Methods 203 (2022) 378–382

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

# WHISTLE server: A high-accuracy genomic coordinate-based machine learning platform for RNA modification prediction

Lian Liu [a,1], Bowen Song [b,e,1], Kunqi Chen [c], Yuxin Zhang [c], João Pedro de Magalhães [e], Daniel J. Rigden [d], Xiujuan Lei [a,*], Zhen Wei [c,d,*]

[a] School of Computer Sciences, Shannxi Normal University, Xi'an, Shaanxi 710119, China
[b] Department of Mathematical Sciences, University of Liverpool, L69 7ZB Liverpool, United Kingdom
[c] Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[d] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, L69 7ZB Liverpool, United Kingdom
[e] Institute of Ageing & Chronic Disease, University of Liverpool, L69 7ZB Liverpool, United Kingdom

ARTICLE INFO

*Keywords:*
Genomic coordinate
Web server
Epitranscriptome

ABSTRACT

The primary sequences of DNA, RNA and protein have been used as the dominant information source of existing machine learning tools, especially for contexts not fully explored by wet-experimental approaches. Since molecular markers are profoundly orchestrated in the living organisms, those markers that cannot be unambiguously recovered from the primary sequence often help to predict other biological events. To the best of our knowledge, there is no current tool to build and deploy machine learning models that consider genomic evidence. We therefore developed the WHISTLE server, the first machine learning platform based on genomic coordinates. It features convenient covariate extraction and model web deployment with 46 distinct genomic features integrated along with the conventional sequence features. We showed that, when predicting $m^6A$ sites from SRAMP project, the model integrating genomic features substantially outperformed those based on only sequence features. The WHISTLE server should be a useful tool for studying biological attributes specifically associated with genomic coordinates, and is freely accessible at: www.xjtlu.edu.cn/biologicalsciences/whi2.

## 1. Introduction

The primary sequences of DNA, RNA and protein convey the most fundamental information of the biomolecules, and have been used as the primary information source for machine learning tools in biosciences. To date, a large number of sequence-based methods have been developed to address various life science challenges such as the prediction of biological functions [1] and structures [2–4]. Meanwhile, many tools have been developed, such as bioSeq-Analysis [5], PyFeat [6] and PseKRAAC [7], to facilitate sequence-based feature extraction and machine learning modelling. Together these efforts have achieved great success, especially in obtaining insights into biological contexts that could not be adequately explored through wet-experimental approaches.

RNA modifications increase the structural and functional diversity of RNA molecules [8] and regulate every stage of RNA life [9–12]. Important roles of RNA modifications have been revealed in various diseases [13], cancers [14] and during viral infection [15]. Precise identification of RNA modification sites is thus of crucial importance for understanding the regulatory mechanisms and functionality of various RNAs. To date, a large number of computational approaches have been developed for *in silico* prediction of RNA modification sites from the primary RNA sequences, including: the iRNA series [16–24], SRAMP [25], DeepPromise [26], WHISTLE [27], RNAm5CPred [28], Gene2vec [29], PEA [30], BERMP [31] and PPUS [32]. As reviewed recently [26,33–35], these works have greatly advanced our understanding of the localization of multiple RNA modifications under various biological contexts in different organisms.

Due to limitations in available computational resources and in the learning capability of the machine learning models themselves, the primary sequence itself cannot provide all the information needed for machine learning prediction. In many cases only a fraction of the primary sequence rather than its entirety is used for prediction tasks. Substantial amounts of information are therefore lost during the model

* Corresponding authors at: Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (Z. Wei).
*E-mail addresses:* xjlei@snnu.edu.cn (X. Lei), zhen.wei01@xjtlu.edu.cn (Z. Wei).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.ymeth.2021.07.003
Received 8 June 2021; Received in revised form 28 June 2021; Accepted 5 July 2021
Available online 7 July 2021

selection process, in which the optimal input sequence length is determined. Although the sequences discarded from the analysis could in theory contain useful information as well, that information could not be explicitly extracted with the implemented machine learning models and thus cannot be effectively used. For example, in the problem of modification site prediction [36], many machine learning algorithms consider only 20–50 bp of DNA (or RNA) sequences, with the distant sequences discarded.

