L2S-MirLoc: A Lightweight Two Stage MiRNA Sub-Cellular Localization Prediction Framework

Muhammad Nabeel Asim*†, Muhammad Ali Ibrahim*†, Christoph Zehe‡, Olivier Cloarec‡,
Rickard Sjogren‡, Johan Trygg§¶, Andreas Dengel*†, Sheraz Ahmed*
Email: muhammad_nabeel.asim@dfki.de
*German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
†TU Kaiserslautern, Kaiserslautern, Germany
‡Sartorius Corporate Research, Sartorius Stedim Cellca GmbH, 89081 Ulm, Germany
§ Computational Life Science Cluster (CLiC), Umeå University, 90187 Umeå, Sweden
¶ Sartorius Corporate Research, Sartorius Stedim Data Analytics, 90333 Umeå, Sweden

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534015

I. Abstract

A comprehensive understanding of miRNA sub-cellular localization may leads towards better understanding of physiological processes and support the fixation of diverse irregularities present in a variety of organisms. To date, diverse computational methodologies have been proposed to automatically infer sub-cellular localization of miRNAs solely using sequence information, however, existing approaches lack in performance. Considering the success of data transformation approaches in Natural Language Processing which primarily transform multi-label classification problem into multi-class classification problem, here, we introduce three different data transformation approaches namely binary relevance, label power set, and classifier chains. Using data transformation approaches, at 1st stage, multi-label miRNA sub-cellular localization problem is transformed into multi-class problem. Then, at 2nd stage, 3 different machine learning classifiers are used to estimate which classifier performs better with what data transformation approach for hand on task. Empirical evaluation on independent test set indicates that L2S-MirLoc selected combination based on binary relevance and deep random forest outperforms state-of-the-art performance values by significant margin.

II. Introduction

MiRNAs are special ncRNAs [1] which comprise of 22-30 nucleotides [2] and regulate over 60% of mammalian transcriptome. MiRNAs substantially contribute to diverse cellular processes by regulating gene expression [3], [4]. In combination with Argonaute proteins, miRNAs form central component of the miRNA induced silencing complex (mi-RISC) which drives the regulation of a variety of intracellular processes. MiRNAs do not only operate as an element of RISC within the cytoplasm [5], but also exist in other cellular compartments such as nucleus [6], [7], mitochondria [7], [8], nucleolus [9], extracellular vesicles [10], and exosomes [11].

Sub-cellular localization of miRNAs is very critical to their core functionality [12] especially their presence in nucleus [13] and their aptitude to direct RNA-target cleavage [14]. For example, miRNAs localized to nucleus participate in mitosis or gene expression regulation [15]. Similarly, miRNA sub-cellular localization is necessarily required to supervise several physiological processes that occur inside sub-cellular organelles such as mitochondrial metabolism is performed by mito-miRNAs, synaptic plasticity is conducted by endosomal miRNAs. MiRNAs sub-cellular localization also influences the cellular processes involved in development, proliferation, digestion, and differentiation in organisms as well as post-transcriptional regulation of genes [16]. More recently, it has been found that few miRNAs largely impact the epigenetic regulation functionality and nucleus [17], [18].

To understand diverse biological roles of miRNAs and their associations with different diseases, comprehending their sub-cellular localization is very critical. Development of computer aided program which can precisely predict sub-cellular localization of miRNA using only sequence information is need of of the hour.

To date, according to our best knowledge, there exist 2 miRNA sub-cellular localization predictors namely MIRLocator [19] and MirLocPredictor [20]. While MIRLocator [19] sequence-to-sequence neural network relies on predefined label order, MirLocPredictor [20] only takes position of amino acids into and entirely neglect the frequency of amino acids while learning the representation of miRNA sequences. The paper in hand leverages physico-chemical property based representation scheme to effectively encode miRNA sequence nucleotides.

Considering the confined performance of existing computational approaches, here, we present a robust yet lightweight two stage miRNA sub-cellular localization prediction framework namely L2S-MirLoc. Instead of treating miRNA sub-cellular localization sequences as multi-label classification problem, we introduce 3 data transformation approaches including binary relevance, label power set, classifier chain at first stage which transform the multi-label problem into multi-class problem. At second stage, L2S-MirLoc performs extensive experimentation

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

over benchmark core dataset by combining 3 data transformation approaches with 3 lightweight machine learning classifiers (random forest, support vector machine and naive bayes) to find optimal combination of data transformation and classification approach that can precisely infer the sub-cellular location of miRNA sequences. Performance of selected combination is compared with existing computational predictors over benchmark core dataset and independent test set. Contributions of this work can be summarized as:

1) Instead of using statistically rigorous and complex feature encoding scheme, utilization of a lightweight physico-chemical property based encoding scheme to learn optimal representation of miRNA sequences.
2) We introduce 3 different data transformation approaches including binary relevance, label power set, and classifier chain for miRNA sub-cellular localization task.
3) L2S-MirLoc framework performs extensive experimentation with 3 data transformation and 3 most widely used machine learning classifiers including random-forest, Support vector machine, and Naive bayes to find that which data transformation approach performs best with what classifier for miRNA sub-cellular localization task.
4) Using parameter conscious and memory efficient lightweight machine learning classifiers, a comprehensive performance comparison of L2S-MirLoc framework selected combination is performed with existing computational predictors using benchmark core dataset and independent test set in terms of 6 distinct evaluation metrics namely accuracy, precision, recall, F1-score, hamming loss, and area under receiver operating characteristics.

