Enhancer-DSNet: A Supervisedly Prepared Enriched Sequence Representation for the Identification of Enhancers and their Strength

Muhammad Nabeel Asim[1], Muhammad Imran Malik[3], Andreas Dengel[1], and Sheraz Ahmed[1]

[1] German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany Muhammad_Nabeel.Asim@dfki.de
[2] University of Kaiserslautern (TU Kaiserslautern), 67663 Kaiserslautern, Germany
[3] National Center for Artificial Intelligence (NCAI), National University of Sciences and Technology, Islamabad, Pakistan

Abstract. Identification of enhancers and their strength prediction plays an important role in gene expression regulation and currently an active area of research. However, its identification specifically through experimental approaches is extremely time consuming and labor-intensive task. Several machine learning methodologies have been proposed to accurately discriminate enhancers from regulatory elements and to estimate their strength. Existing approaches utilise different statistical measures for feature encoding which mainly capture residue specific physicochemical properties upto certain extent but ignore semantic and positional information of residues. This paper presents "Enhancer-DSNet", a two-layer precisely deep neural network which makes use of a novel k-mer based sequence representation scheme prepared by fusing associations between k-mer positions and sequence type. Proposed Enhancer-DSNet methodology is evaluated on a publicly available benchmark dataset and independent test set. Experimental results over benchmark independent test set indicate that proposed Enhancer-DSNet methodology outshines the performance of most recent predictor by the figure of 2%, 1%, 2%, and 5% in terms of accuracy, specificity, sensitivity and matthews correlation coefficient for enhancer identification task and by the figure of 15%, 21%, and 39% in terms of accuracy, specificity, and matthews correlation coefficient for strong/weak enhancer prediction task.

Keywords: Enhancer Identification · Strong Enhancer · Weak Enhancers · Enhancer Classification · Deep Enhancer Predictor · Enhancer Strength Identification · Enriched K-mers

1 Introduction

Enhancers are functional cis elements which belong to diverse subgroups (e.g strong enhancer, weak enhancers, poised enhancers, and inactive enhancers), where each type of enhancer is associated with multifarious biological activities

[2]. Mainly, in gene expression regulation, enhancers play an indispensable role for the generation of proteins and RNA [3] and ensure very close relationship between biological processes [21]. Enhancers impact cell growth, cell differentiation, cell carcinogenesis, virus activity, and tissue specificity through enhancing genes transcription [21]. Enhancer may be located in separate chromosome or 20 kb far away from genes [4] as compared to promoters which are usually located around start transcriptional sites of genes. Building on these locational differences, identifying enhancers is widely considered far more challenging than promoters. Discriminating enhancers from regulatory elements, estimating their location and overall strength are few most promising tasks which can facilitate deeper comprehension of eukaryotic spatiotemporal gene regulation and evolution of diseases [4].

Initially, enhancers were discovered through typical experimental approaches [5], [6]. Former approach used to identify enhancers by utilizing their association with transcriptional factor [7], whereas, latter approach leveraged DNase-I hypersensitivity. While former approach under detected enhancers [8] as all enhancers are not occupied by transcription factors, latter approach over detected as it classified even DNA segments or non-enhancers as enhancers [4], [9]. Although subsequent methodologies of genome wide mapping of histone modifications [10], [11], [16] decently alleviated high false positive and false negative rate of initial experimental techniques for the discovery of promoters and enhancers. However, these approaches are rigorously expensive, time, and resource consuming. Due to these shortcomings and with the influx of high throughput biological data related to enhancers, demand of robust computational methodologies capable to differentiate enhancers from regulatory elements and estimate their strength got significantly rocketed.

Up to this date, several computational methodologies have been proposed to discriminate enhancers from non-enhancers in genome such as CSI-ANN [13], RFECS [16], EnhancerFinder [11], EnhancerDBN [18], and BiRen [19]. Proposed predictors differ in terms of feature encoding and classifier. For example, CS1-ANN [13] utilized data transformation approach for samples formulation and Artificial Neural Network (ANN) for classification. Likewise, EnhancerFinder [11] incorporates evolutionary conservation knowledge into sample formulation and a combination of several kernel learning approaches for classification. EnhancerDBN [18] makes use of deep belief network (DBN), RFEC [16] utilizes random forest classifier [20], whereas BiRen [19] leverages deep learning approaches to accelerate predictive performance. These approaches only capable to discriminate enhancers from regulatory elements in genome. Therefore, robust enhancer determinant and strength prediction approaches are still scarce. iEnhancer-2L [4] is the very first tool developed to discover enhancer along with their strength using solely sequence information and it has been extensively utilized for genome analysis. To further improve the performance at both layers, more computational methodologies have been developed afterward which have improved iEnhancer-2L [4] methodology further by using the combination of statistical measures to better represent physico-chemical properties such as

