are incorrectly predicted. Furthermore, the F-measure and ACC are also used to evaluate the prediction performance of computational methods. The F-measure is calculated from the harmonic mean of precision (P) and recall (R) ($F = 2 * P * R/(P + R)$).

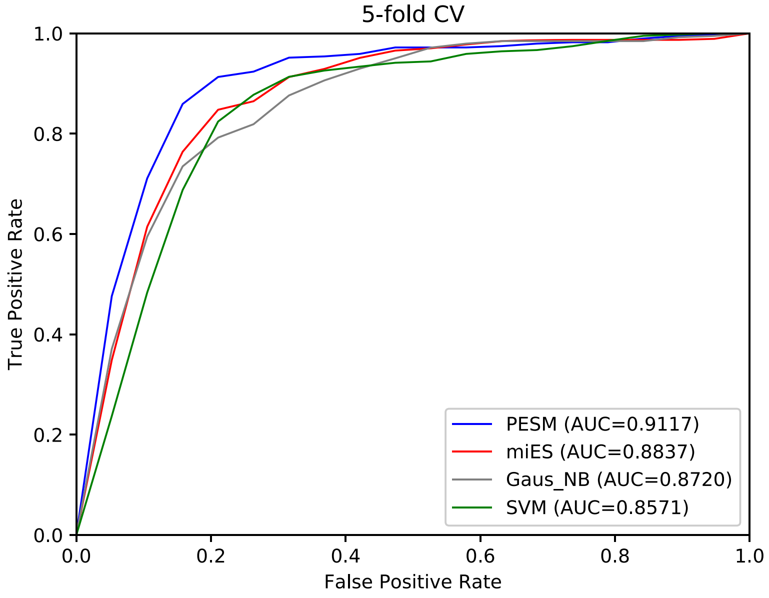

Fig. 1 The ROC plot of the four computational methods with on the 5-fold cross validation

Table 2 The ACC and F-measure values of four computational methods with on the 5-fold cross validation

Method	ACC	F-measure
PESM	0.8516	0.8572
miES	0.8263	0.8326
GaussianNB	0.8000	0.8093
SVM	0.8206	0.8271

Comparison with other competing methods

In this study, we compare our method to other three competing methods which include miES, GaussianNB and SVM. MiES was a computational method for miRNA essentiality prediction, which only uses sequence features of known essential miRNAs. In addition, GaussianNB and SVM are the typical classification models. Figure 1 plots the ROC curve and shows the AUC values of four computational methods. In terms of AUC, our method obtains the best prediction performance as its AUC value is 0.9117, compared with other methods (miES: 0.8837, GaussianNB: 0.8720 and SVM: 0.8571).

In addition, Table 2 shows the ACC and F-measure values of four methods with the 5CV validation. We can see from Table 2 that our method obtains the best prediction performance (ACC:0.8516 and F-mearsure:0.8572), compared with other methods (miES (ACC:0.8263 and F-mearsure:0.8326), GaussianNB (ACC:0.8000 and F-mearsure:0.8093) and SVM (ACC:0.8206 and F-mearsure:0.8271)).

Relative importance of the features

In order to demonstrate the newly added features in the prediction method, we further analyze the relative importance of all 38 features. Figure 2 plots the relative importance of the features, which is computed by the XGBoost package. We can see from Fig. 2 that 4 newly added features are ranked top 10 based on the relative importance, which include $\%CC$ in mat, $P(s)$, $nQ(s)$ and $D(s)$. These 4 added features rank 6, 4, 3 and 5, respectively. It also demonstrates that the newly added features can reflect the intrinsic characteristics of miRNAs and help improve the performance of predicting essential miRNAs.

Parameter analysis for γ, λ, K and T

In this study, we analyze four parameters, including the regularization terms on the number of regions (γ), on the sum of squared scores (λ), the maximum depth of regression trees (K) and the number of trees (T). The default values of γ, λ, K and T are 0, 0.1, 6 and 1000, respectively. We conduct the 5CV to evaluate the prediction performance of PESM. In addition, one of four parameters is