III. Materials and Methods

This section briefly describes three different data transformation approaches which are used to transform multi-label miRNA sequence data into multi class dataset. It also precisely discuss 3 machine learning classifiers used to evaluate the impact of data transformation approaches for miRNA sub-cellular localization prediction. Further experimental benchmark core dataset and independent test set used to evaluate the performance of 9 different experimental settings of proposed L2S-MirLoc framework are also summarized in this section.

Figure 1 illustrates the complete workflow of proposed two-stage miRNA sub-cellular localization framework L2S-MirLoc. Computational framework L2S-MirLoc operates on raw miRNA sequences and generate 1-mers by sliding a 1-dimensional window over sequences with the stride size of 1. Statistical representation of 4 basic nucleotides is learned using physico-chemical property (Electronion Interaction PseudoPotentials (EIIP)) based encoding scheme. Then, multi-label miRNA sequence vectors are passed to 3 different data transformation approaches which transform multi-label classification problem into multi-class classification problem. Finally, transformed sequence vectors are passed to 3 different machine learning classifiers which infer the sub-cellular location of miRNA sequences by extracting crucial hidden relationships between nucleotides. Finally, L2S-MirLoc selects optimal combination of data transformation and machine learning classifier L2S-MirLoc from 9 different experimental settings using F1-score as evaluation criteria considering the efficacy of F1-score over other evaluation metrics. To prove the effectiveness of optimal combination selected by L2S-MirLoc, performance of selected combination is compared with existing computational predictors using independent test set in terms of 6 different evaluation metrics for the task of miRNA sub-cellular localization prediction. Different phases of proposed computational framework L2S-MirLoc are briefly described in following subsections.

A. Learning Statistical Representation of MiRNA Sequences

Statistical representation of miRNA sequences is generated using a physico-chemical property named Electronion Interaction PseudoPotentials (EIIP) which represents the dispersion of electronion energies across biomedical sequence. EIIP statistical representation generation scheme was originally introduced by Nair et al. [21] to encode DNA sequences. As compared to other statistical representation generation schemes based on K-mer frequency, transfer learning, EIIP is very lightweight yet powerful and efficient approach as it neither requires extensive pre-training on biomedical data nor tuning of different residue position and context related hyper-parameters [22]–[25]. Many state-of-the-art statistical representation generation toolkits including Pse-in-One2.0 [22], BioSeq-Analysis2.0 [23], PyFeat [25], iLearn [24], and so-forth have applied EIIP residue values (A, 0.1260; C, 0.1340; G, 0.0806; and T, 0.1335) to represent only DNA sequences. However, more recently, Dou et al. [26] has investigated the suitability of 8 different feature encoding approaches including EIIP residue encoding scheme to effectively capture the inherent relationships of RNA sequence residues for the task of m5c modification prediction [26]. In order to generate the encodings of nucleotides present in RNA sequences using EIIP, considering Thymine (T) equivalent to Uracil (U), EIIP values of A=0.1260, C=0.1340, G=0.0806, and U=0.1335 were utilized. Building on existing work [26] and pre-dominant utilization of EIIP to locate methylation sites, promoters, and enhancers [27]–[29], here we evaluate the effectiveness of EIIP encoding scheme for miRNA sub-cellular localization prediction where a 27-dimensional vector for every miRNA sequence is obtained using following mathematical expression.

$$V_{EIIP} = [EIIP_A \oplus EIIP_C \oplus EIIP_G \oplus EIIP_U] \qquad (1)$$

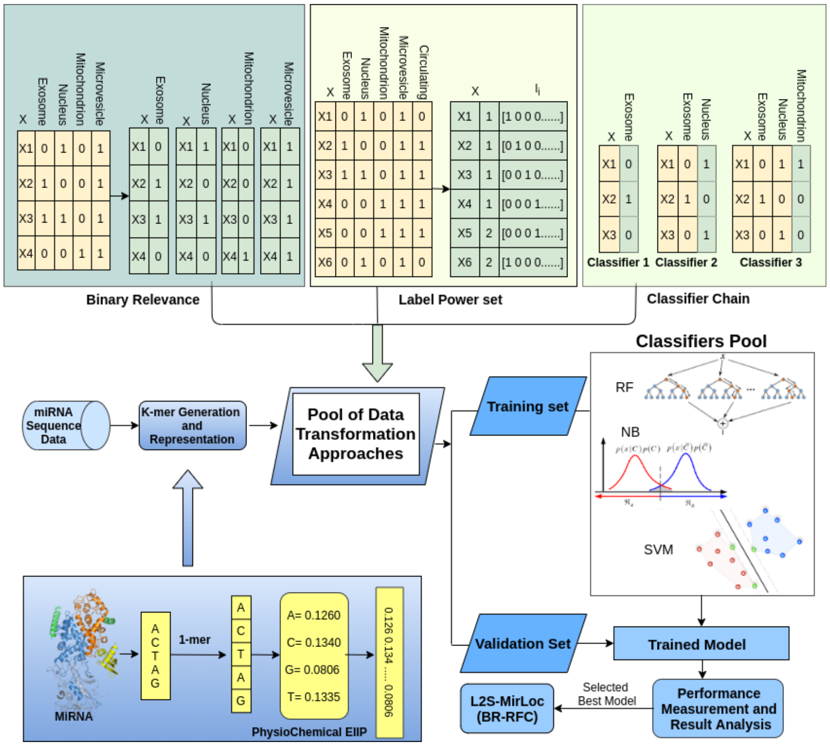


Fig. 1: Graphical illustration of L2S-MirLoc framework that facilitates to find optimal combination of data transformation and classification approach to develop data transformation based end to end multi-label classification system

B. Data Transformation Approaches

Data transformation approaches mainly transform the multi-label classification problem into multi-class classification problem without loosing miRNA sequence to sub-cellular location relationship and interdependencies between sub-cellular locations. Over the period, several data transformation approaches have been presented such as Binary Relevance [30], Ranking using Pairwise Comparison [31], Label Powerset [32], Calibrated Ranking using Pairwise Comparison [33], and Classifier Chains [34]. Considering the wide success of 3 problem transformation approaches including Binary Relevance [30], Classifier Chains [34], and Label Powerset [32] in NLP, here L2S-MirLoc computational framework leverage only these 3 approaches to transform multi-label miRNA sub-cellular localization problem into multi-class problem. All 3 data transformation approaches are briefly described below.