EnhancerPred [21], iEnhancer-PsedeKNC [4] iEnhancer-EL [24], Tan et al. Enhancer [25], and EnhancerPred2.0 [27]. Up to date, only one recently proposed approach namely "iEnhancer-5Step" [26] makes use of SVM classifier and unsupervisedly prepared neural k-mer embeddings to better capture local patterns for the task of enhancer determinant and strength prediction.

Nevertheless, still a lot of improvement in performance is required as these approaches produce confined performance especially in distinguishing strong enhancers from weak enhancers. To develop an optimal machine learning model for enhancer identification and strength prediction task, most crucial step is to encode biomedical sequence into fixed-size low dimensional vectors. In this context, few sequence encoding approaches including Local Descriptor, Conjoint Triad (CT), Auto Covariance (AC), and PSE-KNC [24] have been utilized where residual oriented physico-chemical properties are taken into account. But, the major downfalls for such manually curated feature vectors are, these approaches fail to take semantic information of residues into account (such as residues order) in sequences and also neglect noteworthy information from large number of unlabelled biomedical sequences that can assist the classifier to better identify class boundaries. To overcome these shortcomings upto certain extent, Le et al. [26] have recently employed neural word embeddings prepared in an unsupervised manner. Although unsupervised k-mer embeddings capture semantic information of k-mers, however they still lack to associate inherent k-mer relationships with sequence type keeping within low-dimensional vector space. To fully reap the benefits of neural word embeddings for creating an optimal representation of k-mers present in sequences, we present a novel k-mer based sequence representation scheme which prepares the sequence embeddings in a supervised manner where we fuse the alliance of k-mers with sequence type. To evaluate the effectiveness of presented enriched sequence representation, we present a two-layer classification methodology (Enhancer-DSNet) based on linear classifier and perform experimentation over a publicly available benchmark dataset and independent test set for the task of enhancer determinant and strength prediction task. We have obtained excellent predictive accuracy, outperformed various combinations of machine learning algorithms, commonly-used sequence encoding schemes, and unsupervisedly prepared k-mer embeddings with significant margins.

2 Materials And Methods

This section discusses proposed two-layer classification methodology "Enhancer-DSNet", benchmark dataset and independent test set used for experimentation, and evaluation measures.

3 Proposed Enhancer-DSNet Methodology

With the huge success of pre-trained neural word embeddings over diversified NLP tasks [23], biomedical researchers have extensively utilized distributed representations in different biomedical tasks [28]. These embeddings are usually

prepared in an unsupervised manner by training a shallow neural network on gigantic sequence corpora. Pre-trained neural k-mer embeddings are semantically meaningful low dimensional dense representation of k-mers present in the sequences. Although neural k-mer embeddings prepared in an unsupervised manner create proximal representation of highly similar k-mer groups in embedding space and have shown good performance in different biomedical tasks such as sequences structural similarity estimation [29], and transmembrane prediction [22]. However, these embeddings still lack to associate class information with distinct arrangements of nucleotides present in sequences, a phenomena that can significantly raise the classifier performance [28].

Considering relationships between distinct k-mers largely depend on k-mer size and sequence type, we have generated k-mer embeddings in a supervised manner. Unlike trivial neural k-mer embeddings, here, we improve k-mer representation by creating associations between k-mers positions and sequence type.