Importantly, distinct biological events are profoundly orchestrated in the living organisms, making them powerful predictors of each other. Although it is often expected that biological events should ultimately be encoded in the primary sequences, in practice, many of them could not be explicitly recovered from sequence-based analysis alone, and incorporating additional biological evidence can often boost the accuracy of prediction tools. However, as biological evidence (such as transcriptome annotation) is usually indexed with genomic coordinates, associating biological evidence (such as whether a given sequence lies within a 5′UTR or a CDS) to arbitrary biological sequences is often non-trivial and time consuming due to the involvement of genome aligners, and can be complicated by ambiguous multi-mapping scenarios.

We previously developed WHISTLE, a high-accuracy prediction framework for N6-methyladenosine ($m^6A$) RNA methylation site prediction [27]. Although WHISTLE considers only 41 bp of RNA sequence, by incorporating 35 additional genomic features its performance is among the best of $m^6A$ predictors, and is comparable to the most recent deep learning methods that require thousands of nucleotide of sequences as the input and were equipped with advanced encoding schemes [26]. Furthermore, we have shown that the WHISTLE framework can be successfully migrated to other prediction problems, including other RNA modifications (7-MethylGuanine, Pseudouridylation and N1-methyladenosine) [37–39], in non-human organism (mouse) [40], and for predicting marks located on lncRNAs and introns [41,42]. In all these studies, we demonstrated that the prediction performance achieved from genomic features alone is already comparable to sequence-based models; and that models combining both sequence and genomic features consistently yielded high-accuracy prediction results that are substantially better than those based on sequence information only. It is evident that additional biological evidence can be a valuable complement to sequence information in various prediction tasks; however, to the best of our knowledge, none of the existing sequence-based feature extraction tools try to recover higher-level biological evidences (such as transcriptome annotation, miRNA binding) from the input of primary sequences. Bioinformatics tools that enable extraction of biological evidence should greatly facilitate machine learning projects in biosciences.

By extending our previous work, we present here WHISTLE Server, a high-accuracy genomic coordinate-based machine learning platform for unleashing predictive power beyond the primary nucleic acid sequences. The inputs of WHISTLE server are genomic coordinates rather than the primary sequences so as to facilitate downstream genomic evidence extraction. It features convenient online (or offline) extraction of 46 distinct genomic features along with the conventional sequence features for both human and mouse. Importantly, the platform supports straightforward prediction model construction and online deployment for private or public usage. As case studies, we showed additionally that, when predicting $m^6A$ sites from SRAMP, the model integrating additional genomic features substantially outperformed those based on only sequence features. When only using genomic features, its performance is far better than using sequence features. Furthermore, it is possible to build and deploy a prediction model as a web app with just a few clicks with our server. Please refer to Fig. 1 for the overall design of the WHISTLE Server.

## 2. Material and methods

Rather than analyzing the biological entities in the sequence space only, we also consider the genome space along with various biological annotations and datasets mapped to it for additional information. Compared with the widely adopted sequence-based systems, our framework has clear advantages when dealing with entities whose sequences can be mapped to one or multiple genomic coordinates by providing additional genomic features extracted from the biological space; however, this approach will fail for entities that cannot be mapped to the genome, e.g., a piece of virus DNA, or when the biological