1) Binary Relevance: In Binary relevance [30] methodology, a multi-label categorization problem is divided into several binary categorization problems. Afterwards, for each binary classification problem, a binary classification algorithm is trained which predicts the presence or absence of respective class label. At the end, all binary predictions are concatenated to generate multiple labels for given corpus instance. The only problem binary relevance approach faces is that it does not take into account label dependencies. To illustrate binary relevance better, a hypothetical example is given in Figure 1.

2) Label PowerSet: In Label Power-set methodology [32] a multi-label problem is transformed into traditional multi-class problem by assigning a new single label against every distinct combination of corpus labels. Then, traditional classifier is trained and during inference predicted label is mapped again to a set of pre-defined labels. This approach takes into account label interdependencies but for datasets that have huge label cardinality, a handful of single labels get very few samples for training. To illustrate label power set better, a hypothetical example is given in Figure 1

3) Classifier Chains: Classifier Chains [34] approach resolves the labels co-relation problem faced by binary relevance technique. In this method, all binary classifiers are connected in random order, and the labels predicted by previous classifier is incorporated as additional information in subsequent classifier. This is achieved by expanding the feature vector which is linked to each classifier with the values of former labels in randomly-order chain during training phase. Nevertheless as the labels are randomly ordered, thus it can cause poor categorization performance. To illustrate classifier chains better, a hypothetical example is given in Figure 1

C. Machine Learning Classifiers

After transforming multi-label miRNA sub-cellular localization problem into multi-class problem, L2S-MirLoc makes use of 3 most widely used classifiers including deep random forest, support vector machine and naive bayes.

Deep forest consists of a collection of random forests where every random forest is trained on 27-dimensional sequence vectors to obtain a 2-dimensional vector containing class probabilities. Unlike deep neural networks which demands comprehensive training examples and a careful hyper-parameter tuning, deep forest proves really effective in acquiring hyper-level representation at low cost. Deep forest has shown great performance in diverse bioinformatics tasks [35]–[37]. Building on the significance of aggregating diversified estimators in meta-learning [38], [39], deep forest combines 100 random forests where estimator determined distribution is merged with input features before sending forward as inputs to following layer. In order to generate final prediction, Maximum agglomerated figure computed over all 2-dimensional sub-cellular location probability vectors is treated as final prediction to evaluate the performance of deep forest.

Support vector machine is another estimator widely used for multifarious tasks such as outlier detection, regression, and classification in diverse fields of genomics, proteomics, bioinfomormatics and NLP [40]–[42]. It is classified as a discriminative classifier because it maps every instance as a co-ordinate in a high-dimensional space and distinguish different class groups using hyperplane. Further, L2S-MirLoc leverages Naive bayes classifier based on bayes theorem which is extensively utilized in multifarious NLP and bioinformatics tasks.

D. Benchmark Dataset

To assess the performance of proposed two stage miRNA sub-cellular location prediction approaches, we leverage a public benchmark corpus and independent test set given by Xiao et al. [19] and Asim et al. [20].

Using RNA locate database, they compiled miRNA sub-cellular localization dataset having 1,048 sequences annotated against Exosome, Mitochondrion, Cytoplasm, Circulating, Microvesicle, and Nucleus. In benchmark dataset and independent test set, each human miRNA sequence consists of 4 nucleotides A, U, C, G and average sequence length lies around 27 nucleotides.

To provide insights into label cardinality and density, a comprehensive multi-dimensional analysis of benchmark dataset is carried, findings of which are presented in Table I and graphical illustrations 2. Table I summarizes the distribution of 6 sub-cellular locations and sequence identity distribution in terms of respective cellular compartments. Whereas, in Figure 2, pie chart illustrates that most of the eukaroytic miRNA sequences belong to single sub-cellular compartment succeeded by bi-cellular miRNA sequences which make up of 0.2% of total eukaroytic miRNA sequences. In order to analyze miRNA sequence-subcellular location distribution, bar graph of Figure 2 indicates the total count of miRNA samples belonging to every sub-cellular location in terms of unique colors. Finally, to facilitate the significant information concerning how often different sub-cellular locations have shown up jointly in benchmark core dataset, bi-subcellular and tri-subcellular localization based confusion matrix are given in Figure 2.