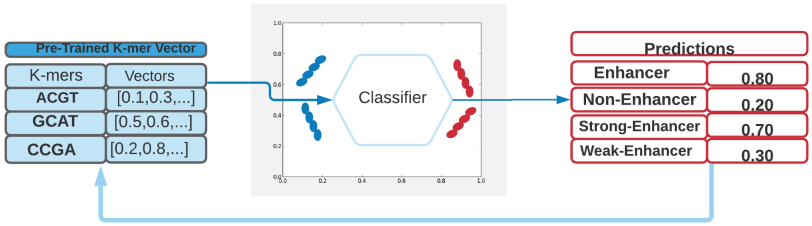


Fig. 1: Supervisedly Prepared Neural K-mer Embeddings

To generate sequence embeddings, k-mer embeddings are concatenated through summation. In this manner, we are accurately capturing semantic information and local patterns present in sequences. Also, we are computing sequences similarity correctly within low dimensional space revealing functional relationship, while making sure that computation relies on set of features pertinent to hand on problem. Architecture of proposed two-layer Enhancer-DSNet approach is illustrated in Figure 2. Where firstly, overlapped k-mers of each sequence is generated by sliding a window across the sequence with stride size of 1. Afterward, overlapped k-mers of sequences are passed to embedding layer, where 100 dimensional vectors are generated for each overlapped k-mers. All k-mer vectors are then aggregated to generate 100 dimensional vector for the whole sequence. In order to avoid over fitting the model, a dropout layer with dropout rate of 0.5% is utilized. After dropout layer, softmax classifier is used to incorporate label information into the sequence vectors by updating model parameters. In this manner, we ensure that, on independent test set, model is able to extract meaningful patterns through which classifier will better discriminate the sequences at both layers.

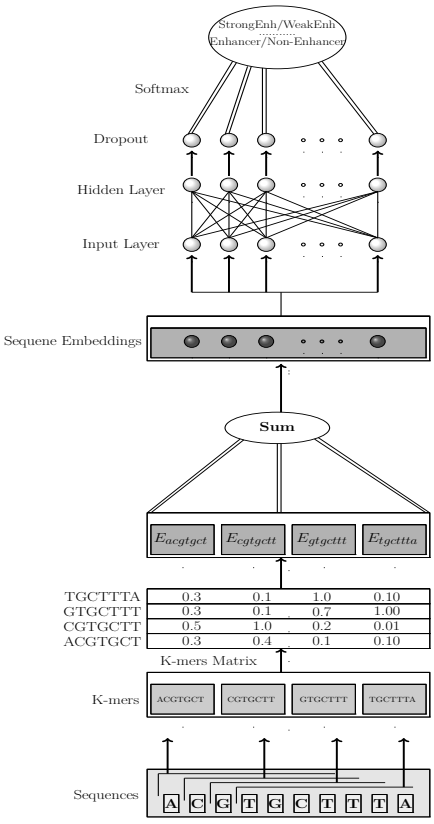


Fig. 2: Architecture of Proposed Two-Layer Classification Methodology "Enhancer-DSNet"

4 Benchmark Dataset

To evaluate the integrity of proposed Enhancer-DSNet approach, experimentation is performed on a publicly available benchmark dataset and independent test set [4]. These resources have been utilized in previous studies to evaluate enhancer determinant and strength prediction approaches [21, 4, 27, 24, 25]. Enhancer and non-Enhancer discrimination benchmark dataset has 2968 samples, out of which 1484 samples are enhancers and 1484 samples are non-enhancers. Out of 1484 enhancer samples, 742 samples are strong enhancers and remaining 742 samples are weak enhancers. While enhancer/non-enhance dataset is used to discriminate enhancers from non-enhancers, strong/weak enhancer subset formulated using enhancer samples is further used to estimate the strength of enhancers. Besides benchmark dataset, an independent test set is also publicly available which contain 400 samples, out of which 200 samples are enhancers and remaining 200 samples are non-enhancers. From 200 enhancer samples, 100 sam-

ples are strong enhancers and remaining 100 samples are weak enhancers. Just like benchmark dataset, enhancer/non-enhancer independent test set is used to for enhancer/non-enhancer prediction task, whereas strong/weak enhancer subset formulated using enhancer samples of independent test is used to estimate the strength of enhancers. Detailed formulation of benchmark and independent test set have been clearly elaborated in Liu et al. [24] work, hence there in no need to repeat here.

