

RESEARCH ARTICLE

Extremely-randomized-tree-based Prediction of N⁶-Methyladenosine Sites in *Saccharomyces cerevisiae*

Rajiv G. Govindaraj¹, Sathiyamoorthy Subramaniyam^{2,3} and Balachandran Manavalan^{4,*}

¹HotSpot Therapeutics, 50 Milk Street, 16th Floor, Boston, MA 02109, USA; ²Research and Development Center, In-silicogen Inc., Yongin-si 16954, Gyeonggi-do, Republic of Korea; ³Department of Biotechnology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu 641048, India; ⁴Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea

Abstract: Introduction: N⁶-methyladenosine (m6A) is one of the most common post-transcriptional modifications in RNA, which has been related to several biological processes. The accurate prediction of m6A sites from RNA sequences is one of the challenging tasks in computational biology. Several computational methods utilizing machine-learning algorithms have been proposed that accelerate *in silico* screening of m6A sites, thereby drastically reducing the experimental time and labor costs involved.

Methodology: In this study, we proposed a novel computational predictor termed ERT-m6Apred, for the accurate prediction of m6A sites. To identify the feature encodings with more discriminative capability, we applied a two-step feature selection technique on seven different feature encodings and identified the corresponding optimal feature set.

Results: Subsequently, performance comparison of the corresponding optimal feature set-based extremely randomized tree model revealed that Pseudo k-tuple composition encoding, which includes 14 physicochemical properties significantly outperformed other encodings. Moreover, ERT-m6Apred achieved an accuracy of 78.84% during cross-validation analysis, which is comparatively better than recently reported predictors.

Conclusion: In summary, ERT-m6Apred predicts *Saccharomyces cerevisiae* m6A sites with higher accuracy, thus facilitating biological hypothesis generation and experimental validations.

ARTICLE HISTORY

Received: November 01, 2019

Revised: December 28, 2019

Accepted: January 24, 2020

DOI:

10.2174/1389202921666200219125625

Keywords: Extremely randomized tree, feature optimization, N⁶-methyladenosine sites, cross-validation, RNA sequences, *Saccharomyces cerevisiae*.

1. INTRODUCTION

Post-transcriptional modifications in RNA are the variations that occur on a newly transcribed primary RNA transcript. To date, approximately 150 kinds of RNA modifications have been determined [1, 2]. The most abundant RNA modification is N⁶-methyladenosine (m6A), which is prevalent among viruses, plants, insects, mammals, and eukaryotes such as yeast [3-7]. m6A denotes the methylation at N-6 position of adenosine nucleotide catalyzed by a methyltransferase complex and this reaction is reversible by demethylases (ALKBH5 and FTO). m6A modification has been involved in a series of biological processes, such as mRNA exporting, nascent mRNA synthesis, splicing events, nuclear translation, and translocation [8-10]. Importantly, unusual modifications of m6A have been associated with several diseases, including prostate cancer, thyroid tumor, leukemia, etc. [11-13]. Therefore, accurate identification of m6A modification sites would be of great benefit for cell biologists to better understand the disease mechanism.

Several experimental approaches, including high performance liquid chromatography [14], next-generation sequencing technologies [15, 16], and two-dimensional thin layer chromatography [17] have been widely applied in the identification of m6A sites. Particularly, next-generation sequencing is not available for large-scale genomic sequences' m6A identification. Overall, these experimental approaches are time-consuming and cost-ineffective, when applied on large-scale genome analysis. Therefore, the development of an accurate and efficient computational method for m6A identification is necessary to complement experimental approaches.

Previous decade has witnessed tremendous growth in the development of various machine-learning (ML)-based methods to predict m6A sites from RNA sequences in different species, such as *Homo sapiens*, *Saccharomyces cerevisiae*, *Mus musculus*, and *Arabidopsis thaliana*. In this study, we focused on *S. cerevisiae* because it has been widely recognized as an attractive model organism. To date, 14 prediction models have been developed for this species to predict m6A sites. Chen *et al.*, [18] proposed the first predictor, where they constructed a reliable benchmark dataset of 1307 positive samples (m6A sites) and an equal number of negative samples (non-m6A sites) for *S. cerevisiae* based on the experimental data [19]. They developed a predictor called

*Address correspondence to this author at the Department of Physiology, Ajou University School of Medicine, 16499, Suwon, Republic of Korea; Tel/Fax: +82-31-219-4555/ +82-31-219-5049; E-mail: bala@ajou.ac.kr

“iRNA-Methyl” by employing SVM and pseudo-nucleotide composition features, that achieved an accuracy of 65.59%. Notably, iRNA-Methyl benchmark dataset acted as a base for the development of other methods, including m6Apred [20], pRNAm-PC [21], RNA-MethylPred [22], TargetM6A [23], M6A-HPCS [24], RAM-ESVM [25], RAM-NPPS [26], M6APred-EL [27], iRNA(m6A)-PseDNC [28], iMethyl-STNC [29], BERMP [30], M6AMRFS [31], and iRNA-Freq [32]. Researchers employed different approaches and feature encoding schemes to improve the predictive performance. A detailed description of the performance of each method and their approaches have been provided in the recent review [33]. Although substantial growth has been made, it still remains a challenging task to extract appropriately useful features to accurately differentiate m6A sites from non-m6A sites.

In the current study, we proposed a novel computational approach called ERT-m6Apred. Firstly, 14 physicochemical properties (PCPs) were incorporated into Pseudo k-tupler composition (PseKNC) and were considered as the features to differentiate m6A from non-m6A samples. To optimize the feature space, we combined F-score and sequential forward search (SFS) using extremely randomized tree (ERT) for enhancing the ability of the feature representation. Our benchmark result shows that the proposed method ERT-m6Apred can achieve improved performance when compared to the existing methods.