MiRNA Location Distribution					
Exosome	Cytoplasm	Mitochondrion	Microvesicle	Circulating	Nucleus
869	209	338	348	513	349
MiRNA Sequence Identity Distribution					
Uni-Label	Bi-Label	Tri-Label	Tetra-Label	Penta-Label	Hexa-Label
424	233	128	78	64	120

TABLE I: Characteristics of Benchmark MiRNA Sub-Cellular Localization Dataset [19]

Considering the fact that addition of new miRNA sequences is inevitable in genomic repositories, following the process given by Xiao et al. [19], Asim et al. [20] acquired miRNA sequence IDs from RNALocate database which were inserted over the duration of 2 years, using which new miRNA sequences were extracted from miRBase database. Using newly acquired sequences, Asim et al. [20] prepared an independent test set for the task of miRNA sub-cellular localization prediction. Mainly, this independent test set has 77 miRNA sequences where 45 miRNA sequences belong to precisely uni-cellular compartment, 16 miRNA sequences belong to bi-cellular compartments, 8 of them belong to tri-cellular compartments, 5 miRNA sequences belong to tetra-cellular compartments, 2 miRNA sequences belong to penta-cellular compartments, and only 1 miRNA sequence belong to hexa-cellular compartments.

E. Evaluation Criterion

Following the evaluation criteria used by Asim et al. [20], here we use 6 different example based evaluation measures namely accuracy, precision, recall, F1-score, hamming loss and area under receiver operating characteristics, details of which can be seen from Asim et al. [20] work.

IV. Experimental Setup

L2S-MirLoc computational framework is developed for the task of miRNA sub-cellular localization prediction using python and scikit-learn application programming interface (API). L2S-MirLoc framework supports the selection of a variety of kernels and grid-search based parameter optimization. For the hand on task, we use support vector machine with rbf kernel, regularization parameter value of 1.0, and degree of 3. For naive bayes classifier, alpha of 1.0 is used, whereas for random forest, 100 estimators are used with gini as split criteria.

We have experimented with 3 data transformation approaches using 3 different classifiers which make up to 9 different combination for end-to-end multi-label miRNA sub-cellular localization prediction. In order to better

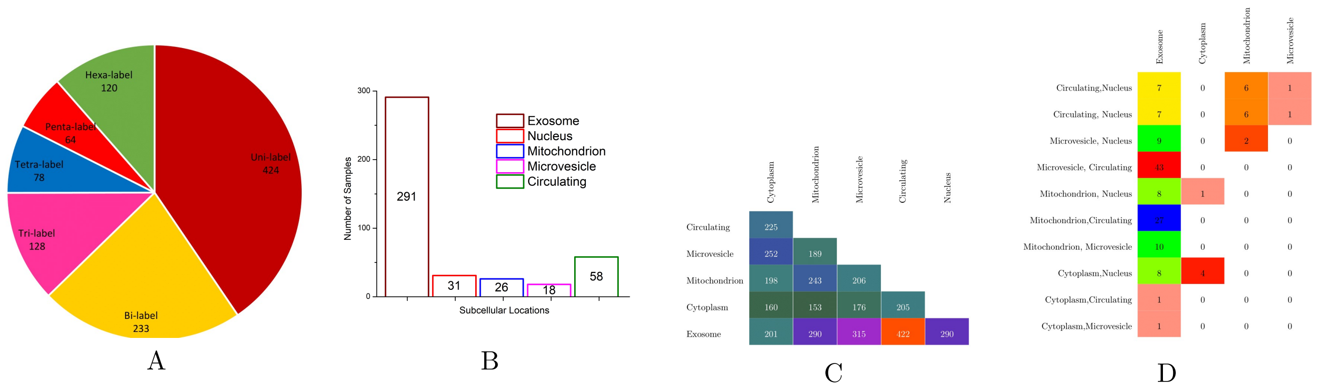


Fig. 2: (A) Descriptive Statistic of Core MiRNA subCellular Localization Dataset Segregation of MiRNA Sequences in terms of Label Cardinality (B) MiRNA Sequences Belong to every Sub-cellular Location (C) Dense Bi-Label Confusion Matrix (D) Dense Tri-Label Confusion Matrix

illustrate the results of extensive experimentation, as a naming convection, acronym of each data transformation approach is combined with precise name of every classifier. For example, when miRNA sequences are transformed using binary relevance approach at 1st stage and support vector machine classifier is used at 2nd stage for miRNA sub-cellular location prediction, then this combination is represented as BR-SVC, for naive bayes classifier BR-NB, and for random forest classifier BR-RFC. Similar naming convention is used for combinations produced by 2 other data transformation approaches.

V. Results

This section performs a comprehensive performance comparison of 9 different experimental settings by applying 10-fold cross validation over benchmark core dataset. Optimal combination of data transformation and machine learning classifier is found from 9 different experimental settings (generated by the combinations of 3 data transformation approaches with 3 classifiers) using F1-score as evaluation metric considering its efficacy over other evaluation metrics. After selecting best combination, to prove the integrity of best combination, a fair performance comparison with 8 other methods is also performed over independent test set.