5 Evaluation Metrics

Following evaluation criteria of previous studies related to the classification of enhancer and other regulatory elements, and estimating the strength of enhancers [21, 4, 27, 24, 25], here we have used 4 different evaluation measures (sensitivity, specificity, accuracy, and matthews correlation coefficient) to perform a fair performance comparison of proposed approach with state-of-the-art approaches. As these measures are briefly described in previous studies [21, 4] so here we just give a short description. To provide intuitive understanding for readers, evaluation metrics along with mathematical expressions are briefly described below:

$$f(x) = \begin{cases} \text{Accuracy (ACC)} = 1 - (O_{-}^{+} + (O_{+}^{-})/(O^{+} + O^{-}) & 0 \leq \text{Acc} \leq 1 \\ \text{Specificity (SP)} = 1 - (O_{+}^{-}/O^{-}) & 0 \leq \text{SP} \leq 1 \\ \text{Sensitivity (SN)} = 1 - -(O_{-}^{+}/O^{+}) & 0 \leq \text{SN} \leq 1 \\ \text{MCC} = 1 - (O_{-}^{+}/O^{+} + O_{+}^{-}/O^{-})/\sqrt{(1+O_{+}^{-} - O_{-}^{+}/O^{+})(1+O_{-}^{+} - O_{+}^{-}/O^{-}} & -1 \leq \text{MCC} \leq 1 \end{cases} \tag{1}$$

Here, O^{+} infers total positive class observations investigated, O^{-} represents total negative class observations investigated. While, number of positive class observations predicted correctly and are negative class observations predicted correctly. Whereas, represent positive class observation incorrectly predicted as negative and are negative class observations mis-classified as positive.

6 Experimental Setup And Results

This section illustrates experimental details and briefly describes Results of proposed Enhancer-DSNet approach.

To generate sequence embeddings of benchmark dataset in a supervised manner, and to perform experimentation over over benchmark dataset and independent test set, we have used Pytorch API. To generate supervised sequence vectors, we have trained the newly developed skip-gram model for 30 epochs with 0.008 learning rate and adam optimizer. Experimentation for both enhancer identification and strength prediction tasks is performed using 7-mer enriched sequence vectors.

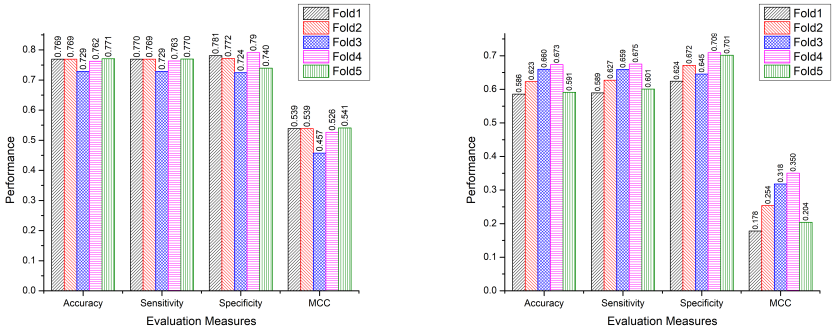
6.1 Results

Here, we briefly describe and compare the performance of proposed Enhancer-DSNet methodology with state-of-the-art Enhancer determinant and strength prediction approaches using cross validation and benchmark independent test set.

Cross-validation In order to better evaluate the performance of a classifier by eliminating biasness towards the split of dataset, most widely used re-sampling approach is called cross-validation. In k-fold cross validation, one can split a dataset into k number of groups, for example, 5-fold cross validation will segregate entire dataset into 5 groups where each group will be splitted into train, and test sets to to train and test the model. In this manner, each group of limited data samples take part in training and testing processes. Another similar unbiased performance estimator is jackknife test where training is performed over entire dataset except one observation of a dataset which is iteratively used to test the model. In comparison to cross-validation, jackknife test is quite expensive to compute especially for large datasets and it has also high variance as datasets used to estimate classifier performance are quite similar. Hence, k-fold cross validation is widely considered a better estimator of bias and variance as it is a well compromise among computational requirements and impartiality. Existing enhancer and non-enhancer discriminator and enhancer strength predictor approaches (EnhancerPred [21], iEnhancer-PsedeKNC [4] iEnhancer-EL [24], EnhancerPred2.0 [27]) utilized jackknife test to evaluate the performance of their models on a benchmark dataset. However, most recent Tan et al. predictor [25] performance is evaluated using 5 fold cross validation. Following Tan et al. [25] work, in our experimentation, we have also used 5-fold cross validation on a benchmark dataset. So here, using 5-fold cross validation, we perform performance comparison of Enhancer-DSNet with most recent Tan et al. predictor [25].

Figures 3a 3b illustrate the performance of Enhancer-DSNet across 5-folds on a benchmark dataset of enhancer/non-enhancer and strong/weak enhancer prediction task. To sum up, performance of Enhancer-DSNet remains consistent across 5-folds when evaluated in terms of 4 distinct evaluation metrics.