2. MATERIALS AND METHODS

2.1. Dataset

Chen *et al.*, [18] constructed a benchmark dataset containing 2614 sequences from *S. cerevisiae*. Of those, 1307 sequences are positive samples (m6A sites) and 1307 sequences are negative samples (non-m6A sites). Notably, both negative and positive samples are 51-base pair (bp) long with the adenosine in the center position. Furthermore, sequence similarity of this dataset is lower than 85%. This dataset has been widely used in the development of various prediction models [8, 21, 34]. For fair comparison with the existing methods, we also employed the same benchmark dataset for model development. The benchmark dataset can be downloaded from previous work: <http://server.malab.cn/m6APred-EL/>.

2.2. Feature Encoding Scheme

2.2.1. Pseudo k-tupler Composition (PseKNC)

In this work, we utilized Type-II PseKNC to represent RNA samples, which has been widely applied in previous methods [34-36]. This encoding can efficiently capture short-range and long-range information of RNA sequences. For a given sequence, it is denoted by:

$$R = [n_1, n_2, \dots, n_{4^k}, n_{4^k+1}, \dots, n_{4^k-\lambda}, n_{4^k+\lambda}]$$

where,

$$\begin{cases} \frac{f_u}{\sum_{a=1}^{4^k} f_a + \omega \sum_{b=1}^{\lambda} \tau_b}, (1 \leq u \leq 4) \\ \frac{\omega \theta_{u-4^k}}{\sum_{a=1}^{4^k} f_a + \omega \sum_{b=1}^{\lambda} \tau_b}, (4^k \leq u \leq 4^k + \lambda) \end{cases}$$

where λ denotes the counted rank of the correlations along an RNA sequence, f_u is the normalized frequency of the u^{th} k -tuple nucleotide in RNA segment, ω is the weight factor, and τ_b is defined as:

$$\tau_b = \frac{1}{L-b-1} \sum_{a=1}^{L-b-1} \tau(R_a R_{a+1}, R_{a+b} R_{a+b+1}), (b = 1, 2, \dots, \lambda; \lambda < L)$$

The correlation function $\tau(R_a R_{a+1}, R_{a+b} R_{a+b+1})$ is defined as:

$$\tau(R_a R_{a+1}, R_{a+b} R_{a+b+1}) = \frac{1}{\mu} \sum_{v=1}^{\mu} [P_v(R_a R_{a+1}) - P_v(R_{a+b} R_{a+b+1})]^2$$

where $\mu = 14$ physicochemical properties (PCPs), which includes 'Tilt', 'Roll', 'Rise', 'Shift', 'Slide', 'Twist', 'GC content', 'Adenine content', 'Uracil content', 'Stacking energy', 'Bending stiffness', 'Electron_interaction', 'Enthalpy', and 'Entropy.' We note that the above PCPs are based on DNA, which were modified accordingly to RNA sequence (replacing one base pair from Uracil to Thymine). L is the sequence length. $P_v(R_a R_{a+1})$ is the numerical value of the v^{th} physicochemical index of dinucleotide $R_a R_{a+1}$ at position a and $R_{a+b} R_{a+b+1}$ denotes the corresponding value of the dinucleotide $R_{a+b} R_{a+b+1}$ at position $a+b$. Based on our preliminary analysis of these three parameters (λ , ω , and k), we set $\lambda=1.0$, $\omega = 0.8$, and $k = 6$ to calculate the Type-II PseKNC that generates 4098-dimensional feature vector.

2.2.2. Feature Optimization

In general, the feature vectors with higher dimensions (>100) have noisy and irrelevant information [37, 38]. Therefore, it is necessary to exclude that information and improve classification accuracy, which is regarded as one of the key steps in ML-based model development. A two-step feature selection scheme was applied that has been widely used in computational biology [39-42]. Firstly, ranking the original features by F-score algorithm and generate a ranked feature list (from highest to the lowest ranked features) [43]. Secondly, two features were selected at each time from the ranked list, and added consecutively to ERT and developed their corresponding models by 10-fold cross-validation (CV). Eventually, the feature set equivalent to the model with the highest accuracy was considered as optimal features.

The F-score of the k^{th} feature is defined as:

$$F\text{-score}(k) = \frac{(\bar{x}_k^{(+)} - \bar{x}_k)^2 + (\bar{x}_k^{(-)} - \bar{x}_k)^2}{\frac{1}{m^+ - 1} \sum_{l=1}^{n^+} (\bar{x}_{k,l}^{(+)} - \bar{x}_k^{(+)})^2 + \frac{1}{m^- - 1} \sum_{l=1}^{n^-} (\bar{x}_{k,l}^{(-)} - \bar{x}_k^{(-)})^2}$$

where \bar{x}_k , $\bar{x}_k^{(-)}$, and $\bar{x}_k^{(+)}$ denote an average of the k^{th} feature in the combined (both negative and positive), negative, and positive datasets, respectively. m^- and m^+ represent the number of negative and positive samples, respectively. $\bar{x}_{k,l}^{(-)}$ and $\bar{x}_{k,l}^{(+)}$ represent the k^{th} feature of l^{th} negative instance and k^{th} feature of l^{th} positive instance, respectively.

2.2.3. Extremely Randomized Tree

Guertz *et al.* proposed ERT method in 2006 [44], which has been widely applied in various fields [45-51]. A brief description of how we implemented ERT in this study has been described previously [52, 53]. Grid search approach

was implemented to optimize three ERT parameters, which include the number of randomly selected features (*mtry*), number of trees (*ntree*), and minimum number of samples required to split an internal node (*nsplit*). The three parameters search ranges were: $3 \leq mtry \leq 12$ with a step size of 1, $40 \leq ntree \leq 1000$ with a step size of 20, and $4 \leq nsplit \leq 15$ with a step size of 1.

2.2.4. Cross-validation

We carried out 10-fold CV test to assess model performance. In 10-fold CV, the benchmark dataset was randomly separated into 10 subgroups with roughly equal size [54-64], with each subgroup containing the same number of m6A and non-m6A samples [65, 66]. One of the subgroups was considered as the validation set to assess the trained model and the remaining subgroups were used to train the model. This step was repeated 10 times by considering each one of the subgroups used at least once as a validation [67-69]. The performance of 10 validation sets was averaged and provided an assessment of the global performance.