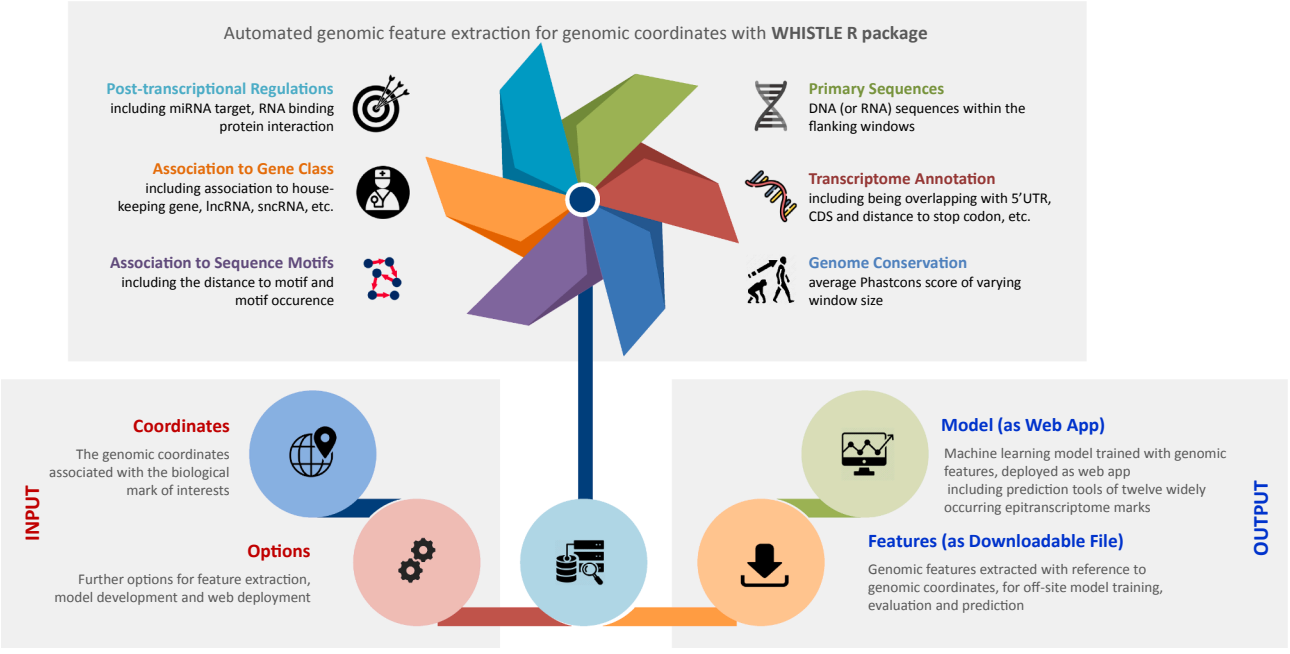


**Fig. 1.** The overall design of WHISTLE Server. WHISTLE Server supports convenient genomic feature extraction and prediction model development from the input of genomic coordinates of interests. A total of 46 genomic features can be extracted for human and mouse, concerning transcriptome annotation, genome conservation, post-transcriptional regulation, motif, gene class as well as the conventional sequence features. User can develop prediction models and conveniently deploy them online as web app for private or public usage.

space has not been annotated, e.g., for a less studied non-model organism. Luckily, as more and more experimental datasets and biological knowledge are accumulated, genomic features in general are likely to be more and more complete and effective. The WHISTLE Server supports two core functions: (1) batch genomic feature extraction for a set of genomic coordinates and (2) convenient high-accuracy prediction model construction and web deployment (as web app) for private or public usage.

We previously developed WHISTLE [27], a high-accuracy predictor of $m^6A$ RNA methylation sties. Although the WHISTLE framework is based on rather short input sequences (around 41 bp) and classic machine learning algorithm (random forest), by taking advantage of the additional genomic features extracted with respect to genomic coordinates, it achieved a state-of-the-art performance comparable to the most advanced deep-learning algorithms [26], which usually require long input sequence (typically 1 kb of nucleotides) and carefully tuning of the hyper-parameters. The parameter tuning of deep learning models often make them less convenient to migrate to address new problems. In contrast, migration of the WHISTLE framework is extremely convenient and straightforward, as shown in our previous work for 7-methylguanine [37], Pseudouridylation [38] and N1-methyladenosine [39]. To demonstrate the effectiveness of our approach, we showed with the following case study that the predictors constructed by the WHISTLE Server (leveraging genomic features) achieved high-accuracy prediction performance on $m^6A$ modification sites from SRAMP project.