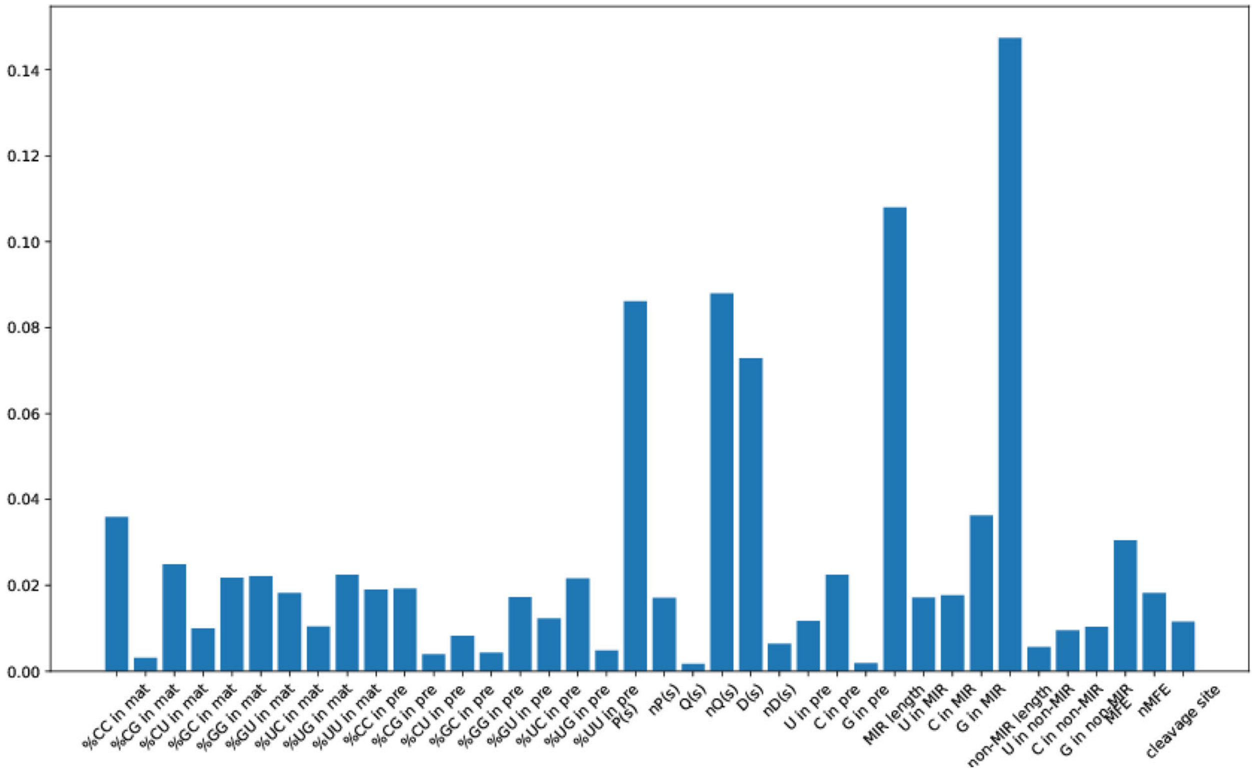

Fig. 2 The relative importance of all 38 features. pre-miR means pre-miRNA; MIR means mature miRNA; non-MIR means non-mature-miRNA

Table 3 The prediction performances of PESM with different settings of λ

λ	0.25	0.50	0.75	1.0
AUC	0.9116	0.9116	0.9116	0.9117
λ	1.25	1.50	1.75	2.0
AUC	0.9083	0.9041	0.9025	0.9041

analyzing while the other three parameters are set to be the default values.

The default value of γ is 0 in the XGBoost package. We also compute the prediction performance of PESM with the parameter γ in the set 0, 0.1, 0.2 according to reference [70]. The AUC values of our method are 0.9117, 0.9133 and 0.9100. In this study, we set the value of parameter γ to 0 based on our experiment results and the default value in the XGBoost package.

We evaluate the prediction performance of PESM when parameter λ ranges from 0.25 to 2.0 with the increment of 0.25. We can see from Table 3 that PESM can achieve the best prediction performance when it is set to 1.0 which is also the default value of XGBoost package. Therefore, we set λ to 1.0 in this study.

Furthermore, in the XGBoost package, the default value of parameter K is 6. Table 4 describes the AUC values obtained by PESM when K ranges from 3 to 9. We can see from Table 4 that our method can obtain the best prediction performance when K is set to be 7, and obtain reliable prediction performances when K ranges from 5 to 7. Therefore, by considering the default value in the XGBoost package and our experiments results, we set K to 6 in this study.