2.2.5. Performance Assessment

The commonly used four sets of metrics in the fields of computational biology and bioinformatics [57, 70-77] were utilized to quantitatively evaluate the performance of the proposed method. These metrics included Matthews correlation coefficient (MCC), accuracy (ACC), sensitivity (SN), and specificity (SP), which were computed as follows:

$$\left\{ \begin{array}{l} SN = \frac{TP}{TP + FN} \\ SP = \frac{TN}{TN + FP} \\ ACC = \frac{TP + TN + FN + FP}{TP + TN + FN + FP} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \end{array} \right.$$

Where TP and TN respectively represent the number of m6A samples and the number of non-m6A samples correctly predicted. FP and FN represent the numbers of m6A or non-m6A samples wrongly predicted, respectively.

3. RESULTS AND DISCUSSION

3.1. Exploration of Different Feature Encodings in m6A Site Prediction

To identify the appropriate feature encoding in m6A site prediction seven feature encodings were explored, including PseKNC, k-mer (concatenation of mono-, di-, tri-, tetra-, and penta-nucleotide composition), binary profile (BPF), electron-ion interaction pseudo potential (PseEIIP), numerical representation of nucleotides (NUM), ring-function-hydrogen-chemical (RFHC) properties, and a combination of dinucleotide binary encoding, and local position-specific dinucleotide frequency (DPE_LPF). A detailed description of the other six feature encodings (except PseKNC) has been provided in previous studies [68, 78, 79] and we modified according to RNA sequence. Seven prediction models were developed and the performances are shown in Fig. (1). As shown in Fig. (1), we observed that four encodings (BPF, RFHC, DPE_LPF, NUM) achieved a similar performance, whose ACC in the range of 73.8-75.5%; (ii) the next two encodings (Kmer and PseKNC) achieved the same performance (ACC of 71.8%), which is significantly lower than the above four encodings, and EIIP achieved the worst performance. Finally, none of the featured encodings achieved significant performance over others. It should be noted that all these encodings (except NUM and EIIP) have greater than 100-dimensional feature vector. Hence, we applied SFS to see whether any encoding has an advantage over others in terms of performance.

3.2. Optimal Feature Selection of Each Encoding and their Performance Comparison

Fig. (2A) shows the ACC curves with gradual increment of features for seven encodings. The result shows that all

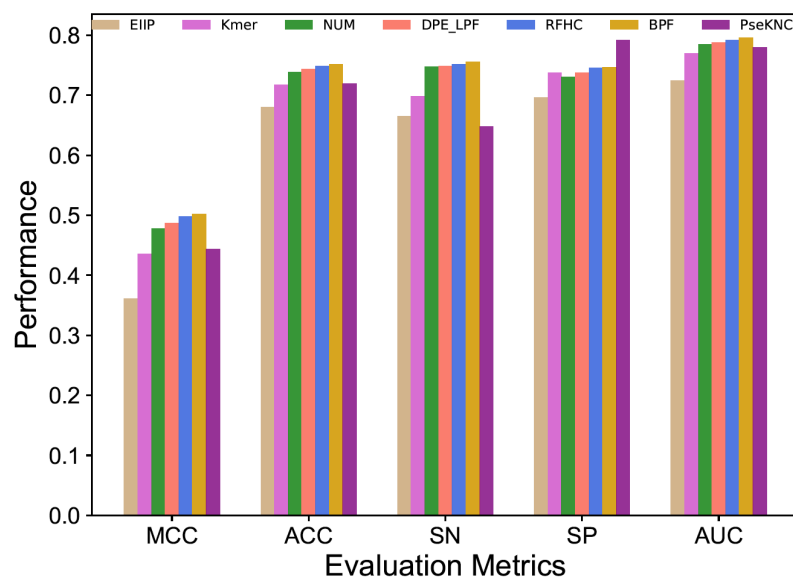


Fig. (1). Performance comparison of seven different feature encodings in terms of the Matthews Correlation Coefficient (MCC), Accuracy (ACC), Sensitivity (SN), Specificity (SP) and the Area Under the Curve (AUC). (A higher resolution / colour version of this figure is available in the electronic copy of the article).

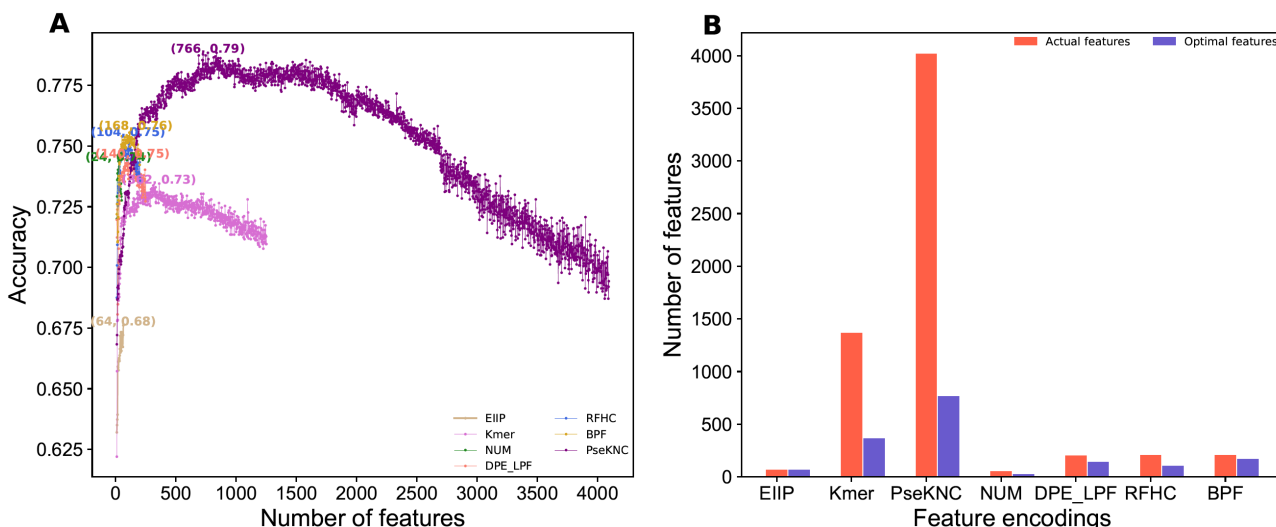


Fig. (2). (A) Sequential forward search for distinguishing between m6A and non-m6A sites for seven different encodings. (B) Comparison of the original and optimal features dimension. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