### 2.1. Datasets and model

The development of an RNA modification site prediction model requires transcriptome-wide profiling data at base-resolution for training and testing purposes. The same number of negative sites as positive sites were randomly selected, either from the same nucleotide on the same transcript (full transcript mode) or the same nucleotide on the same mature RNA (mature mRNA mode). The mature RNA mode was considered here to eliminate potential bias towards mature transcripts, which can be induced by the poly-A selection step performed when preparing the RNA library in most high-throughput profiling approaches. Note that the negative data generation and motif-based site selection can be performed automatically by the WHISTLE Server with options specified from the user.

### 2.2. Genomic feature extraction

The WHISTLE Server supports the online extraction of 46 genomic features in batch for genomic coordinates of human and mouse. These genomic features cover various biological evidence, including transcriptome annotation, genome conservation, post-transcriptional regulation, association to gene class, association to sequence motif and the conventional primary sequences. The genomic features can be broadly grouped into 3 categories: 1) Topological features related to transcript landmarkers. 2) Range based RNA annotations that are RBP and microRNA binding sites which are previously reported to relate to $m^6A$ turn-over. 3) The sequence derived features such as sequence contents, sequence conservation, and sequence motifs. The topological features are derived from Bioconductor package GenomicFeatures. The range based RNA annotations are generated by overlapping the modification sites with the binding regions. The sequence features are calculated using the Bioconductor package BSgenome. The additional genomic features can be derived from a more flexible combination of existing topological features, such as the overlapping between the exons containing RBP regions. Additionally, more transcript annotations can be added from other databases, such as the transcript expression patterns & functional annotations. Many of those cannot be easily retrieved from the immediate primary sequences (see Supplementary Table S1 for details). Besides, we also provided the WHISTLE R package, which can be installed on a local computer for large-scale automated genomic feature extraction under the R environment. To the best our knowledge, there are no existing tools that support batch genomic coordinate-based feature extraction. As these genomic features can carry predictive power beyond the primary sequences, it is expected that they may be used to construct more advanced and more accurate prediction models offsite which would out-perform conventional sequence-based prediction models.

### 2.3. Prediction model construction and web deployment

In addition to genomic feature extraction, the WHISTLE Server also supports convenient construction and web deployment of high-accuracy predictors from a set of genomic coordinates that are associated with the attribute of interests, e.g., genomic coordinates of RNA modifications (RNA methylation sites). To do so, a few options need to be specified by the user to customize the related procedures, i.e., negative data generation, genomic features extraction, classification algorithm selection and the options related to web app deployment. When it is known that the attribute of interests is restricted to a specific sequence pattern, e.g., the DRACH motif of N6-methyladenosine RNA methylation, coordinates that do not comply with the sequence motif can optionally be discarded from the analysis. WHISTLE supported five classification algorithms for model training, random forest (RF), support vector machine (SVM), K-nearest neighbor (KNN), logistic regression (LR) and eXtreme Gradient Boosting (XGBoost). RF is a popular machine learning algorithm used to predict $m^6A$ RNA methylation, which was applied in SRAMP to predict mammalian $m^6A$ sites. SVM is another algorithm applied in computational biology, based on which the methods of MethyRNA [43] and RAM-ESVM [44] were developed to predict RNA methylation sites. KNN is one of the most powerful methods in the data mining classification technology, and LR is a method with a simple algorithm and a high performance. XGBoost is frequently used in competitions and industry, and can be effectively applied to the tasks of classification, regression and ranking, it was used in M6AMRFS [45] to predict $m^6A$ sites in multiple species based on the sequence features.