Finally, Table 5 shows the prediction performance of PESM when the tree number T is set to 100, 500, 1000, 1500, 2000. We can see from Table 5 that PESM obtain the reliable prediction performance when T is selected from one of set 1000, 1500, 2000. Therefore, we also set the default value of T to 1000 in this study.

Discussion

With the development of biotechnology, studies have shown that miRNAs participate in many biological processes, such as cell growth, cell death and so on. Furthermore, miRNAs also play important roles in human diseases, especially the complex diseases, such as cancer. Therefore, the study of miRNA and disease associations has become a main research topic in bioinformatics.

Table 4 The prediction performances of PESM with different settings of K

K	3	4	5	6	7	8	9
AUC	0.9066	0.9067	0.9116	0.9117	0.9133	0.9058	0.9053

Table 5 The prediction performances of PESM with different settings of T

T	100	500	1000	1500	2000
AUC	0.8958	0.9068	0.9117	0.9141	0.9113

Based on the more systematic understanding of miRNAs, studies further demonstrate that some miRNA molecules are essential to the disease development. The essential miRNAs are necessary to manifest principles of disease mechanisms. Therefore, identifying the essential miRNAs is very appealing.

Conclusion

In this study, we have developed a computational method (PESM) to predict the essentiality of miRNAs. PESM integrates the 38 sequence and structural features of miRNAs. Then it further uses the gradient boosting machines to compute the predicted scores of essential miRNAs. The experiment results with the 5-fold cross validation show that the prediction performance of PESM is superior to other competing methods, including the state-of-art method miES. Finally, we have analyzed the relative importance of all features by the XGBoost package, and the results demonstrate that the newly added features can further improve the prediction performances.

Although our method can effectively predict the essential miRNAs and non-essential miRNAs, its limits should be addressed in the future. First, the non-essential miRNAs in the current benchmark dataset are randomly selected. Second, the more features of miRNAs also should be designed, such as topological features of miRNAs. Finally, other similarity-based methods [74], collaborative metric learning methods [75] and deep learning methods [76, 77] should be adopted. We would provide a more effective computational method to predict essential miRNAs by addressing above limitations in the future.

Abbreviations

5CV: 10-fold cross validation; ACC: Accuracy; AUC: Area under the receiver operating curve; FPR :False positive rate; GaussianNB: Gaussian Naive Bayes; MFE: minimum free energy; SVM: Support vector machine; TPR: True positive rate

Acknowledgements

The authors are very grateful to the anonymous reviewers for their constructive comments which have helped significantly in revising this work.

Authors' contributions

GD and JW conceived the project; CY designed the experiments; and CY performed the experiments; CY, GD and FXW wrote the paper. All authors read and approved the final manuscript.

Funding

Publication costs are funded by National Natural Science Foundation of China under Grant No.61962050. The authors would like to express their gratitude for the support from the National Natural Science Foundation of China(No.61772552, No.61420106009, No.61832019 and No.61962050), 111 Project

(No. B18059) and Hunan Provinvial Science and Technology Program (No. 2018WK4001).

Availability of data and materials

The datasets and source codes are available at https://github.com/bioinfomaticsCSU/PESM.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Hunan Provincial Key Lab on Bioinformtics, School of Computer Science and Engineering, Central South University, 932 South Lushan Rd, 410083 ChangSha, China. [2]School of Computer and Information,Qiannan Normal University for Nationalities, Longshan Road, 558000 DuYun, China. [3]Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, SKS7N5A9 Saskatoon, Canada.

Received: 28 November 2019 Accepted: 21 February 2020
Published online: 18 March 2020

