OXFORD

Briefings in Functional Genomics, 2016, 1–7

doi: 10.1093/bfgp/elw005
Technology review

Application of supervised machine learning algorithms for the classification of regulatory RNA riboswitches

Swadha Singh* and Raghvendra Singh*

Corresponding authors: Swadha Singh, Plant Molecular Biology Lab, CSIR-National Botanical Research Institute, Rana Pratap Marg, Lucknow, India. E-mail: swadha22@gmail.com; and Raghvendra Singh, Center of Bioinformatics, Institute of Interdisciplinary Studies, Nehru Science Center, University of Allahabad, Allahabad-211002, India. Tel.: +919453751992; E-mail: raghvendra1986singh@gmail.com
*These authors contributed equally to this work.

Abstract

Riboswitches, the small structured RNA elements, were discovered about a decade ago. It has been the subject of intense interest to identify riboswitches, understand their mechanisms of action and use them in genetic engineering. The accumulation of genome and transcriptome sequence data and comparative genomics provide unprecedented opportunities to identify riboswitches in the genome. In the present study, we have evaluated the following six machine learning algorithms for their efficiency to classify riboswitches: J48, BayesNet, Naïve Bayes, Multilayer Perceptron, sequential minimal optimization, hidden Markov model (HMM). For determining effective classifier, the algorithms were compared on the statistical measures of specificity, sensitivity, accuracy, F-measure and receiver operating characteristic (ROC) plot analysis. The classifier Multilayer Perceptron achieved the best performance, with the highest specificity, sensitivity, F-score and accuracy, and with the largest area under the ROC curve, whereas HMM was the poorest performer. At present, the available tools for the prediction and classification of riboswitches are based on covariance model, support vector machine and HMM. The present study determines Multilayer Perceptron as a better classifier for the genome-wide riboswitch searches.

Key words: riboswitches; machine learning algorithms; support vector machine; hidden Markov model; Multilayer Perceptron

Introduction

Riboswitches are complex folded structural elements located in the noncoding regions of mRNAs, which act as receptors for specific metabolic or metal ion ligand and modulate the transcription or translation of the gene/s present in transcript [1–3]. The riboswitches (50–250 nt in length) are predominantly present in the 5' untranslated region of mRNA [4, 5], and are composed of two main components: an 'aptamer region' that selectively binds to a metabolite and alters the conformation of 'expression platform', regulating the gene expression [6]. The prevalence of the RNA-based genetic control elements have been notified in prokaryotes and eukaryotes: bacteria, archaea, fungi, algae and land plants [2, 7]. It has also been recognized in human cells [8]. Riboswitches participate in several kinds of mechanisms for gene regulation [9]. They act as metabolite sensors and perform feedback regulation of several metabolic pathways in prokaryotes and eukaryotes [10]. The regulatory RNA switches that react with many different types of molecular and environmental signals are believed to facilitate in empowering the bacteria to respond to several environmental changes [4]. In addition, some of the riboswitches are known to behave like ribozymes, i.e. they self-cleave after a specific conformation is generated in the mRNA [11]. Some of the riboswitches are located in introns and influence expression via regulating mRNA splicing in eukaryotes [12].

Riboswitches influence cell physiology, differentiation and development and thus are present across all domains of life [13]. The RNA genetic control element has several advantages over other regulatory elements such as capability to operate in protein-independent manner, faster regulatory responses, easier transfer to other organisms and flexible combination of sensing and regulatory domains. Therefore, the structured mRNA elements are an attractive target for the development of small-molecule responsive gene-expression systems for gene function studies, synthetic biology and other biotechnological

Swadha Singh is Computational Biologist. She works on NGS data, machine learning algorithm to address key biological questions.
Raghvendra Singh did D.Phil. in Bioinformatics from University of Allahabad. His current interest is to explore epigenetic modulator as drug target.

applications [14, 15]. For example, riboswitches could be useful as promising targets for antibacterial and antifungal agents [16–18]. Artificial engineering of riboswitches could provide enormous opportunities for the manipulation of expression of desired genes [14, 19].