Finally, a predictor based on both genomic and sequence features will be generated and deployed online, which can predict the attribute associated with the input genomic coordinates. To the best our knowledge, there are no other such tools available for constructing predictors directly from genomic coordinates. The prediction performance of the web app is provided along with the web app page, including the sensitivity (Sn), specificity (Sp), accuracy (ACC), Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUC), all obtained from a 5-fold cross validation using the complete training data. In the process of constructing the model, WHISTLE used sequence features, genomic features and combined features to train the model respectively, and then takes the model with the best results of 5-fold cross validation as the final training model. Additionally, the users can specify whether the deployed web app should be publicly available or for private usage only. For a private web app, a link will be sent to only the designated email address; while all the public web apps will be displayed on the WHISTLE Server website, available for all to use.

### 2.4. Web interface implementation and programming environment

Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP) were applied in the construction of web interfaces. Genomic features for human and mouse were extracted with our newly developed WHISTLE package using the input genomic coordinates. Model construction and performance analysis were performed under R environment with customized scripts.

## 3. Results

To show the advantages brought by the additional genomic features,

we used the default setting of the WHISTLE server (46 genomic features, 21 bp of sequences with chemical property and nucleotide frequency encoding, random forest classifier), and constructed a prediction model for the datasets from SRAMP. Because our previous work, such as intronic $m^6A$ sites prediction, lncRNA methylation sites prediction and $m^1A$ prediction, all used RF classifier for prediction, so the default classifier here is RF. We then compared the performance of the predictors based on different feature sets (sequence features alone, genomic features alone, or both sequence and genomic features) in a 5-fold cross validation. In order to measure the prediction effect of the model, we used the measurements of sensitivity (Sn), specificity (Sp), accuracy (ACC), and Matthews correlation coefficient (MCC) to show the results of the model. In addition, we calculated the areas under the curves (as called "AUC") to evaluate the prediction performance. AUC was used as the main metrics for its non-parametric nature. The performance was obtained with 5-fold cross validation using the complete datasets, which were provided at WHISTLE Server website.

As shown in Table 1, under the mature RNA mode, genomic features alone are more powerful than sequence features (AUC = 0.802), which is superior to the prediction performance of SRAMP (AUC = 0.797). Importantly, the prediction performance by genomic features was better than combining both sequence and genomic features (AUC = 0.7865), demonstrating that the genomic features provided additional information given the sequence-based variables and played a more important role in the prediction.

The genomic features are even more evident under the full transcript mode. As shown in Table 2, the model based on genomic features (AUC = 0.9908) are substantially more accurate than sequence features (AUC = 0.8034).

## 4. Discussion

We present here WHISTLE Server, the first genomic coordinate-based machine learning platform to enhance prediction accuracy by integrating sequence-based features and higher-level annotations. Different from sequence-based platform such as bioSeq-Analysis [5], the WHISTLE server integrates additional biological evidence (such as transcriptome annotation, microRNA binding) with respect to genomic intervals, enabling higher prediction accuracy compared to purely sequence-based approaches. Importantly, it supports convenient prediction model construction and web deployment for private or public usage with just a few clicks. Together, it should make a useful resource to study biological attributes associated with specific genomic coordinates.

Future pathways to improve the WHISTLE Server could include: (1) Support for more species and more genome assemblies. Even though our computational framework is likely to be effective on most well annotated organisms, it currently supports only human (assembly hg19) and mouse (mm10). (2) Support for more feature encoding schemes and more machine learning algorithms. WHISTLE Server currently supports 5 classification algorithms: random forest (RF), support vector machine (SVM), K-nearest neighbor (KNN), logistic regression (LR) and eXtreme Gradient Boosting (XGBoost). (3) Coverage of more genomic features that may contribute to the prediction tasks. (4) Support for more intelligent ways of combining the sequence and genomic information for various prediction tasks, such as with multimodal techniques as used in Bichrom [46] and DeepRiPe [47].