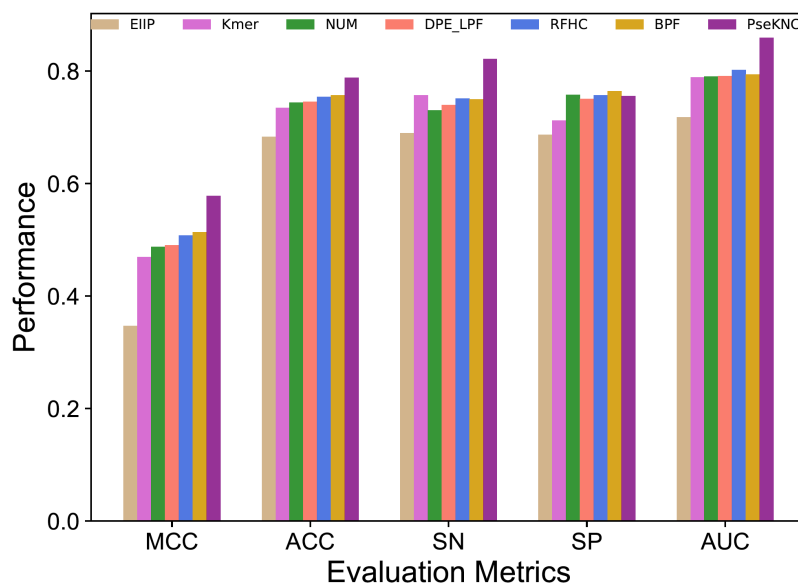


Fig. (3). Performance comparison of seven different feature encodings based on optimal feature set encodings in terms of Accuracy (ACC), the Matthews Correlation Coefficient (MCC), Sensitivity (SN), Specificity (SP) and the Area Under the Curve (AUC). (A higher resolution / colour version of this figure is available in the electronic copy of the article).

encodings follow the same pattern, where ACC curve gradually improved, reached its highest point and then deteriorated upon the increment of features. Remarkably, two-step feature selection procedure significantly improved the prediction performance in two encodings (Kmer and PseKNC), slight improvement in four encodings (BPF, DPE_LPF, RFHC and NUM), and no improvement in EIIP encoding when compared to their control (using all features). Next, we examined the highest ACC achieved by each encoding corresponding optimal features. Result shows that Kmer, PseKNC, BPF, RFHC, LPDF, and NUM encodings respectively considered 26.60%, 18.7%, 82.36% 50.98%, 70.0%, and 47.1% as the optimal features when compared to their original dimension (Fig. 2B). It appears that all features are equally important in case of EIIP encoding, because of its best performance with all features.

Next, we compared the optimal feature-based performances. Fig. (3) shows that PseKNC achieved the ACC of 78.84%, which is 3.17-10.74% higher than other encodings, thus indicating that PseKNC has more discriminative capability to identify m6A sites. Hence, we selected PseKNC-based model namely, ERT-m6Apred, as the final one.

3.3. Comparison with the Existing Methods

To see whether our proposed method is better than the existing predictors, we compared our method with the four recent methods, namely M6AMRFS, BERMP, iMethyl-STTNC and iRNA-Freq. Notably, all these methods were trained and validated on the same benchmark dataset as employed in this study, indicating a fair comparison. As shown in Table 1, our method ERT-m6Apred achieved an ACC of 78.84% and MCC of 0.578. Explicitly, ACC and MCC were

Table 1. Performance comparison of the proposed predictor with the existing methods.

Methods	MCC	ACC	SN	SP
M6A-ERT	0.578	0.788	0.822	0.755
M6AMRFS	0.485	0.743	0.752	0.733
BERMP	0.430	0.713	0.730	0.696
iMethyl-STTNC	0.380	0.698	0.703	0.682
iRNA-Freq	0.550	0.771	0.836	0.719

First column represents the method name. The second, third, fourth, and fifth respectively represents Matthews Correlation Coefficient (MCC), Accuracy (ACC), Sensitivity (SN) and Specificity (SP).

1.7-9.0% and 2.8-20% higher than the existing methods, thus indicating that our predictor ERT-m6Apred is more accurate and balanced than the existing predictors in m6A site prediction.

CONCLUSION

In this study, we reported a novel predictor for m6A site prediction from RNA sequence. To the best of our knowledge, this is the first instance where ERT and PseKNC were employed in m6A site prediction. Comparative analysis of CV results shows that the proposed predictor is better than existing predictors. Although our proposed method showed improved performance over other methods, it still has room for improvement. Recently, several novel computational approaches have been proposed in computational biology [68, 78, 80-85] to identify function from the sequence. Hence, developing a novel prediction model by utilizing such approaches may improve the prediction performance.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available in M6APred-EL database at <http://server.malab.cn/M6APred-EL/> [27].

FUNDING

This work has been supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science and Technology [2018R1D1A1B07049572].