Table 1 reports the average of performance figures produced by 5-fold cross validation at layer 1 and 2 in terms of accuracy, specificity, sensitivity and matthews correlation coefficient (mcc). As is indicated by the Table 1, for enhancer/non-enhancer prediction task (layer-1), proposed Enhancer-DSNet outshines Tan et al. Enhancer [25] by the figure of 3% in terms of sensitivit, 2% in terms of accuracy and 2% in terms of matthews correlation coefficient. However, for strong/weak enhancer prediction task (layer-2), proposed Enhancer-DSNet outsperforms Tan et al. Enhancer [25] with a huge margin across 4 different evaluation metrics. Enhancer-DSNet significantly superior performance overshadows most recent Tan et al. Enhancer [25] performance by the figure of 17% in terms of sensitivity, 29% in terms of specificity, 4% in terms of accuracy, and 6% in terms of mcc.



(a) Enhancer/Non-Enhancer Prediction (b) Strong/Weak Enhancer Prediction

Fig. 3: Performance of Enhancer-DSNet Produced Over 5-Folds For Layer 1 and 2 In Terms Of Accuracy, Specificity, Sensitivity, and MCC

Classifiers	Sensitivity	Specificity	Accuracy	MCC
1st Layer (Enhancer/Non-Enhancer)				
Enhancer-DSNet	**0.76**	**0.76**	**0.76**	**0.52**
Tan et al. Enhancer [25]	0.73	0.76	0.74	0.50
2nd Layer (Strong Enhancer/Weak Enhancer)				
Enhancer-DSNet	**0.63**	**0.67**	**0.63**	**0.26**
Tan et al. Enhancer [25]	0.80	0.38	0.59	0.20

Table 1: Performance Comparison of Enhancer-DSNet With Most Recent Tan et al. Enhancer [25] Using 5-Fold Cross Validation For Enhancer/Non-Enhancer and Strong/Weak Enhancer Prediction Task

Performance over Benchmark Independent Test Set Table 2 reports the performance of proposed Enhancer-DSNet and existing predictors produced over independent test set for enhancer/non-enhancer and independent subset for strong/weak enhancer prediction tasks in terms of accuracy, specificity, sensitivity, and matthews correlation coefficient. According to the Table 2, at layer-1, among all existing predictors excluding most recent Tan et al. Enhancer [25], and -iEnhancer-EL [24] mark better performance across most evaluation metrics. Here, proposed Enhancer-DSNet outperforms most recent Tan et al. Enhancer [25] by the figure of 2%, 1%, 2%, and 5% in terms of sensitivity, specificity, accuracy, and mcc and second best performing -iEnhancer-EL [24] by the figure of 7%, 3%, and 6% in terms of sensitivity, accuracy, and mcc. Whereas, at layer-2, once again proposed Enhancer-DSNet outshines most recent Tan et al. Enhancer [25] by the promising figure of 21% in terms of specificity, 15% in terms of accuracy, and 39% in terms of mcc, and second best performing predictor -iEnhancer-EL [24] by the figure of 29% in term of sensitivity, 22% in terms of accuracy, and 48% in terms of mcc.

Classifiers	Sensitivity	Specificity	Accuracy	MCC
1st Layer (Enhancer/Non-Enhancer)				
Enhancer-DSNet	**0.78**	**0.77**	**0.78**	**0.56**
Tan et al. Enhancer [25]	0.76	0.76	0.76	0.51
iEnhancer-EL [24]	0.71	0.79	0.75	0.50
iEnhancer-2L[4]	0.71	0.75	0.73	0.46
EnhancerPred [21]	0.74	0.75	0.74	0.48
2nd Layer (Strong Enhancer/Weak Enhancer)				
Enhancer-DSNet	**0.83**	**0.67**	**0.83**	**0.70**
Tan et al. Enhancer [25]	0.83	0.46	68.49	0.31
iEnhancer-EL [24]	0.54	0.68	0.61	0.22
iEnhancer-2L[4]	0.47	0.74	0.61	0.22
EnhancerPred [21]	0.45	0.65	0.55	0.10

Table 2: Performance Comparison of Enhancer-DSNet With Existing Enhancer/Non-Enhancer and Strong/Weak Enhancer Predictors Over Independent Test Set