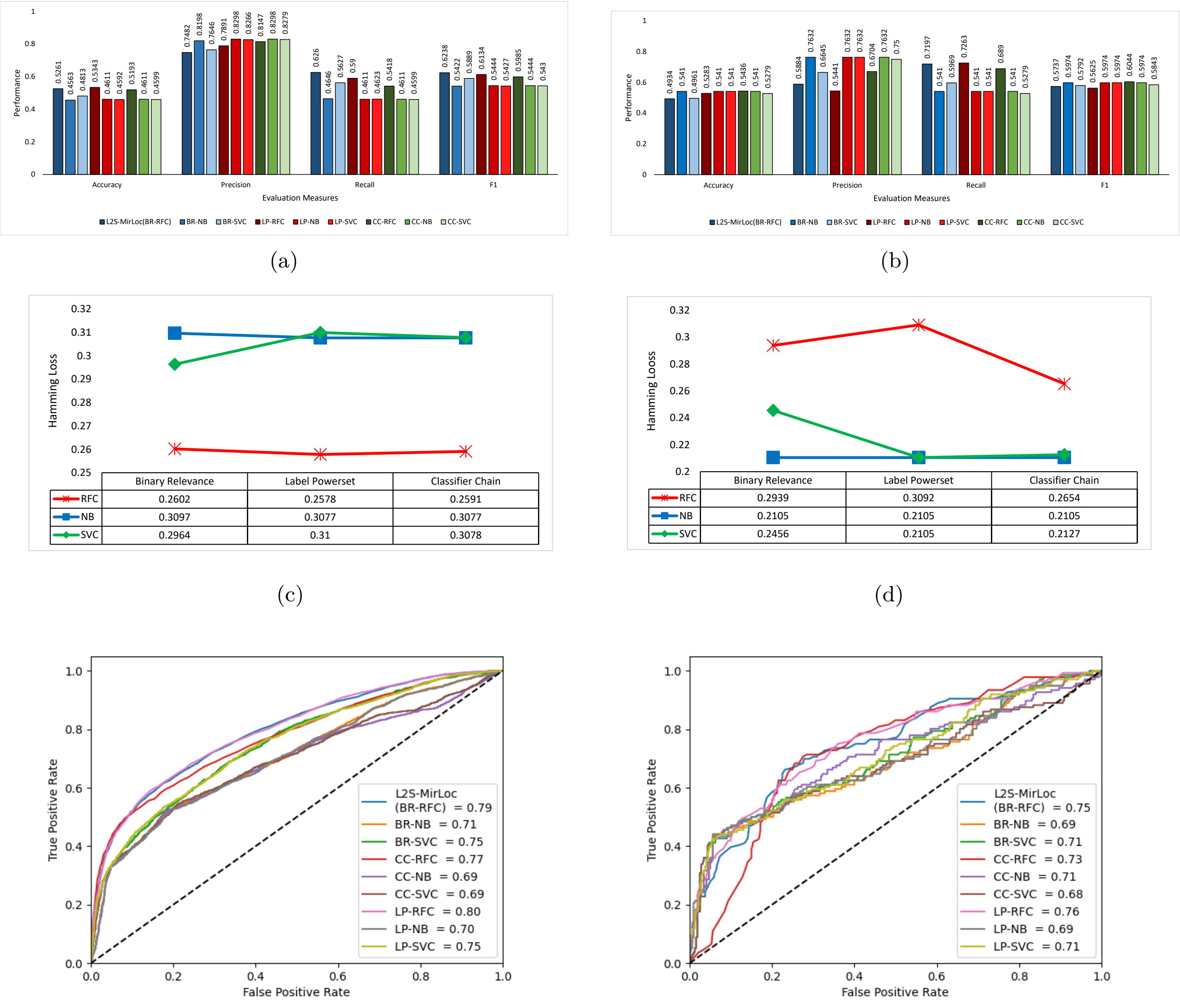
Figures 3a, and 3b show the performance of 9 different settings across both core dataset and independent test set in terms of accuracy, precision, recall, and F1-score. As extensive experimentation over benchmark core dataset using L2S-MirLoc framework reveals that from all 9 combinations, binary relevance data transformation approach in combination with deep random forest produces best performance across most evaluation metrics, therefore this optimal combination is referred as L2S-MirLoc (BR-RFC) in following discussion.

It is evident from the Figure 3a that, over core miRNA sub-cellular localization dataset, all 3 data transformation approaches achieve top accuracy, recall, and F1-score with random-forest classifier followed by SVM classifier. Whereas, all 3 data transformation approaches attain better precision using naive bayes classifier. From all 3 data transformation approaches, binary relevance achieve better performance as compared to label power set and classifier chains.

Among all 9 different combinations, L2S-MirLoc (BR-RFC) achieves best performance followed by label power set using random forest (LP-RFC) across most evaluation metrics. Although Label power set using random forest classifier (LP-RFC) achieves slightly better accuracy and precision, however it is biased towards type I and type II error. This is why it attains a lower F1-score than L2S-MirLoc (BR-RFC). Overall, L2S-MirLoc (BR-RFC) marks more stable and better performance.

On the other hand, over independent test set (Figure 3b), binary relevance achieves better performance using naive bayes classifier (BR-NB) across most evaluation metrics. Label power set marks similar performance with naive bayes (LP-NB) and SVM (LP-SVC) classifier, slightly better than random forest (LP-RFC) classifier. Classifier chain marks better accuracy and precision with naive bayes (CC-NB) classifier, recall and F1-score with random forest (CC-RFC) classifier. Although label power set using random forest classifier (LP-RFC) marks slightly better accuracy and recall than L2S-MirLoc (BR-RFC), however F1-score comparison proves the biaseness of former approach towards Type I error.