A high-level conservation in the ribonucleotide sequences of aptamer and secondary structures is useful for their computational identification in genomic sequences. Several classes of riboswitches have been discovered across many species [20]. The structural identification and distribution analysis of diverse RNA switches in the genomic databases is desirable in describing RNAs in terms of both structure and function [21]. The enormous growth of genome and transcriptome sequence data and recent studies on comparative genomics have experienced increasing number of riboswitches in the genome, and several of them (more than two dozen) have been experimentally characterized [22]. Several databases have been established compiling the information about different functional RNA molecules [23]. Some of the significant computation tools implemented for the analysis of riboswitches are Riboswitch finder [24], RibEx [25] and RiboSW [26]. These computation tools are based on covariance model (CM), support vector machine (SVM) and hidden Markov model (HMM) algorithm, respectively [24–26]. These existing programs are based on the principal of multiple sequence alignment to find out conserved sequence of previously reported riboswitches in position-independent manner. All the reported programs are limited for specific type of riboswitches. In this study, we are taking into account bi-nucleotide conservation to classify the riboswitches.

In the present study, we have evaluated the different machine learning algorithms, such as J48, BayesNet, Naïve Bayes, Multilayer Perceptron, sequential minimal optimization (SMO), HMM, for their efficiency to be used in the structural classification of riboswitches. The classification results of each algorithm were presented in confusion matrix, which shows the number of correctly and incorrectly classified instances of riboswitches. The confusion matrix was used for calculating specificity, sensitivity, accuracy, F-measure and receiver operating characteristic (ROC) plot analysis, which are the major parameters for comparing the performance evaluation of machine learning algorithms.

Material and methods

Data mining

The training database was generated by obtaining riboswitch sequences from FTP site of Rfam database, which is a comprehensive collection of noncoding RNA gene families as well as *cis* regulatory RNA elements [20]. By executing a PERL script, 16 538 riboswitch sequences were fetched with Rfam IDs, and the sequences were clustered in the 16 reported riboswitch classes (Supplementary Table S1).

Feature vector computation

Feature vector computation was done by calculating the frequency of mono- and di-nucleotides in the riboswitch sequences of each class. A total of 20 features were computed by using the PERL script. The frequency of mono- and di-nucleotides was as follows: A, T, G, C, AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG and UU. These feature vectors were used as an input for the model optimization.

Machine learning

The data mining software, Waikato Environment for Knowledge Analysis (WEKA) version 3.6.10 [27], was used for the data preprocessing, feature selection and classification experiments.

Partitioning of the data set

The feature vector was partitioned into data sets of different sizes—binary and multi-class data sets. The binary data set contained the feature vectors of two classes of riboswitch (Supplementary Table S2). The multi-class data set was further subdivided into three types: data sets with feature vectors of 4, 8 and 16 riboswitch classes (Supplementary Table S2). These divisions and subdivisions were made to examine the efficiency of the algorithms in small and large data sets.

Data preprocessing

The feature vector data set (.txt format) was converted to comma-separated value files using weka.core.converters to prepare the input files in an WEKA readable format.

Algorithms for classification

The six algorithms, J48, BayesNet, Naïve Bayes, Multilayer Perceptron, SMO and HMM (details are given in Supplementary File S3), were compared with one another for their capability to be used as a classifier of riboswitches.

Evaluation of classifiers

For the performance evaluation of all the classifiers, the confusion matrix was used to calculate sensitivity, specificity and accuracy, and to draw ROC curve [28, 29]. Efficiency of classifying algorithms in terms of positive class labels was drawn from sensitivity, whereas the negative class labels were drawn from specificity. F-score combines sensitivity and specificity by harmonic mean. The accuracy, sensitivity, specificity and F-score are calculated by the standard formula given below.

$$Accuracy = (TP + FP)/(TP + FP + TN) \quad \text{(i)}$$

$$Sensitivity = TP/(TP + FN) \quad \text{(ii)}$$

$$Specificity = TN/(FP + TN) \quad \text{(iii)}$$

$$F - score = 2 \times TP/2TP + FN + FP \quad \text{(iv)}$$

Here, *TP* denotes number of true-positive rate; *FP* denotes number of false-positive rate; *TN* denotes number of true-negative rate; and *FN* denotes number of false-negative rate.