Author contributions

BS and LL initialized the project; XL and ZW designed the research plan; ZW constructed the genomic features considered in site prediction; LL performed the site prediction; BS and LL built the website, LL drafted the manuscript. All authors read, critically revised and approved the final manuscript.

**Table 1**
Performance based on different feature sets under mature RNA mode.

| Method | Features | Sn | Sp | ACC | MCC | AUC |
|---|---|---|---|---|---|---|
| WHISTLE | Combined | 0.7722 | 0.6971 | 0.7347 | 0.4707 | 0.7865 |
| | Sequence | 0.7668 | 0.7028 | 0.7348 | 0.4706 | 0.6632 |
| | Genomic | 0.6481 | 0.6536 | 0.6508 | 0.3017 | 0.802 |

**Table 2**
Performance based on different feature sets under full transcript mode.

| Method | Features | Sn | Sp | ACC | MCC | AUC |
|---|---|---|---|---|---|---|
| WHISTLE | Combined | 0.9763 | 0.9303 | 0.9533 | 0.9076 | 0.989 |
| | Sequence | 0.974 | 0.9439 | 0.959 | 0.9184 | 0.8034 |
| | Genomic | 0.7283 | 0.7296 | 0.7289 | 0.4579 | 0.9908 |

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work has been supported by National Natural Science Foundation of China [61902230 to L.L., 61972451 to X.L.]; China Postdoctoral Science Foundation [2018M640949 to L.L.]; Fundamental Research Funds for the Central Universities [GK202103091 to L.L., GK201901010 to X.L.].

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2021.07.003.

## References

[1] The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong, Nucleic Acids Res. 47 (D1) (2018) D330–D338.
[2] A.W. Senior, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (7792) (2020) 706–710.
[3] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics 16 (4) (2000) 404–405.
[4] J.S. Reuter, D.H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis, BMC Bioinf. 11 (1) (2010) 129.
[5] B. Liu, X. Gao, 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 Res. 47 (20) (2019).
[6] K.D. Meyer, et al., 5' UTR m(6)A promotes cap-independent translation, Cell 163 (4) (2015) 999–1010.
[7] Y. Zuo, et al., PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition, Bioinformatics 33 (1) (2016) 122–124.
[8] H. Grosjean, Fine-Tuning of RNA Functions by Modification and Editing, Springer, 2005.
[9] S. Zaccara, R.J. Ries, S.R. Jaffrey, Reading, writing and erasing mRNA methylation, Nat. Rev. Mol. Cell Biol. 20 (10) (2019) 608–624.
[10] H.-C. Duan, Y. Wang, G. Jia, Dynamic and reversible RNA N6-methyladenosine methylation, Wiley Interdiscip. Rev.: RNA (2019).
[11] S. Delaunay, M. Frye, RNA modifications regulating cell fate in cancer, Nat. Cell Biol. 21 (5) (2019) 552–559.
[12] I.A. Roundtree, et al., Dynamic RNA modifications in gene expression regulation, Cell 169 (7) (2017) 1187–1200.
[13] E. Destefanis, et al., A mark of disease: how mRNA modifications shape genetic and acquired pathologies, RNA 27 (4) (2020) rna.077271.120.
[14] I. Barbieri, T. Kouzarides, Role of RNA modifications in cancer, Nat. Rev. Cancer 20 (6) (2020) 303–322.
[15] K. Tsai, B.R. Cullen, Epigenetic and epitranscriptomic regulation of viral replication, Nat. Rev. Microbiol. (2020) 1–12.
[16] W.-R. Qiu, et al., iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier, Med. Chem. 13 (8) (2017) 734–743.
[17] H. Yang, et al., iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens, J. Comput. Biol. 25 (11) (2018) 1266–1277.