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

- [1] Maden, B. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Progress in nucleic acid research and molecular biology*; Elsevier, **1990**, Vol. 39, pp. 241-303.
- [2] Wang, X.; Lu, Z.; Gomez, A.; Hon, G.C.; Yue, Y.; Han, D.; Fu, Y.; Parisien, M.; Dai, Q.; Jia, G.; Ren, B.; Pan, T.; He, C. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, **2014**, *505*(7481), 117-120. <http://dx.doi.org/10.1038/nature12730> PMID: 24284625
- [3] Yue, Y.; Liu, J.; He, C. RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev.*, **2015**, *29*(13), 1343-1355. <http://dx.doi.org/10.1101/gad.262766.115> PMID: 26159994
- [4] Wei, C.M.; Gershowitz, A.; Moss, B. 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry*, **1976**, *15*(2), 397-401. <http://dx.doi.org/10.1021/bi00647a024> PMID: 174715
- [5] Zhong, S.; Li, H.; Bodi, Z.; Button, J.; Vespa, L.; Herzog, M.; Fray, R.G. MTA is an Arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell*, **2008**, *20*(5), 1278-1288. <http://dx.doi.org/10.1105/tpc.108.058883> PMID: 18505803
- [6] Bodi, Z.; Button, J.D.; Grierson, D.; Fray, R.G. Yeast targets for mRNA methylation. *Nucleic Acids Res.*, **2010**, *38*(16), 5327-5335. <http://dx.doi.org/10.1093/nar/gkq266> PMID: 20421205
- [7] Clancy, M.J.; Shambaugh, M.E.; Timpte, C.S.; Bokar, J.A. Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res.*, **2002**, *30*(20), 4509-4518. <http://dx.doi.org/10.1093/nar/gkf573> PMID: 12384598
- [8] Liu, N.; Pan, T. N6-methyladenosine-encoded epitranscriptomics. *Nat. Struct. Mol. Biol.*, **2016**, *23*(2), 98-102. <http://dx.doi.org/10.1038/nsmb.3162> PMID: 26840897
- [9] Edupuganti, R.R.; Geiger, S.; Lindeboom, R.G.H.; Shi, H.; Hsu, P.J.; Lu, Z.; Wang, S-Y.; Baltissen, M.P.A.; Jansen, P.W.T.C.; Rossa, M.; Müller, M.; Stunnenberg, H.G.; He, C.; Carell, T.; Vermeulen, M. N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis. *Nat. Struct. Mol. Biol.*, **2017**, *24*(10), 870-878. <http://dx.doi.org/10.1038/nsmb.3462> PMID: 28869609
- [10] Slobodin, B.; Han, R.; Calderone, V.; Vrieland, J. A. O.; Loayza-Puch, F.; Elkon, R.; Agami, R. Transcription impacts the efficiency of mRNA translation via co-transcriptional N6-adenosine methylation. *Cell*, **2017**, *169*(2), 326-337 e12. <http://dx.doi.org/10.1016/j.cell.2017.03.031>
- [11] Akilzhanova, A.; Nurkina, Z.; Momynaliev, K.; Ramanculov, E.; Zhumadilov, Z.; Rakhypbekov, T.; Hayashida, N.; Nakashima, M.; Takamura, N. Genetic profile and determinants of homocysteine levels in Kazakhstan patients with breast cancer. *Anticancer Res.*, **2013**, *33*(9), 4049-4059. PMID: 24023349