Further, performance of 3 different data transformation approaches is compared using random forest, support vector machine and naive bayes classifiers in terms of hamming loss (Figures 3c, 3d). Analysis of hamming loss figures produced by diverse approaches over benchmark core dataset and independent test set (Figures 3c, 3d) reveals that among all classifiers, random forest achieves lowest hamming loss figures across both datasets. Further, random forest marks better hamming loss with classi-



(a)

(b)

	Binary Relevance	Label Powerset	Classifier Chain
RFC	0.2602	0.2578	0.2591
NB	0.3097	0.3077	0.3077
SVC	0.2964	0.31	0.3078

(c)

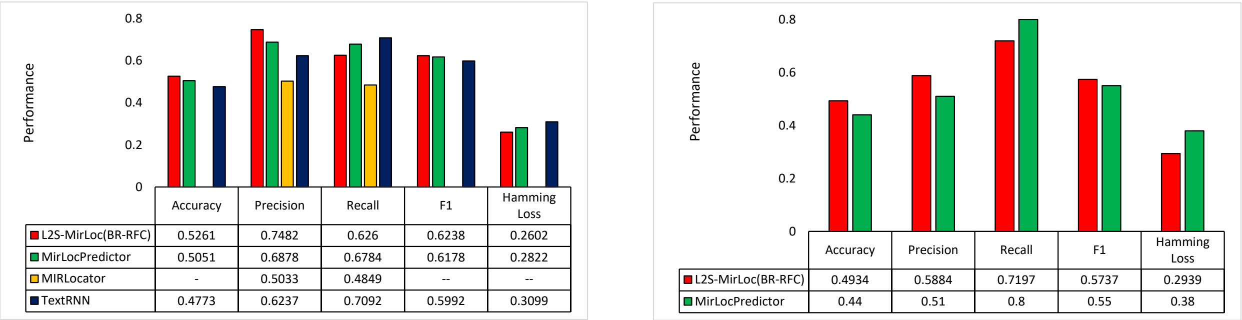
	Binary Relevance	Label Powerset	Classifier Chain
RFC	0.2939	0.3092	0.2654
NB	0.2105	0.2105	0.2105
SVC	0.2456	0.2105	0.2127

(d)

(e)

(f)

Fig. 3: Performance Comparison of 9 different data transformation based classification approaches in terms of accuracy precision, recall and f1 on core and independent test set respectively (A-B). Comparison of hamming loss on core and independent dataset respectively (C-D). (E-F) illustrates AU-ROC values produced by 9 approaches over core and independent dataset respectively.



	Accuracy	Precision	Recall	F1	Hamming Loss
■ L2S-MirLoc(BR-RFC)	0.5261	0.7482	0.626	0.6238	0.2602
■ MirLocPredictor	0.5051	0.6878	0.6784	0.6178	0.2822
■ MIRLocator	-	0.5033	0.4849	--	--
■ TextRNN	0.4773	0.6237	0.7092	0.5992	0.3099

	Accuracy	Precision	Recall	F1	Hamming Loss
■ L2S-MirLoc(BR-RFC)	0.4934	0.5884	0.7197	0.5737	0.2939
■ MirLocPredictor	0.44	0.51	0.8	0.55	0.38

(a) Core Dataset

(b) Independent Test Set

Fig. 4: (A-B) Performance Comparison of L2S-MirLoc (BR-RFC) with Existing MiRNA Sub-Cellular Localization Predictors using Core Dataset and Independent Test Set in terms of 5 Evaluation Metrics.

fier chain transformation approach. While Naive bayes achieves almost same hamming loss across both core dataset and independent test set, support vector machine achieves lowest hamming loss using binary relevance and label powerset over core dataset and independent test set respectively. From all data transformation approaches, overall classifier chain marks slightly better hamming loss than label power set and binary relevance approaches.

In order to prove the generalize ability of proposed L2S-MirLoc (BR-RFC) based on binary relevance and deep random forest, performance of binary relevance, label power set, classifier chain is compared in terms of area under receiver operating characteristics (AU-ROC) (Figures 3e, 3f). Figure 3e depicts that, over benchmark core dataset, all 3 data transformation approaches attain top AU-ROC figure using random forest classifier followed by SVM classifier. Performance of label power set using random forest classifier and proposed L2S-MirLoc (BR-RFC) based on binary relevance go almost hand in hand in terms of AU-ROC. Overall, both binary relevance and label power set attain similar degree of separability across all 3 classifiers and both achieve the peak of almost 80% using random forest classifier. Likewise AU-ROC trend analysis for independent test set (Figure 3f) reveals that, both binary relevance and label power set achieve the peak AU-ROC score of 75% using random-forest classifier as compared to 72% achieved by classifier chain approach. Further, binary relevance and label power set attain almost same degree of separability across all 3 classifiers. Overall, from all data transformation approaches, binary relevance and label power set mark better performance that classifier chains. Whereas, among all classifiers, random forest takes the lead followed by SVM classifier.

In a nutshell, physico-chemical property based nucleotide encoding scheme effectively captures comprehensive relationships between nucleotides which eventually help the classifiers in identifying sub-cellular compartments of miRNA sequences. Among all data transformation approaches, overall binary relevance achieves better performance over both core dataset and independent test. From all machine learning classifiers, random forest achieves best performance followed by SVM across most evaluation metrics.

A. L2S-MirLoc (BR-RFC) performance comparison with state-of-the-art miRNA subcellular location prediction approaches

To prove the integrity of optimal combination of data transformation and machine learning classifier L2S-MirLoc (BR-RFC), here we compare the performance of L2S-MirLoc (BR-RFC) with existing computational approaches for the task of miRNA sub-cellular localization prediction.

Figures 4a, and 4b compare the performance of L2S-MirLoc (BR-RFC) with existing computational predictors using benchmark core miRNA sub-cellular localization dataset and independent test set in terms of accuracy, precision, recall, F1-score, and hamming loss. As is indicated by the Figure 4a, over benchmark core miRNA sub-cellular localizataion dataset, L2S-MirLoc (BR-RFC) outshines state-of-the-art MirLocPredictor [20] by the accuracy figure of 2% precision figure of 6% F1-score of 1%, and hamming loss of 2%. Further, over core dataset, L2S-MirLoc (BR-RFC) outshines MIRLocator [19] by the promising figure of 25% and 13% in terms of precision and recall respectively. Whereas on benchmark independent miRNA sub-cellular localization test set (Figure 4b), L2S-MirLoc (BR-RFC) outperforms previous best MirLocPredictor [20] across 4 evaluation metrics with even more promising margin. More specifically, it attains the accuracy improvement of 5%, precision improvement of 8%, F1-score improvement of 2%, and hamming loss improvement of 9%.