The methodology adopted in this study has been schematically represented in Figure 1.

Results and discussion

Comparison of predictive performance

Feature selection is a prerequisite for model building, as it not only avoids the overfitting of the models, but also helps in enhancing the robustness of the selected feature subsets [30]. A total of 20 types of feature vectors were computed on the basis of the frequency of mono- and di-nucleotides in the data of each riboswitch class. The binary data set (with feature vectors of two riboswitch classes) and multi-class data sets

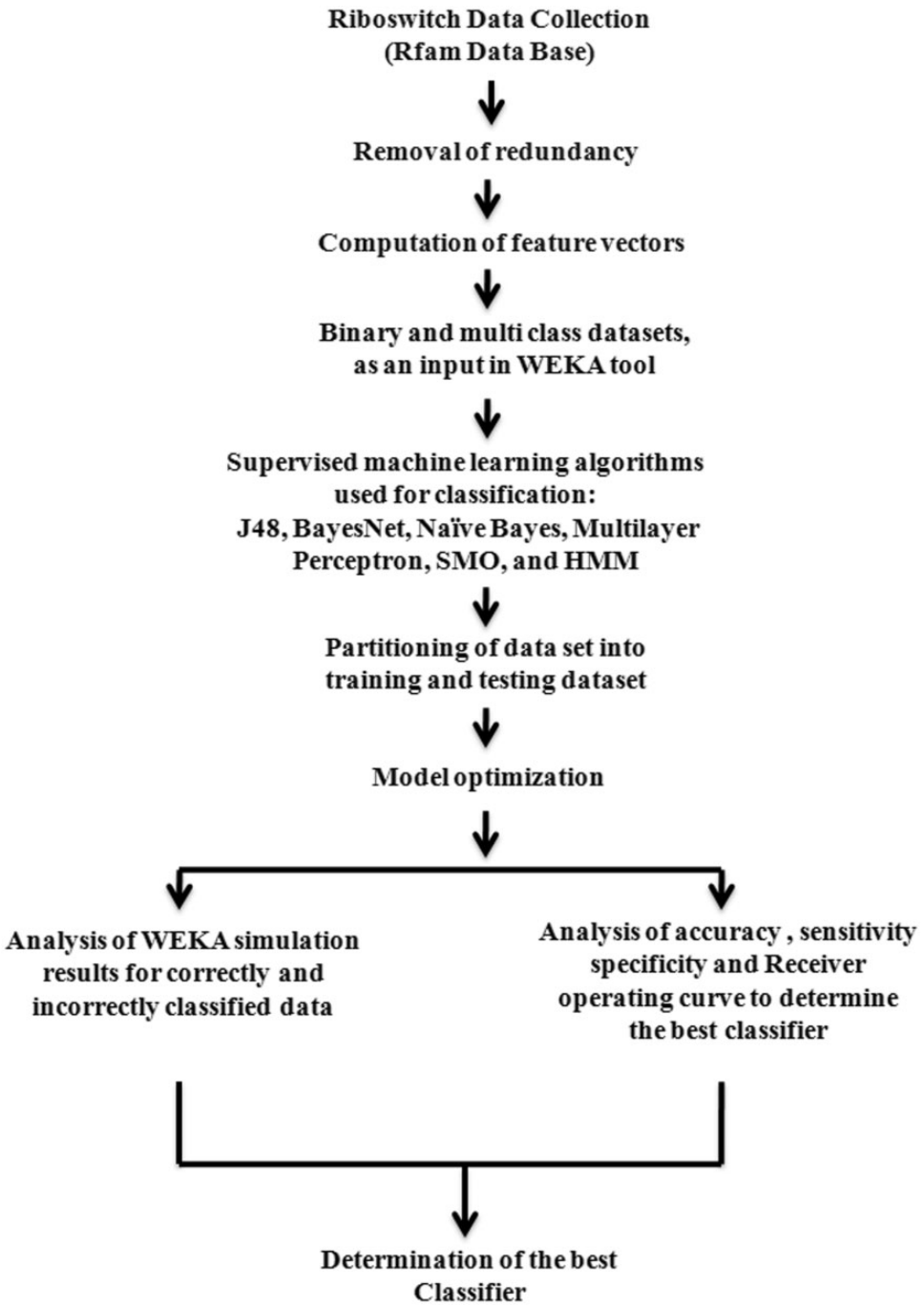

Figure 1. Schematic representation of the computational method used to compare the prediction algorithms.