- [12] Machiela, M.J.; Lindström, S.; Allen, N.E.; Haiman, C.A.; Albanes, D.; Barricarte, A.; Berndt, S.I.; Bueno-de-Mesquita, H.B.; Chanock, S.; Gaziano, J.M.; Gapstur, S.M.; Giovannucci, E.; Henderson, B.E.; Jacobs, E.J.; Kolonel, L.N.; Krogh, V.; Ma, J.; Stampfer, M.J.; Stevens, V.L.; Stram, D.O.; Tjønneland, A.; Travis, R.; Willett, W.C.; Hunter, D.J.; Le Marchand, L.; Kraft, P. Association of type 2 diabetes susceptibility variants with advanced prostate cancer risk in the Breast and Prostate Cancer Cohort Consortium. *Am. J. Epidemiol.*, **2012**, *176*(12), 1121-1129. <http://dx.doi.org/10.1093/aje/kws191> PMID: 23193118
- [13] Heiliger, K.-J.; Hess, J.; Vitagliano, D.; Salerno, P.; Braselmann, H.; Salvatore, G.; Ugolini, C.; Summerer, I.; Bogdanova, T.; Unger, K.; Thomas, G.; Santoro, M.; Zitzelsberger, H. Novel candidate genes of thyroid tumourigenesis identified in Trk-T1 transgenic mice. *Endocr. Relat. Cancer*, **2012**, *19*(3), 409-421. <http://dx.doi.org/10.1530/ERC-11-0387> PMID: 22454401
- [14] Zheng, G.; Dahl, J.A.; Niu, Y.; Fedorcsak, P.; Huang, C.-M.; Li, C.J.; Vågbo, C.B.; Shi, Y.; Wang, W.-L.; Song, S.-H.; Lu, Z.; Bosmans, R.P.; Dai, Q.; Hao, Y.J.; Yang, X.; Zhao, W.M.; Tong, W.M.; Wang, X.J.; Bogdan, F.; Furu, K.; Fu, Y.; Jia, G.; Zhao, X.; Liu, J.; Krokkan, H.E.; Klungland, A.; Yang, Y.G.; He, C. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell*, **2013**, *49*(1), 18-29. <http://dx.doi.org/10.1016/j.molcel.2012.10.015> PMID: 23177736
- [15] Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; Sorek, R.; Rechavi, G. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **2012**, *485*(7397), 201-206. <http://dx.doi.org/10.1038/nature11112> PMID: 22575960
- [16] Meyer, K.D.; Saletore, Y.; Zumbo, P.; Elemento, O.; Mason, C.E.; Jaffrey, S.R. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, **2012**, *149*(7), 1635-1646. <http://dx.doi.org/10.1016/j.cell.2012.05.003> PMID: 22608085
- [17] Keith, G. Mobilities of modified ribonucleotides on two-dimensional cellulose thin-layer chromatography. *Biochimie*, **1995**, *77*(1-2), 142-144. [http://dx.doi.org/10.1016/0300-9084\(96\)88118-1](http://dx.doi.org/10.1016/0300-9084(96)88118-1) PMID: 7599271
- [18] Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K.C. iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **2015**, *490*, 26-33. <http://dx.doi.org/10.1016/j.ab.2015.08.021> PMID: 26314792
- [19] Schwartz, S.; Agarwala, S.D.; Mumbach, M.R.; Jovanovic, M.; Mertins, P.; Shishkin, A.; Tabach, Y.; Mikkelsen, T.S.; Satija, R.; Ruvkun, G.; Carr, S.A.; Lander, E.S.; Fink, G.R.; Regev, A. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, **2013**, *155*(6), 1409-1421. <http://dx.doi.org/10.1016/j.cell.2013.10.047> PMID: 24269006
- [20] Chen, W.; Tran, H.; Liang, Z.; Lin, H.; Zhang, L. Identification and analysis of the N(6)-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci. Rep.*, **2015**, *5*, 13859. <http://dx.doi.org/10.1038/srep13859> PMID: 26343792
- [21] Liu, Z.; Xiao, X.; Yu, D.-J.; Jia, J.; Qiu, W.-R.; Chou, K.-C. pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal. Biochem.*, **2016**, *497*, 60-67. <http://dx.doi.org/10.1016/j.ab.2015.12.017> PMID: 26748145
- [22] Jia, C.-Z.; Zhang, J.-J.; Gu, W.-Z. RNA-MethylPred: a high-accuracy predictor to identify N6-methyladenosine in RNA. *Anal. Biochem.*, **2016**, *510*, 72-75. <http://dx.doi.org/10.1016/j.ab.2016.06.012> PMID: 27338301
- [23] Li, G.-Q.; Liu, Z.; Shen, H.-B.; Yu, D.-J. TargetM6A: Identifying N⁶-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. *IEEE Trans. Nanobioscience*, **2016**, *15*(7), 674-682. <http://dx.doi.org/10.1109/TNB.2016.2599115> PMID: 27552763
- [24] Zhang, M.; Sun, J.-W.; Liu, Z.; Ren, M.-W.; Shen, H.-B.; Yu, D.-J. Improving N(6)-methyladenosine site prediction with heuristic selection of nucleotide physical-chemical properties. *Anal. Biochem.*, **2016**, *508*, 104-113. <http://dx.doi.org/10.1016/j.ab.2016.06.001> PMID: 27293216
- [25] Chen, W.; Xing, P.; Zou, Q.; Detecting, N. Detecting N⁶-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep.*, **2017**, *7*, 40242. <http://dx.doi.org/10.1038/srep40242> PMID: 28079126
- [26] Xing, P.; Su, R.; Guo, F.; Wei, L.; Identifying, N. Identifying N⁶-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci. Rep.*, **2017**, *7*, 46757. <http://dx.doi.org/10.1038/srep46757> PMID: 28440291
- [27] Wei, L.; Chen, H.; Su, R. iM6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids*, **2018**, *12*, 635-644. <http://dx.doi.org/10.1016/j.omtn.2018.07.004> PMID: 30081234
- [28] Chen, W.; Ding, H.; Zhou, X.; Lin, H.; Chou, K.-C. iRNA(m6A)-PseDNC: identifying N⁶-methyladenosine sites using pseudo dinucleotide composition. *Anal. Biochem.*, **2018**, *561-562*, 59-65. <http://dx.doi.org/10.1016/j.ab.2018.09.002> PMID: 30201554
- [29] Akbar, S.; Hayat, M. iMethyl-STTNC: Identification of N⁶-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. *J. Theor. Biol.*, **2018**, *455*, 205-211. <http://dx.doi.org/10.1016/j.jtbi.2018.07.018> PMID: 30031793
- [30] Huang, Y.; He, N.; Chen, Y.; Chen, Z.; Li, L. BERMP: a cross-species classifier for predicting m⁶A sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci.*, **2018**, *14*(12), 1669-1677. <http://dx.doi.org/10.7150/ijbs.27819> PMID: 30416381
- [31] Qiang, X.; Chen, H.; Ye, X.; Su, R.; Wei, L. M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Front. Genet.*, **2018**, *9*, 495. <http://dx.doi.org/10.3389/fgene.2018.00495> PMID: 30410501
- [32] Zhuang, Y.Y.; Liu, H.J.; Song, X.; Ju, Y.; Peng, H. A linear regression predictor for identifying N⁶-methyladenosine sites using frequent gapped K-mer pattern. *Mol. Ther. Nucleic Acids*, **2019**, *18*, 673-680. <http://dx.doi.org/10.1016/j.omtn.2019.10.001> PMID: 31707204
- [33] Zhu, X.; He, J.; Zhao, S.; Tao, W.; Xiong, Y.; Bi, S. A comprehensive comparison and analysis of computational predictors for RNA N⁶-methyladenosine sites of *Saccharomyces cerevisiae*. *Brief. Funct. Genomics*, **2019**, *18*(6), 367-376. <http://dx.doi.org/10.1093/bfpg/ely018> PMID: 31609411
- [34] Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iDNA6mA-PseKNC: Identifying DNA N⁶-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*, **2019**, *111*(1), 96-102. <http://dx.doi.org/10.1016/j.ygeno.2018.01.005> PMID: 29360500
- [35] Dao, F.Y.; Lv, H.; Wang, F.; Feng, C.Q.; Ding, H.; Chen, W.; Lin, H. Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics*, **2019**, *35*(12), 2075-2083. <http://dx.doi.org/10.1093/bioinformatics/bty943> PMID: 30428009
- [36] Lou, C.; Zhao, J.; Shi, R.; Wang, Q.; Zhou, W.; Wang, Y.; Wang, G.; Huang, L.; Feng, X.; Zhou, F. sefOri: selecting the best-engineered sequence features to predict DNA replication origins. *Bioinformatics*, **2019**, *36*(1), 49-55. <http://doi.org/10.1093/bioinformatics/btz506>
- [37] Tan, J.X.; Li, S.H.; Zhang, Z.M.; Chen, C.X.; Chen, W.; Tang, H.; Lin, H. Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.*, **2019**, *16*(4), 2466-2480. <http://dx.doi.org/10.3934/mbe.2019123> PMID: 31137222
- [38] Yang, H.; Yang, W.; Dao, F.Y.; Lv, H.; Ding, H.; Chen, W.; Lin, H. A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief. Bioinform.*, **2019**, *bbz123*. <http://dx.doi.org/10.1093/bib/bbz123> PMID: 31633777
- [39] Xu, Z.-C.; Feng, P.-M.; Yang, H.; Qiu, W.-R.; Chen, W.; Lin, H. iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics*, **2019**, *35*(23), 4922-4929. <http://dx.doi.org/10.1093/bioinformatics/btz358> PMID: 31077296
- [40] Manavalan, B.; Lee, J. SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics*, **2017**, *33*(16), 2496-2503. <http://dx.doi.org/10.1093/bioinformatics/btx222> PMID: 28419290
- [41] Lai, H.Y.; Zhang, Z.Y.; Su, Z.D.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: A computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids*, **2019**, *17*, 337-346. <http://dx.doi.org/10.1016/j.omtn.2019.05.028> PMID: 31299595
- [42] Feng, C.Q.; Zhang, Z.Y.; Zhu, X.J.; Lin, Y.; Chen, W.; Tang, H.; Lin, H. iTerm-PseKNC: a sequence-based tool for predicting bacte-