References

1. Bartel DP. Micrornas: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.
2. Ambros V. micrornas: tiny regulators with great potential. Cell. 2001;107(7):823–6.
3. Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded rna. Nature. 2004;431(7006):343.
4. Wen D, Danquah M, Chaudhary AK, Mahato RI. Small molecules targeting microrna for cancer therapy: Promises and obstacles. J Control Rel. 2015;219:237–47.
5. Chen X, Xie D, Zhao Q, You Z-H. Micrornas and complex diseases: from experimental results to computational models. Brief Bioinforma. 2019;20(2):515–39.
6. Almeida MI, Nicoloso MS, Zeng L, Ivan C, Spizzo R, Gafà R, Xiao L, Zhang X, Vannini I, Fanini F, et al. Strand-specific mir-28-5p and mir-28-3p have distinct effects in colorectal cancer cells. Gastroenterology. 2012;142(4):886–96.
7. Schultz J, Lorenz P, Gross G, Ibrahim S, Kunz M. Microrna let-7b targets important cell cycle molecules in malignant melanoma cells and interferes with anchorage-independent growth. Cell Res. 2008;18(5):549.
8. Tsai K-W, Wu C-W, Hu L-Y, Li S-C, Liao Y-L, Lai C-H, Kao H-W, Fang W-L, Huang K-H, Chan W-C, et al. Epigenetic regulation of mir-34b and mir-129 expression in gastric cancer. Int J Cancer. 2011;129(11):2600–10.
9. Gorur A, Fidanci SB, Unal ND, Ayaz L, Akbayir S, Yaroglu HY, Dirlik M, Serin MS, Tamer L. Determination of plasma microrna for early detection of gastric cancer. Mol Biol Rep. 2013;40(3):2091–6.
10. Weidhaas J. Using micrornas to understand cancer biology. Lancet Oncol. 2010;11(2):106–7.
11. Bartel DP. Metazoan micrornas. Cell. 2018;173(1):20–51.
12. Lu W, You R, Yuan X, Yang T, Samuel EL, Marcano DC, Sikkema WK, Tour JM, Rodriguez A, Kheradmand F, et al. The microrna mir-22 inhibits the histone deacetylase hdac4 to promote t h 17 cell–dependent emphysema. Nat Immunol. 2015;16(11):1185.
13. Dooley J, Garcia-Perez JE, Sreenivasan J, Schlenner SM, Vangoitsenhoven R, Papadopoulou AS, Tian L, Schonefeldt S, Serneels L, Deroose C, et al. The microrna-29 family dictates the balance between homeostatic and pathological glucose handling in diabetes and obesity. Diabetes. 2016;65(1):53–61.
14. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. Nucleic Acids Res. 2013;42(D1):1070–4.
15. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. mir2disease: a manually curated database for microrna deregulation in human disease. Nucleic Acids Res. 2008;37(suppl_1):98–104.
16. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y, et al. dbdemc: a database of differentially expressed mirnas in human cancers. In: BMC Genomics, vol. 11. BioMed Central; 2010. p. 5. https://doi.org/10.1186/1471-2164-11-s4-s5.
17. Wang D, Gu J, Wang T, Ding Z. Oncomirdb: a database for the experimentally verified oncogenic and tumor-suppressive micrornas. Bioinformatics. 2014;30(15):2237–8.
18. Lan W, Wang J, Li M, Liu J, Wu F-X, Pan Y. Predicting microrna-disease associations based on improved microrna and disease similarities. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2018;15(6):1774–82.
19. You Z-H, Huang Z-A, Zhu Z, Yan G-Y, Li Z-W, Wen Z, Chen X. Pbmda: A novel and effective path-based computational model for mirna-disease association prediction. PLoS Comput Biol. 2017;13(3):1005455.