(with feature vectors of 4, 8 and 16 riboswitch classes) were used as an input of WEKA tool.

The six algorithms, J48, BayesNet, Naïve Bayes, Multilayer Perceptron, SMO and HMM, were used for the classification of riboswitches. J48 follows a simple algorithm of C4.5 decision tree learning [31]. Bayonet is a probabilistic model to classify the features. Classification is done by applying Bayes theorem to compute probability of riboswitch class [32]. Naïve Bayes is simple probabilistic classifiers that work on the principal of Bayes' theorem [33]. Multilayer Perceptron is an Artificial Neural Network model, which computes a single output from multiple instances by forming linear combination [34]. SVM is a supervised learning model. SMO has advantage over SVM in terms of replacement of all missing values globally and transform nominal attributes into binary values [35]. HMM is statistical Markov model, a process where the state depends on previous states in nondeterministic way [36].

The binary and multi-class data sets were divided into training (60%) and testing (40%) data sets. The model optimization was done by the training data set, whereas the performance of the model was analyzed by the testing data set. The WEKA simulation results were examined for correctly and incorrectly classified data (Figure 2). In case of binary data set, >95% of the riboswitch candidates were correctly classified by all the

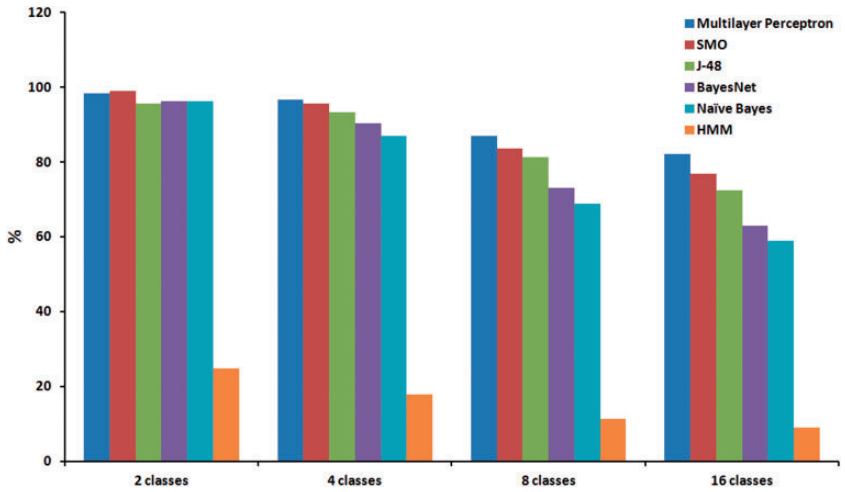

Figure 2. Comparison of the correctly classified riboswitches by the classifiers J48, BayesNet, Naïve Bayes, Multilayer Perceptron, SMO and HMM, in the data sets comprising 2, 4, 8 and 16 classes of riboswitches. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)