- rial transcriptional terminators. *Bioinformatics*, **2019**, 35(9), 1469-1477.
<http://dx.doi.org/10.1093/bioinformatics/bty827> PMID: 30247625
- [43] Ding, H.; Li, D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids*, **2015**, 47(2), 329-333.
<http://dx.doi.org/10.1007/s00726-014-1862-4> PMID: 25385313
- [44] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.*, **2006**, 63(1), 3-42.
<http://dx.doi.org/10.1007/s10994-006-6226-1>
- [45] Paul, A.; Furmanchuk, A.; Liao, W.K.; Choudhary, A.; Agrawal, A. Property prediction of organic donor molecules for photovoltaic applications using extremely randomized trees. *Mol. Inform.*, **2019**, 38(11-12), e1900038.
<http://dx.doi.org/10.1002/minf.201900038> PMID: 31503423
- [46] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. AtbPpred: A robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J.*, **2019**, 17, 972-981.
<http://dx.doi.org/10.1016/j.csbj.2019.06.024> PMID: 31372196
- [47] Nattee, C.; Khamsemanan, N.; Lawtrakul, L.; Toochinda, P.; Han-nongbua, S. A novel prediction approach for antimalarial activities of Trimethoprim, Pyrimethamine, and Cycloguanil analogues using extremely randomized trees. *J. Mol. Graph. Model.*, **2017**, 71, 13-27.
<http://dx.doi.org/10.1016/j.jmgm.2016.09.010> PMID: 27835827
- [48] Soltaninejad, M.; Yang, G.; Lambrou, T.; Allinson, N.; Jones, T.L.; Barrick, T.R.; Howe, F.A.; Ye, X. Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI. *Int. J. CARS*, **2017**, 12(2), 183-203.
<http://dx.doi.org/10.1007/s11548-016-1483-3> PMID: 27651330
- [49] Xia, B.; Zhang, H.; Li, Q.; Li, T. PETS: a stable and accurate predictor of protein-protein interacting sites based on extremely-randomized trees. *IEEE Trans. Nanobioscience*, **2015**, 14(8), 882-893.
<http://dx.doi.org/10.1109/TNB.2015.2491303> PMID: 26529772
- [50] Scalzo, F.; Hamilton, R.; Asgari, S.; Kim, S.; Hu, X. Intracranial hypertension prediction using extremely randomized decision trees. *Med. Eng. Phys.*, **2012**, 34(8), 1058-1065.
<http://dx.doi.org/10.1016/j.medengphy.2011.11.010> PMID: 22401795
- [51] Marée, R.; Geurts, P.; Wehenkel, L. Random subwindows and extremely randomized trees for image classification in cell biology. *BMC Cell Biol.*, **2007**, 8(Suppl. 1), S2.
<http://dx.doi.org/10.1186/1471-2121-8-S1-S2> PMID: 17634092
- [52] Manavalan, B.; Subramaniyam, S.; Shin, T.H.; Kim, M.O.; Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.*, **2018**, 17(8), 2715-2726.
<http://dx.doi.org/10.1021/acs.jproteome.8b00148> PMID: 29893128
- [53] Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.*, **2018**, 9, 276.
<http://dx.doi.org/10.3389/fphar.2018.00276> PMID: 29636690
- [54] Charoenkwan, P.; Kanthawong, S.; Schaduagrath, N.; Yana, J.; Shoombuatong, W. PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method. *Cells*, **2020**, 9(2), E353.
<http://dx.doi.org/10.3390/cells9020353> PMID: 32028709
- [55] Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the *Rosaceae* genome. *Int. J. Biol. Macromol.*, **2019**, S0141-8130(19)38547-2.
<http://dx.doi.org/10.1016/j.ijbiomac.2019.12.009> PMID: 31805335
- [56] Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol Omics*, **2019**, 15(6), 451-458.
<http://dx.doi.org/10.1039/C9MO00098D> PMID: 31710075
- [57] Su, R.; Hu, J.; Zou, Q.; Manavalan, B.; Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.*, **2019**.
<http://dx.doi.org/10.1093/bib/bby124> PMID: 30649170
- [58] Wei, L.; Su, R.; Luan, S.; Liao, Z.; Manavalan, B.; Zou, Q.; Shi, X. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*, **2019**, 35(23), 4930-4937.
<http://dx.doi.org/10.1093/bioinformatics/btz408> PMID: 31099381
- [59] Shoombuatong, W.; Schaduagrath, N.; Pratiwi, R.; Nantasenam, C. THPeP: A machine learning-based approach for predicting tumor homing peptides. *Comput. Biol. Chem.*, **2019**, 80, 441-451.
<http://dx.doi.org/10.1016/j.compbiolchem.2019.05.008> PMID: 31151025
- [60] Schaduagrath, N.; Nantasenam, C.; Prachayasittikul, V.; Shoombuatong, W. ACPred: a computational tool for the prediction and analysis of anticancer peptides. *Molecules*, **2019**, 24(10), E1973.
<http://dx.doi.org/10.3390/molecules24101973> PMID: 31121946
- [61] Laengsri, V.; Shoombuatong, W.; Adirojananon, W.; Nantasenam, C.; Prachayasittikul, V.; Nuchnoi, P. ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC Med. Inform. Decis. Mak.*, **2019**, 19(1), 212.
<http://dx.doi.org/10.1186/s12911-019-0929-2> PMID: 31699079
- [62] Laengsri, V.; Nantasenam, C.; Schaduagrath, N.; Nuchnoi, P.; Prachayasittikul, V.; Shoombuatong, W. TargetAnti-Angio: a sequence-based tool for the prediction and analysis of anti-angiogenic peptides. *Int. J. Mol. Sci.*, **2019**, 20(12), E2950.
<http://dx.doi.org/10.3390/ijms20122950> PMID: 31212918
- [63] Charoenkwan, P.; Schaduagrath, N.; Nantasenam, C.; Piacham, T.; Shoombuatong, W. iQSP: a sequence-based tool for the prediction and analysis of quorum sensing peptides via Chou's 5-steps rule and informative physicochemical properties. *Int. J. Mol. Sci.*, **2019**, 21(1), E75.
<http://dx.doi.org/10.3390/ijms21010075> PMID: 31861928
- [64] Uchida, T.; Furukawa, M.; Kikawada, T.; Yamazaki, K.; Gohara, K. Intracellular trehalose via transporter TRET1 as a method to cryoprotect CHO-K1 cells. *Cryobiology*, **2017**, 77, 50-57.
<http://dx.doi.org/10.1016/j.cryobiol.2017.05.008> PMID: 28552273
- [65] Qiao, Y.; Xiong, Y.; Gao, H.; Zhu, X.; Chen, P. Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinformatics*, **2018**, 19(1), 14.
<http://dx.doi.org/10.1186/s12859-018-2009-5> PMID: 29334889
- [66] He, J.; Fang, T.; Zhang, Z.; Huang, B.; Zhu, X.; Xiong, Y.; Pse, U.I. PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics*, **2018**, 19(1), 306.
<http://dx.doi.org/10.1186/s12859-018-2321-0> PMID: 30157750
- [67] Cao, R.; Freitas, C.; Chan, L.; Sun, M.; Jiang, H.; Chen, Z. Pro-LanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules*, **2017**, 22(10), 1732.
<http://dx.doi.org/10.3390/molecules22101732> PMID: 29039790
- [68] Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids*, **2019**, 18, 131-141.
<http://dx.doi.org/10.1016/j.omtn.2019.08.011> PMID: 31542696
- [69] Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.*, **2018**, 16, 412-420.
<http://dx.doi.org/10.1016/j.csbj.2018.10.007> PMID: 30425802
- [70] Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D-C. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.*, **2019**, 20(8), 1964.
<http://dx.doi.org/10.3390/ijms20081964> PMID: 31013619
- [71] Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget*, **2017**, 9(2), 1944-1956.
 PMID: 29416743
- [72] Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.*, **2018**, 9, 476.
<http://dx.doi.org/10.3389/fmicb.2018.00476> PMID: 29616000
- [73] Xu, Q.; Xiong, Y.; Dai, H.; Kumari, K.M.; Xu, Q.; Ou, H.-Y.; Wei, D.-Q. PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.*, **2017**, 417, 1-7.
<http://dx.doi.org/10.1016/j.jtbi.2017.01.019> PMID: 28099868
- [74] Xiong, Y.; Wang, Q.; Yang, J.; Zhu, X.; Wei, D.Q. PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbi-*