[18] W. Chen, et al., iRNA(m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition, Anal. Biochem. 561–562 (2018). S0003269718307632.
[19] W. Chen, et al., iRNA-Methyl: Identifying N 6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.
[20] W.-R. Qiu, et al., iRNAm 5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (25) (2017) 41178–41188.
[21] W. Chen, et al., iRNA-m2G: identifying N2-methylguanosine sites based on sequence derived information, Mol. Ther. Nucleic Acids 18 (6) (2019) 253–258.
[22] W. Chen, et al., iRNA-m7G: identifying N7-methylguanosine sites by fusing multiple features, Mol. Ther. Nucleic Acids 18 (6) (2019) 269–274.
[23] M. Tahir, H. Tayara, K.T. Chong, iRNA-PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components, J. Theor. Biol. 465 (2019) 1–6.
[24] W. Chen, et al., iRNA-PseU: Identifying RNA pseudouridine sites, Mol. Ther. Nucleic Acids 5 (2016).
[25] Y. Zhou, et al., SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features, Nucleic Acids Res. (2016).
[26] Z. Chen, et al., Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences, Briefings Bioinf. 21 (5) (2019) 1676–1696.
[27] K. Chen, et al., WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach, Nucleic Acids Res. 7 (2019) 7.
[28] T. Fang, et al., RNAm 5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition, Mol. Ther. Nucleic Acids 18 (6) (2019) 739–747.
[29] Q. Zou Sr, et al., Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA, RNA 25 (2) (2018) rna.069112.118.
[30] J. Zhai, et al., PEA: an integrated R toolkit for plant epitranscriptome analysis, Bioinformatics 34 (21) (2018) 3747–3749.
[31] Y. Huang, et al., BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach, Int. J. Biol. Sci. 14 (12) (2018) 1669–1677.
[32] Y.-H. Li, G. Zhang, Q. Cui, PPUS: a web server to predict PUS-specific pseudouridine sites, Bioinformatics 20 (2015) 3362–3364.
[33] X. Zhu, et al., A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae, Brief. Funct. Genomics 18 (6) (2019) elz018.
[34] H. Lv, et al., Evaluation of different computational methods on 5-methylcytosine sites identification, Briefings Bioinf. 21 (3) (2019) bbz048.
[35] X. Chen, et al., RNA methylation and diseases: experimental results, databases, Web servers and computational models, Brief Bioinform. (2017) p. bbx142-bbx142.
[36] C. Ao, L. Yu, Q. Zou, Prediction of bio-sequence modifications and the associations with diseases, Brief. Funct. Genomics 21 (1) (2021) 1–18.
[37] B. Song, et al., m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human, Bioinformatics 36 (11) (2020) 3528–3536.
[38] B. Song, et al., PIANO: a web server for pseudouridine-site (Ψ) identification and functional annotation, Front. Genet. 11 (88) (2020).
[39] L. Lian, et al., ISGm1A: Integration of sequence features and genomic features to improve the prediction of human m1A RNA methylation sites. IEEE Access, 2020: p. 1-1.
[40] B. Song, et al., PSI-MOUSE: Predicting mouse pseudouridine sites from sequence and genome-derived features. Evolut. Bioinf., 2020. 16: p. 1176934320925752.
[41] L. Liu, et al., LITHOPHONE: improving lncRNA methylation site prediction using an ensemble predictor, Front. Genet. 11 (545) (2020).
[42] L. Liu, et al., WITMSG: large-scale prediction of human intronic m6A RNA methylation sites from sequence and genomic features, Curr. Genomics 21 (1) (2020) 67–76.
[43] W. Chen, H. Tang, H. Lin, MethyRNA: a web-server for identification of N(6)-methyladenosine sites, J. Biomol. Struct. Dyn. 35 (3) (2016) 683–687.
[44] C. Wei, P. Xing, Z. Quan, Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines, Sci. Rep. 7 (2017) 40242.
[45] X. Qiang, et al., M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species, Front. Genet. 9 (2018).
[46] D. Srivastava, et al., An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding, Genome Biol. 22 (1) (2021) 20.
[47] M. Safra, et al., The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution, Nature 551 (7679) (2017) 251–255.