VI. Conclusion

Considering the limited performance and deficiency in generalization ability of existing computational predictors, this paper comprehensively evaluates the performance impact of multi-label problem transformation approaches using 3 different classifiers to find optimal combination for the task of miRNA sub-cellular localization prediction. L2S-MirLoc (BR-RFC) leverages very lightweight physico-chemical property based encoding, binary relevance for data transformation, and deep forest for accurate sub-cellular localization of miRNAs. Performance comparison of L2S-MirLoc (BR-RFC) with existing computational MiRNA sub-cellular location predictors using benchmark core dataset and independent test set indicates that proposed L2S-MirLoc (BR-RFC) outperforms physico-chemical property based, nucleotide frequency based, and pre-trained k-mer embeddings based predictors with decent margin in terms of most evaluation metrics. We believe that this study will also supplement the sub-cellular localization research related to other ncRNAs. Further, computational framework L2S-MirLoc (BR-RFC) can be used to find optimal combination of data transformation and machine learning classifier for diverse genomics, proteomics, and bioinformatics tasks.

VII. Acknowledgement

This work is supported by Sartorius Artificial Intelligence Lab (SAIL) project. We thank all members of the Deep Learning Competence Center at the DFKI for their comments and support.

References

[1] M. R. Friedländer, E. Lizano, A. J. Houben, D. Bezdan, M. Báñez-Coronel, G. Kudla, E. Mateu-Huertas, B. Kagerbauer, J. González, K. C. Chen et al., “Evidence for the biogenesis of more than 1,000 novel human micrornas,” Genome biology, vol. 15, no. 4, pp. 1–17, 2014.

[2] V. N. Kim, “Microrna biogenesis: coordinated cropping and dicing,” Nature reviews Molecular cell biology, vol. 6, no. 5, pp. 376–385, 2005.

[3] N. Guzman, K. Agarwal, D. Asthagiri, L. Yu, M. Saji, M. D. Ringel, and M. E. Paulaitis, “Breast cancer–specific mir signature unique to extracellular vesicles includes “microrna-like” trna fragments,” Molecular Cancer Research, vol. 13, no. 5, pp. 891–901, 2015.

[4] S. Jonas and E. Izaurralde, “Towards a molecular understanding of microrna-mediated gene silencing,” Nature reviews genetics, vol. 16, no. 7, pp. 421–433, 2015.

[5] M. A. Carmell, Z. Xuan, M. Q. Zhang, and G. J. Hannon, “The argonaute family: tentacles that reach into rnai, developmental control, stem cell maintenance, and tumorigenesis,” Genes & development, vol. 16, no. 21, pp. 2733–2742, 2002.

[6] C. D. Jeffries, H. M. Fried, and D. O. Perkins, “Nuclear and cytoplasmic localization of neural stem cell micrornas,” Rna, vol. 17, no. 4, pp. 675–686, 2011.

[7] Z. F. Li, Y. M. Liang, P. N. Lau, W. Shen, D. K. Wang, W. T. Cheung, C. J. Xue, L. M. Poon, and Y. W. Lam, “Dynamic localisation of mature micrornas in human nucleoli is influenced by exogenous genetic materials,” PLoS One, vol. 8, no. 8, p. e70869, 2013.

[8] J. A. Makarova, M. U. Shkurnikov, D. Wicklein, T. Lange, T. R. Samatov, A. A. Turchinovich, and A. G. Tonevitsky, “Intracellular and extracellular microrna: an update on localization and biological role,” Progress in histochemistry and cytochemistry, vol. 51, no. 3-4, pp. 33–49, 2016.

[9] J. C. R. Politz, E. M. Hogan, and T. Pederson, “Micrornas with a nucleolar location,” Rna, vol. 15, no. 9, pp. 1705–1715, 2009.

[10] M. Mittelbrunn, C. Gutiérrez-Vázquez, C. Villarroya-Beltri, S. González, F. Sánchez-Cabo, M. Á. González, A. Bernad, and F. Sánchez-Madrid, “Unidirectional transfer of microrna-loaded exosomes from t cells to antigen-presenting cells,” Nature communications, vol. 2, no. 1, pp. 1–10, 2011.

[11] A. Turchinovich, T. R. Samatov, A. G. Tonevitsky, and B. Burwinkel, “Circulating mirnas: cell–cell communication function?” Frontiers in genetics, vol. 4, p. 119, 2013.

[12] A. K. Leung, “The whereabouts of microrna actions: cytoplasm and beyond,” Trends in cell biology, vol. 25, no. 10, pp. 601–610, 2015.

[13] S. A. Khudayberdiev, F. Zampa, M. Rajman, and G. Schratt, “A comprehensive characterization of the nuclear microrna repertoire of post-mitotic neurons,” Frontiers in molecular neuroscience, vol. 6, p. 43, 2013.

[14] K. T. Gagnon, L. Li, Y. Chu, B. A. Janowski, and D. R. Corey, “Rnai factors are present and active in human cell nuclei,” Cell reports, vol. 6, no. 1, pp. 211–221, 2014.