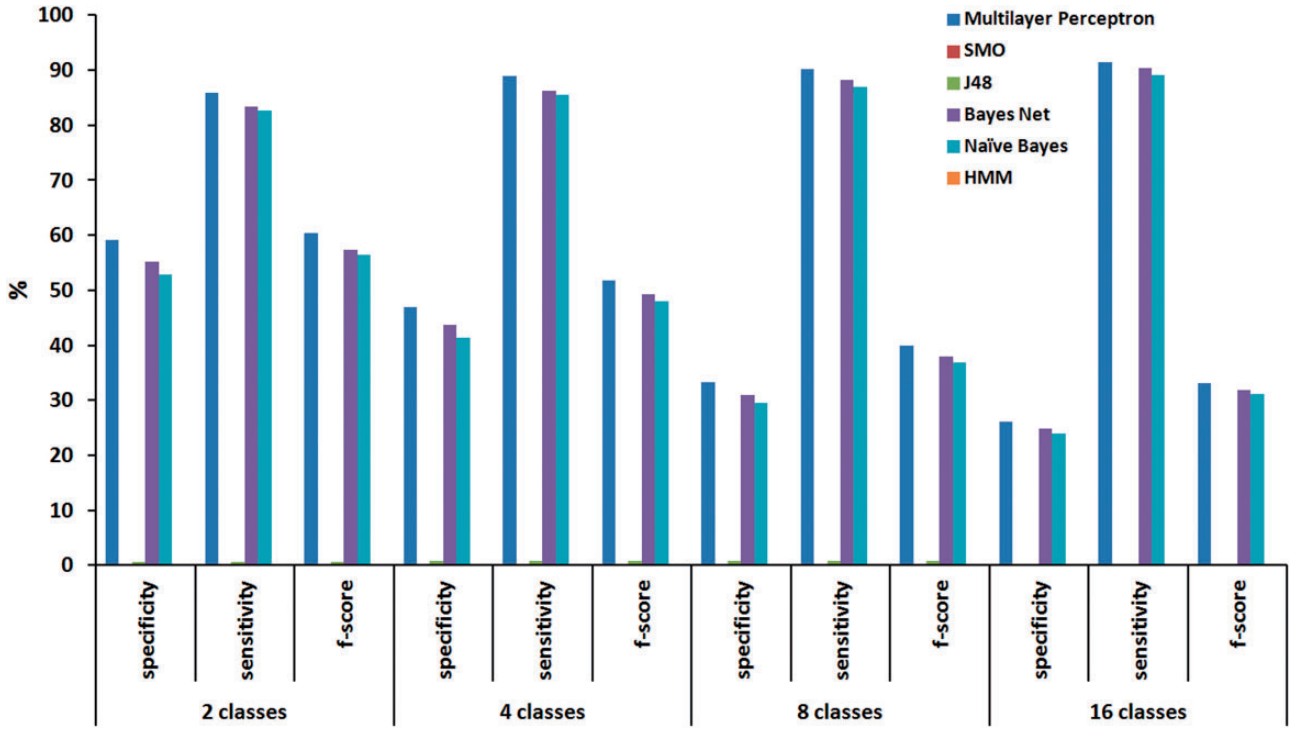

Figure 3. Comparison of specificity, sensitivity and F-score to determine the effective classifier for the identification of riboswitches in the data sets comprising 2, 4, 8 and 16 classes of riboswitches. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)

algorithms except HMM (Figure 2). However, in case of the data set with a mixture of multiple numbers (≥4) of riboswitch types, the Multilayer Perceptron and HMM were the most and least efficient, respectively, in identifying the correct classes of riboswitches (Figure 2).

To assess the performance of the six algorithms, the confusion matrix obtained from the WEKA results was used for statistical measures. The confusion matrix is a layout that represents the instances of predicted classes—true positive, false positive, false negative and true negative. The prediction performance of the six machine learning approaches was compared by calculating specificity, sensitivity and F-score (Figure 3). In these scoring parameters, the best ability to distinguish riboswitch classes was found in Multilayer Perceptron, followed by BayesNet and Naïve Bayes. In case of the data of a mix of 16 classes of riboswitches, the highest values (%) for specificity, sensitivity and F-score were 26.09, 91.42 and 33.05, respectively, in Multilayer Perceptron. However, the values for sensitivity, specificity and F-score values were significantly less in HMM, SMO and J48 (Figure 3). Average predictive accuracy was computed for each algorithm. In case of the data with maximum number of riboswitch classes, the Multilayer Perceptron exhibited highest accuracy, followed by J48 and SMO, whereas HMM was the lowest performer (Figure 4).