- ol.*, **2018**, *9*, 2571.
<http://dx.doi.org/10.3389/fmicb.2018.02571> PMID: 30416498
- [75] Yang, W.; Zhu, X.J.; Huang, J.; Ding, H.; Lin, H. A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.*, **2019**, *14*, 234-240.
<http://dx.doi.org/10.2174/1574893613666181113131415>
- [76] Ding, H.; Yang, W.; Tang, H.; Feng, P.M.; Huang, J.; Chen, W.; Lin, H. PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. *Virol. Sin.*, **2016**, *31*(4), 350-352.
<http://dx.doi.org/10.1007/s12250-016-3740-6> PMID: 27151186
- [77] Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med. Res. Rev.*, **2020**.
<http://dx.doi.org/10.1002/med.21658> PMID: 31922268
- [78] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids*, **2019**, *16*, 733-744.
<http://dx.doi.org/10.1016/j.omtn.2019.04.019> PMID: 31146255
- [79] Wei, L.; Luan, S.; Nagai, L.A.E.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N⁴-methylcytosine sites in multiple species. *Bioinformatics*, **2019**, *35*(8), 1326-1333.
<http://dx.doi.org/10.1093/bioinformatics/bty824> PMID: 30239627
- [80] Wang, J.; Li, J.; Yang, B.; Xie, R.; Marquez-Lago, T.T.; Leier, A.; Hayashida, M.; Akutsu, T.; Zhang, Y.; Chou, K.C.; Selkrig, J.; Zhou, T.; Song, J.; Lithgow, T. Bastion3: a two-layer ensemble predictor of type III secreted effectors. *Bioinformatics*, **2019**, *35*(12), 2017-2028.
<http://dx.doi.org/10.1093/bioinformatics/bty914> PMID: 30388198
- [81] Zhang, Y.; Yu, S.; Xie, R.; Li, J.; Leier, A.; Marquez-Lago, T.T.; Akutsu, T.; Smith, A.I.; Ge, Z.; Wang, J. PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics*, **2019**, *1*, 9.
<http://dx.doi.org/10.1093/bioinformatics/btz629> PMID: 31393553
- [82] Li, F.; Chen, J.; Leier, A.; Marquez-Lago, T.; Liu, Q.; Wang, Y.; Revote, J.; Smith, A.I.; Akutsu, T.; Webb, G.I.; Kurgan, L.; Song, J. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics*, **2019**, btz721.
<http://dx.doi.org/10.1093/bioinformatics/btz721> PMID: 31566664
- [83] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. maHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics*, **2019**, *35*(16), 2757-2765.
PMID: 30590410
- [84] Yu, B.; Qiu, W.; Chen, C.; Ma, A.; Jiang, J.; Zhou, H.; Ma, Q. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics*, **2019**, btz734.
<http://dx.doi.org/10.1093/bioinformatics/btz734> PMID: 31603468
- [85] Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, **2018**, *34*(23), 4007-4016.
<http://dx.doi.org/10.1093/bioinformatics/bty451> PMID: 29868903