[15] C. Catalanotto, C. Cogoni, and G. Zardo, “Microrna in control of gene expression: an overview of nuclear functions,” International journal of molecular sciences, vol. 17, no. 10, p. 1712, 2016.

[16] V. Ambros, “The functions of animal micrornas,” Nature, vol. 431, no. 7006, p. 350, 2004.

[17] Q. Yao, Y. Chen, and X. Zhou, “The roles of micrornas in epigenetic regulation,” Current opinion in chemical biology, vol. 51, pp. 11–17, 2019.

[18] M. Fabbri, F. Calore, A. Paone, R. Galli, and G. A. Calin, “Epigenetic regulation of mirnas in cancer,” Epigenetic Alterations in Oncogenesis, pp. 137–148, 2013.

[19] Y. Xiao, J. Cai, Y. Yang, H. Zhao, and H. Shen, “Prediction of microrna subcellular localization by using a sequence-to-sequence model,” in 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018, pp. 1332–1337.

[20] M. N. Asim, M. I. Malik, C. Zehe, J. Trygg, A. Dengel, and S. Ahmed, “Mirlocpredictor: A convnet-based multi-label microrna subcellular localization predictor by incorporating k-mer positional information,” Genes, vol. 11, no. 12, p. 1475, 2020.

[21] A. S. Nair and S. P. Sreenadhan, “A coding measure scheme employing electron-ion interaction pseudopotential (eiip),” Bioinformation, vol. 1, no. 6, p. 197, 2006.

[22] B. Liu, H. Wu, K.-C. Chou et al., “Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences,” Natural Science, vol. 9, no. 04, p. 67, 2017.

[23] B. Liu, X. Gao, and H. Zhang, “Bioseq-analysis2. 0: an updated platform for analyzing dna, rna and protein sequences at sequence level and residue level based on machine learning approaches,” Nucleic acids research, vol. 47, no. 20, pp. e127–e127, 2019.

[24] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb et al., “ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of dna, rna and protein sequence data,” Briefings in bioinformatics, vol. 21, no. 3, pp. 1047–1057, 2020.

[25] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, and A. Dehzangi, “Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences,” Bioinformatics, vol. 35, no. 19, pp. 3831–3833, 2019.

[26] L. Dou, X. Li, H. Ding, L. Xu, and H. Xiang, “Prediction of m5c modifications in rna sequences by combining multiple sequence features,” Molecular Therapy-Nucleic Acids, vol. 21, pp. 332–342, 2020.

[27] Q. Tang, F. Nie, J. Kang, and W. Chen, “ncpro-ml: An integrated computational tool for identifying non-coding rna promoters in multiple species,” Computational and structural biotechnology journal, vol. 18, pp. 2445–2452, 2020.

[28] W. He, C. Jia, and Q. Zou, “4mcpred: machine learning methods for dna n4-methylcytosine sites prediction,” Bioinformatics, vol. 35, no. 4, pp. 593–601, 2019.

[29] W. He and C. Jia, “Enhancerpred2. 0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection,” Molecular Biosystems, vol. 13, no. 4, pp. 767–774, 2017.

[30] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, “Learning multi-label scene classification,” Pattern recognition, vol. 37, no. 9, pp. 1757–1771, 2004.

[31] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, “Label ranking by learning pairwise preferences,” Artificial Intelligence, vol. 172, no. 16-17, pp. 1897–1916, 2008.

[32] G. Tsoumakas and I. Vlahavas, “Random k-labelsets: An ensemble method for multilabel classification,” in European conference on machine learning. Springer, 2007, pp. 406–417.

[33] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, “Multilabel classification via calibrated label ranking,” Machine learning, vol. 73, no. 2, pp. 133–153, 2008.

[34] J. Read, B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” Machine Learning, vol. 85, no. 3, p. 333, Jun 2011. [Online]. Available: https://doi.org/10.1007/s10994-011-5256-5

[35] W. Wang, Q. Dai, F. Li, Y. Xiong, and D.-Q. Wei, “Mlcdforest: multi-label classification with deep forest in disease prediction for long non-coding rnas,” Briefings in Bioinformatics, 2020.

[36] B. Yu, C. Chen, Z. Yu, A. Ma, B. Liu, and Q. Ma, “Prediction of protein-protein interactions based on elastic net and deep forest,” bioRxiv, 2020.

[37] X. Zeng, Y. Zhong, W. Lin, and Q. Zou, “Predicting disease-associated circular rnas using deep forests combined with positive-unlabeled learning methods,” Briefings in bioinformatics, vol. 21, no. 4, pp. 1425–1436, 2020.

[38] J. Zhang, Z. Li, and S. Chen, “Diversity aware-based sequential ensemble learning for robust anomaly detection,” IEEE Access, vol. 8, pp. 42 349–42 363, 2020.

[39] O. Sagi and L. Rokach, “Ensemble learning: A survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1249, 2018.

[40] N. Stephenson, E. Shane, J. Chase, J. Rowland, D. Ries, N. Justice, J. Zhang, L. Chan, and R. Cao, “Survey of machine learning techniques in drug discovery,” Current drug metabolism, vol. 20, no. 3, pp. 185–193, 2019.

[41] B. Bağiröz, E. Doruk, and O. Yildiz, “Machine learning in bioinformatics: Gene expression and microarray studies,” in 2020 Medical Technologies Congress (TIPTEKNO). IEEE, 2020, pp. 1–4.

[42] Y. Jiao and P. Du, “Performance measures in evaluating machine learning based bioinformatics predictors for classifications,” Quantitative Biology, vol. 4, no. 4, pp. 320–330, 2016.