20. Luo H, Wang J, Li M, Luo J, Peng X, Wu F-X, Pan Y. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics. 2016;32(17):2664–71.
21. Yan C, Wang J, Ni P, Lan W, Wu F-X, Pan Y. Dnrlmf-mda: predicting microrna-disease associations based on similarities of micrornas and diseases. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16(1):233–43.
22. Chen X, Wang L, Qu J, Guan N-N, Li J-Q. Predicting mirna–disease association based on inductive matrix completion. Bioinformatics. 2018;34(24):4256–65.
23. Chen X, Huang L. Lrsslmda: Laplacian regularized sparse subspace learning for mirna-disease association prediction. PLoS Comput Biol. 2017;13(12):1005912.
24. Chen X, Yin J, Qu J, Huang L. Mdhgi: Matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. PLoS Comput Biol. 2018;14(8):1006418.
25. Tang C, Zhou H, Zheng X, Zhang Y, Sha X. Dual laplacian regularized matrix completion for microrna-disease associations prediction. RNA Biol. 2019;16(5):601–11.
26. Chen X, Zhu C-C, Yin J. Ensemble of decision tree reveals potential mirna-disease associations. PLoS Comput Biol. 2019;15(7):1007209.
27. Wang L, You Z-H, Chen X, Li Y-M, Dong Y-N, Li L-P, Zheng K. Lmtrda: Using logistic model tree to predict mirna-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput Biol. 2019;15(3):1006865.
28. Pasquier C, Gardès J. Prediction of mirna-disease associations with a vector space model. Sci Rep. 2016;6:27036.
29. Chen X, Wu Q-F, Yan G-Y. Rknnmda: ranking-based knn for mirna-disease association prediction. RNA Biol. 2017;14(7):952–62.
30. Chen X, Xie D, Wang L, Zhao Q, You Z-H, Liu H. Bnpmda: bipartite network projection for mirna–disease association prediction. Bioinformatics. 2018;34(18):3178–86.
31. Zhang L, Chen X, Yin J. Prediction of potential mirna–disease associations through a novel unsupervised deep learning framework with variational autoencoder. Cells. 2019;8(9):1040.
32. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential mirna-disease associations. Bioinformatics. 2019;35(22):4730–8.
33. Zheng K, You Z-H, Wang L, Zhou Y, Li L-P, Li Z-W. Mlmda: a machine learning approach to predict and validate microrna–disease associations by integrating of heterogenous information sources. J Transl Med. 2019;17(1):260.
34. Liang C, Yu S, Luo J. Adaptive multi-view multi-label learning for identifying disease-associated candidate mirnas. PLoS Comput Biol. 2019;15(4):1006931.
35. Yin M-M, Cui Z, Gao M-M, Liu J-X, Gao Y-L. Lwpcmf: logistic weighted profile-based collaborative matrix factorization for predicting mirna-disease associations. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/tcbb.2019.2937774.
36. Zheng K, You Z-H, Wang L, Zhou Y, Li L-P, Li Z-W. Dbmda: A unified embedding for sequence-based mirna similarity measure with applications to predict and validate mirna-disease associations. Mol Ther-Nucleic Acids. 2020;19:602–11.
37. Chen X, Li S-X, Yin J, Wang C-C. Potential mirna-disease association prediction based on kernelized bayesian matrix factorization. Genomics. 2020;112(1):809–19.
38. Chen X, Sun L-G, Zhao Y. Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion. Brief Bioinforma. 2020. https://doi.org/10.1093/bib/bbz159.

39. Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btz965.
40. Yang M, Luo H, Li Y, Wang J. Drug repositioning based on bounded nuclear norm regularization. Bioinformatics. 2019;35(14):455–63.
41. Luo H, Li M, Wang S, Liu Q, Li Y, Wang J. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics. 2018;34(11):1904–12.
42. Yang M, Luo H, Li Y, Wu F-X, Wang J. Overlap matrix completion for predicting drug-associated indications. PLoS Comput Biol. 2019;15(12):. https://doi.org/10.1371/journal.pcbi.1007541.
43. Luo H, Li M, Mengyun Y, Wu F-X, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinforma. 2019. https://doi.org/10.1093/bib/bbz176.
44. Lu C, Yang M, Luo F, Wu F-X, Li M, Pan Y, Li Y, Wang J. Prediction of lncrna-disease associations based on inductive matrix completion. Bioinformatics. 2018;34(19):3357–64.
45. Lu C, Yang M, Li M, Li Y, Wu F, Wang J. Predicting human lncrna-disease associations based on geometric matrix completion. IEEE J Biomed Health Inform. 2019. https://doi.org/10.1109/JBHI.2019.2958389.
46. Yan C, Duan G, Wu F, Pan Y, Wang J. Mchmda: Predicting microbe-disease associations based on similarities and low-rank matrix completion. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2926716.
47. Jiang H, Wang J, Li M, Lan W, Wu F, Pan Y. mirtrs: A recommendation algorithm for predicting mirna targets. IEEE/ACM Trans Comput Biol Bioinforma. 2018. (https://doi.org/10.1109/TCBB.2018.2873299).
48. Beermann J, Piccoli M-T, Viereck J, Thum T. Non-coding rnas in development and disease: background, mechanisms, and therapeutic approaches. Physiol Rev. 2016;96(4):1297–325.
49. Li M, Li W, Wu F-X, Pan Y, Wang J. Identifying essential proteins based on sub-network partition and prioritization by integrating subcellular localization information. J Theor Biol. 2018;447:65–73.
50. Li G, Li M, Peng W, Li Y, Pan Y, Wang J. A novel extended pareto optimality consensus model for predicting essential proteins. J Theor Biol. 2019;480:141–9.
51. Song F, Cui C, Gao L, Cui Q. mies: predicting the essentiality of mirnas with machine learning and sequence features. Bioinformatics. 2018;35(6):1053–4.
52. Kozomara A, Griffiths-Jones S. mirbase: annotating high confidence micrornas using deep sequencing data. Nucleic Acids Res. 2013;42(D1):68–73.
53. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. Nucleic Acids Res. 2018;47(D1):1013–7.
54. De Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, Åström G, Babina M, Bertin N, Burroughs AM, et al. An integrated expression atlas of mirnas and their promoters in human and mouse. Nat Biotechnol. 2017;35(9):872.
55. Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, Xiao C-L, Luo F, Wang J. Deepsignal: detecting dna methylation state from nanopore sequencing reads using deep-learning. Bioinformatics. 2019;35(22):4586–95.
56. Liao X, Li M, Junwei L, Zou Y, Wu F-X, Pan Y, Luo F, Wang J. Improving assembly based on read classification. IEEE/ACM Trans Comput Biol Bioinforma. 2020;17(1):177–88.
57. Li T, Zhang X, Luo F, Wu F-X, Wang J. Multimotifmaker: a multi-thread tool for identifying dna methylation motifs from pacbio reads. IEEE/ACM Trans Comput Biol Bioinforma. 2020;17(1):220–5.
58. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, et al. The nuclear rnase iii drosha initiates microrna processing. Nature. 2003;425(6956):415.
59. Nelson P, Kiriakidou M, Sharma A, Maniataki E, Mourelatos Z. The microrna world: small is mighty. Trends Biochem Sci. 2003;28(10):534–40.
60. Kleftogiannis D, Theofilatos K, Likothanassis S, Mavroudi S. Yamipred: A novel evolutionary method for predicting pre-mirnas and selecting relevant features. IEEE/ACM Trans Comput Biol Bioinforma. 2015;12(5):1183–92.
61. Loong SNK, Mishra SK. Unique folding of precursor micrornas: quantitative evidence and implications for de novo identification. Rna. 2007;13(2):170–87.
62. Hofacker IL. Vienna rna secondary structure server. Nucleic Acids Res. 2003;31(13):3429–31.
63. Batuwita R, Palade V. micropred: effective classification of pre-mirnas for human mirna gene prediction. Bioinformatics. 2009;25(8):989–95.
64. Tseng K-C, Chiang-Hsieh Y-F, Pai H, Chow C-N, Lee S-C, Zheng H-Q, Kuo P-L, Li G-Z, Hung Y-C, Lin N-S, et al. microrpm: a microrna prediction model based only on plant small rna sequencing data. Bioinformatics. 2017;34(7):1108–15.
65. Stegmayer G, Yones C, Kamenetzky L, Milone DH. High class-imbalance in pre-mirna prediction: a novel approach based on deepsom. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2017;14(6):1316–26.
66. Friedman J, Hastie T, Tibshirani R, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat. 2000;28(2):337–407.
67. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
68. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminformatics. 2017;9(1):24.
69. Öztürk H, Özgür A, Ozkirimli E. Deepdta: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):821–9.
70. Chen T, He T. Higgs boson discovery with boosted trees. In: NIPS 2014 Workshop on High-energy Physics and Machine Learning. Montreal; 2015. p. 69–80.
71. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Louppe G. Scikit-learn: Machine learning in python. J Mach Learn Res. 2013;12(10):2825–30.
72. Chang C-C, Lin C-J. Libsvm: A library for support vector machines. ACM Trans Intell Syst Technol (TIST). 2011;2(3):27.
73. Chen Q, Lai D, Lan W, Wu X, Chen B, Chen Y-PP, Wang J. lldmsf: Inferring associations between long non-coding rna and disease based on multi-similarity fusion. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2936476.
74. Lan W, Li M, Zhao K, Liu J, Wu F-X, Pan Y, Wang J. Ldap: a web server for lncrna-disease association prediction. Bioinformatics. 2016;33(3):458–60.
75. Luo H, Wang J, Yan C, Li M, Fangxiang W, Yi P. A novel drug repositioning approach based on collaborative metric learning. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2926453.
76. Kong Y, Gao J, Xu Y, Pan Y, Wang J, Liu J. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. Neurocomputing. 2019;324:63–68.
77. An Y, Huang N, Chen X, Wu F, Wang J. High-risk prediction of cardiovascular diseases via attention-based deep neural networks. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2935059.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.