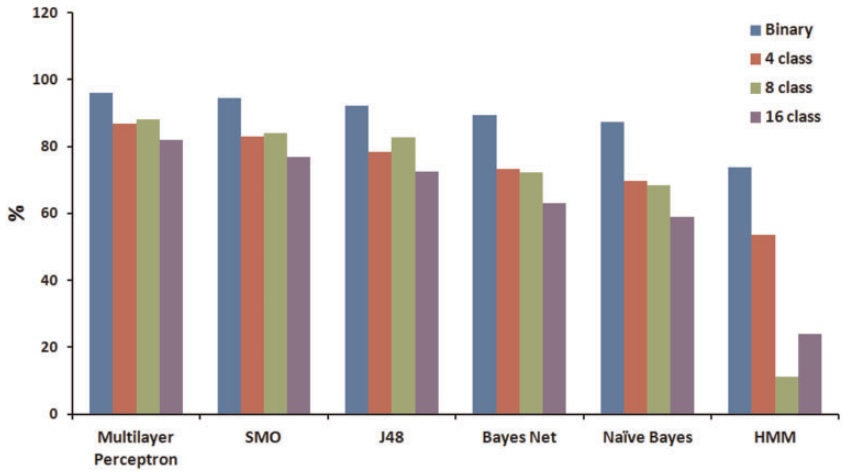

Figure 4. Comparison of accuracy to evaluate the performances of the classifiers, J48, BayesNet, Naïve Bayes, Multilayer Perceptron, SMO and HMM, in the data sets comprising 2, 4, 8 and 16 classes of riboswitches. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)

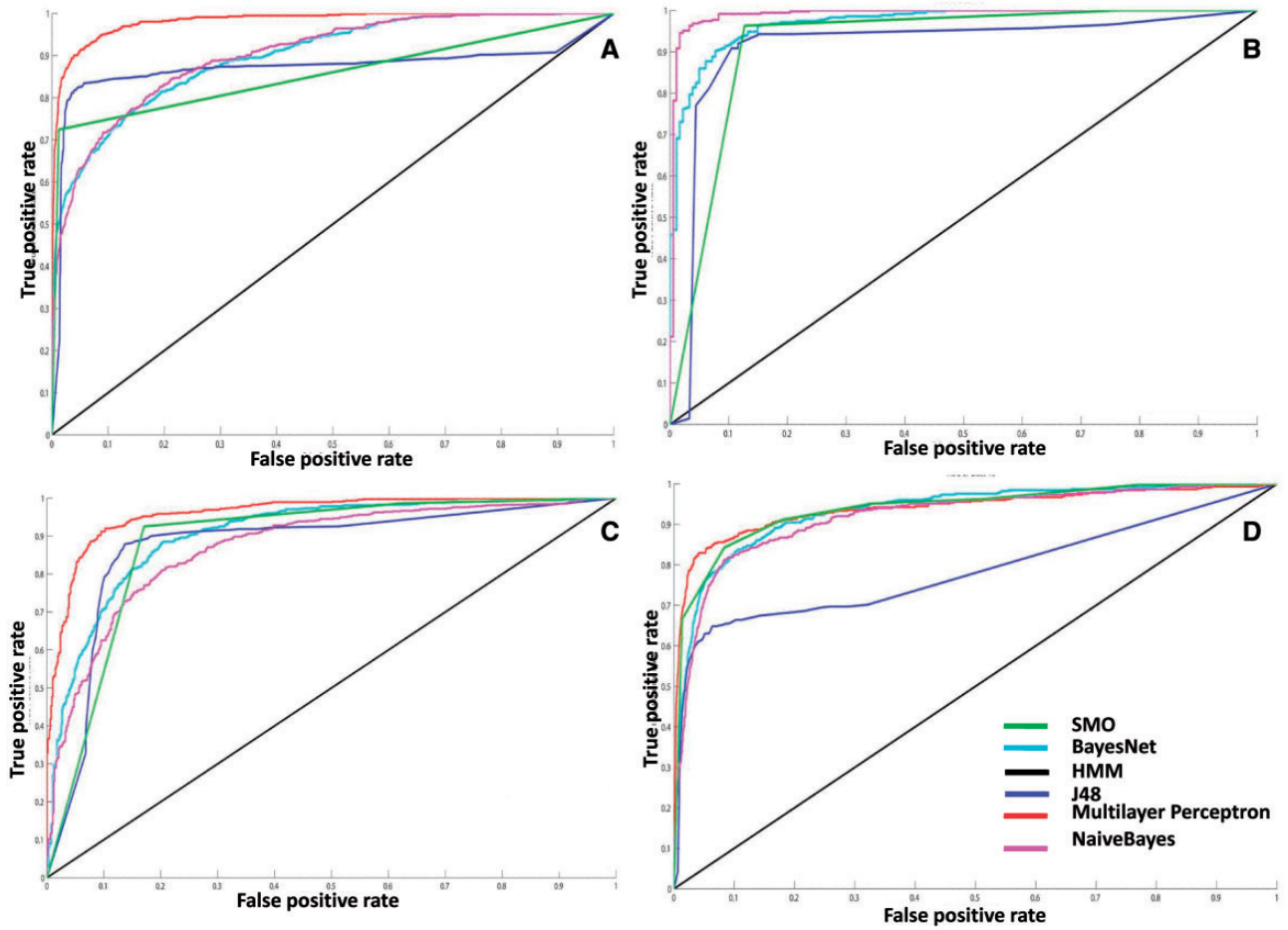

Figure 5. ROC curves for the predictive performance of the classifiers J48, BayesNet, Naïve Bayes, Multilayer Perceptron, SMO and HMM, in the data sets comprising 2 (RF00168, RF00174) (**A**), 4 (RF0105, 1RF01054, RF01055, RF01057) (**B**), 8 (RF00504, RF00521, RF00522, RF00634, RF0105, 1RF01054, RF01055, RF01057) (**C**) and 16 (Supplementary Table S1) (**D**) classes of riboswitches. The ROC curves of all the riboswitch combinations are given in Supplementary Figs S1–S3. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)

ROC graphs were plotted by linking the measurements of false-positive rate on the X-axis and the true-positive rate on the Y-axis for all the classifiers (Figure 5; Supplementary Figures S1–S3). The largest area under the curve was in the case of Multilayer Perceptron, assuring the optimal and robust performance [28] by Multilayer Perceptron, whereas the poorest performance by HMM (Figure 5A–D).

In conclusion, riboswitches are regulatory segments of mRNAs, involved in the fine-tuning of its own expression in response to the effector molecule. Although most of the known

riboswitches occur in bacteria, however, it has been discovered in plants and humans. The increasing automation in the computational biology has played a major role in the recent discoveries that happened in genomic studies. It is desirable to develop tools for the prediction and classification of riboswitches, with high accuracy. At present, the available tools are based on CM, SVM and HMM. Our study established that Multilayer Perceptron is a relatively more reliable classifier for the genome-wide riboswitch searches.

Key Points

- We have evaluated six machine learning algorithms, J48, BayesNet, Naïve Bayes, MultilayerPerceptron, sequential minimal optimization, hidden Markov model (HMM), for their efficiency to classify riboswitches.
- To determine effective classifier, the algorithms were compared on the statistical measures of specificity, sensitivity, accuracy, F-measure and receiver operating characteristic (ROC) plot analysis.
- The classifier Multilayer Perceptron achieved the best performance, with the highest specificity, sensitivity, F-score and accuracy, and with the largest area under the ROC curve, whereas HMM was the poorest performer.
- At present, the available tools for the prediction and classification of riboswitches are based on covariance model, SVM and HMM. The present study determines Multilayer Perceptron as a better classifier for the genome-wide riboswitch searches.

Supplementary data

Supplementary data are available at *Briefings in Functional Genomics and Proteomics* online.

Acknowledgement

We thankful to Dr. Sudhir P. Singh (scientist C) for critical reading and helping in rewriting of the manuscript.

References

1. Mandal M, Breaker RR. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* 2004;**5**(6):451–63.
2. Serganov A, Nudler E. A decade of riboswitches. *Cell* 2013;**152**(1):17–24.
3. Roth A, Breaker RR. The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem* 2009;**78**:305–34.
4. Chen J, Gottesman S. Riboswitch regulates RNA. *Science* 2014;**345**(6199):876–7.
5. Havill JT, Bhatiya C, Johnson SM, *et al*. A new approach for detecting riboswitches in DNA sequences. *Bioinformatics* 2014;**30**(21):3012–19.
6. Robinson CJ, Vincent HA, Wu M-C, *et al*. Modular riboswitch toolsets for synthetic genetic control in diverse bacterial species. *J Am Chem Soc* 2014;**136**(30):10615–24.
7. Sudarsan N, Barrick JE, Breaker RR. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* 2003;**9**(6):644–7.
8. Ray PS, Jia J, Yao P, *et al*. A stress-responsive RNA switch regulates VEGFA expression. *Nature* 2009;**457**(7231):915–19.
9. Peselis A, Serganov A. Themes and variations in riboswitch structure and function. *Biochim Biophys Acta* 2014;**1839**(10):908–18.
10. Bocobza S, Adato A, Mandel T, *et al*. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes Dev* 2007;**21**(22):2874–9.
11. McCown PJ, Winkler WC, Breaker RR. Mechanism and distribution of glmS ribozymes. In: *Ribozymes*. Springer, 2012, 113–29.
12. Li S, Breaker RR. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic Acids Res* 2013;**41**(5):3022–31.
13. Morris KV, Mattick JS. The rise of regulatory RNA. *Nat Rev Genet* 2014;**15**(6):423–37.
14. Berens C, Suess B. Riboswitch engineering—making the all-important second and third steps. *Curr Opin Biotechnol* 2015;**31**:10–15.
15. Groher F, Suess B. Synthetic riboswitches—a tool comes of age. *Biochim Biophys Acta* 2014;**1839**(10):964–73.
16. Lea CR, Piccirilli JA. Turning on'riboswitches to their antibacterial potential. *Nat Chem Biol* 2007;**3**(1):16–17.
17. Sudarsan N, Cohen-Chalamish S, Nakamura S, *et al*. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem Biol* 2005;**12**(12):1325–35.
18. Blount KF, Wang JX, Lim J, *et al*. Antibacterial lysine analogs that target lysine riboswitches. *Nat Chem Biol* 2007;**3**(1):44–9.
19. Wieland M, Hartig JS. Artificial riboswitches: synthetic mRNA-based regulators of gene expression. *Chembiochem* 2008;**9**(12):1873–8.
20. Burge SW, Daub J, Eberhardt R, *et al*. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 2013;**41**:D226–32.
21. Barrick JE, Breaker RR. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol* 2007;**8**(11):R239.
22. Sun EI, Leyn SA, Kazanov MD, *et al*. Comparative genomics of metabolic capacities of regulons controlled by cis-regulatory RNA motifs in bacteria. *BMC Genomics* 2013;**14**(1):597.
23. Chang T-H, Huang H-Y, Hsu JB, *et al*. An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics* 2013;**14(Suppl 2)**:S4.
24. Bengert P, Dandekar T. Riboswitch finder—a tool for identification of riboswitch RNAs. *Nucleic Acids Res* 2004;**32(Suppl 2)**:W154–9.
25. Abreu-Goodger C, Merino E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res* 2005;**33(Suppl 2)**:W690–2.
26. Chang T-H, Huang H-D, Wu L-C, *et al*. Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA* 2009;**15**(7):1426–30.
27. Hall M, Frank E, Holmes G, *et al*. The WEKA data mining software: an update. *ACM SIGKDD Explor Newslett* 2009;**11**(1):10–18.
28. Sun Y, Kamel MS, Wong AK, *et al*. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit* 2007;**40**(12):3358–78.
29. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009;**45**(4):427–37.
30. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**(19):2507–17.

31. Quinlan JR. *C4. 5: Programs for Machine Learning*. Elsevier, 2014.
32. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn* 1997;**29**(2-3):131–63.
33. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc, 1995, 338–45.
34. Rosenblatt F. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. DTIC Document, 1961.
35. Platt J. Fast training of support vector machines using sequential minimal optimization. In: *Advances in Kernel Methods—Support Vector Learning*. 1999, 3.
36. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann Math Stat* 1966;**37**:1554–63.