Results Reproduce Ability Issue It is important to mention that recent enhancer determinant and strength prediction approach namely "i-Enhancer-5Step" is given by Lee et al [26]. Authors have utilized unsupervisedly prepared sequence embeddings by treating each nucleotide as word and entire sequence as sentence. Then, these embeddings are passed to SVM classifier. To re-produce reported results [26], we have performed rigorous experimentation using all mentioned parameters [26], but the performance figures we attained are reasonably low than the reported ones [26]. Also authors of most recent predictor namely iEnhancer-EL [24] did not compare their performance figures with Lee et al. i-Enhancer-5Step [26]. Building on this, we consider the results reported in Lee et al. [26] work are fraudulent. Therefore, similar to iEnhancer-EL [24], we also do not compare the performance of proposed Enhancer-DSNet with Lee et al. i-Enhancer-5Step [26].

7 Conclusion

In the marathon of improving the performance of Enhancer identification and their strength prediction, researchers have predominantly employed physico-chemical properties based, bag-of-words based and unsupervisedly prepared k-mer embeddings with different classifiers. Considering these approaches fail to utilize association of inherent sequence relationships with sequence type, we have fused such association by generating sequence embeddings in a supervised fashion which are later fed to a two-layer classification methodology Enhancer-DSNet based on linear classifier. Over a benchmark dataset, proposed Enhancer-DSNet approach outperforms most recent predictor by the figure of 2%, 3%, 2% in terms of accuracy, sensitivity and mcc for enhancer identification task and by the figure of 29%, 4%, 6% in terms of accuracy, specificity, and mcc for strong/weak

enhancer prediction task. This studly findings has opened new doors of further research where biomedical researchers can utilize supervisedly prepared sequence embeddings to enhance the performance of multifarious biomedical tasks.

References

1. title=EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features, author=Jia, Cangzhi and He, Wenying, journal=Scientific reports, volume=6, pages=38741, year=2016, publisher=Nature Publishing Group.
2. title=Enhancer RNAs: insights into their biological role, author=de Lara, Josué Cortés-Fernández and Arzate-Mejía, Rodrigo G and Recillas-Targa, Félix, journal=Epigenetics insights, volume=12, pages=2516865719846093, year=2019, publisher=SAGE Publications Sage UK: London, England.
3. title=Enhancer prediction in proboscis monkey genome: A comparative study, author=Omar, Norshafarina and Wong, Yu Shiong and Li, Xi and Chong, Yee Ling and Abdullah, Mohd Tajuddin and Lee, Nung Kion, journal=Journal of Telecommunication, Electronic and Computer Engineering (JTEC), volume=9, number=2-9, pages=175–179, year=2017.
4. title=iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, author=Liu, Bin and Fang, Longyun and Long, Ren and Lan, Xun and Chou, Kuo-Chen, journal=Bioinformatics, volume=32, number=3, pages=362–369, year=2016, publisher=Oxford University Press.
5. title=Finding distal regulatory elements in the human genome, author=Heintzman, Nathaniel D and Ren, Bing, journal=Current opinion in genetics & development, volume=19, number=6, pages=541–549, year=2009, publisher=Elsevier.
6. title=High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells, author=Boyle, Alan P and Song, Lingyun and Lee, Bum-Kyu and London, Darin and Keefe, Damian and Birney, Ewan and Iyer, Vishwanath R and Crawford, Gregory E and Furey, Terrence S, journal=Genome research, volume=21, number=3, pages=456–464, year=2011, publisher=Cold Spring Harbor Lab.
7. title=ChIP-seq accurately predicts tissue-specific activity of enhancers, author=Visel, Axel and Blow, Matthew J and Li, Zirong and Zhang, Tao and Akiyama, Jennifer A and Holt, Amy and Plajzer-Frick, Ingrid and Shoukry, Malak and Wright, Crystal and Chen, Feng and others, journal=Nature, volume=457, number=7231, pages=854–858, year=2009, publisher=Nature Publishing Group.
8. title=Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, author=Chen, Jun and Liu, H and Yang, J and Chou, K-C, journal=Amino acids, volume=33, number=3, pages=423–428, year=2007, publisher=Springer.
9. title=iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, author=Liu, Bin and Yang, Fan and Huang, De-Shuang and Chou, Kuo-Chen, journal=Bioinformatics, volume=34, number=1, pages=33–40, year=2018, publisher=Oxford University Press.
10. title=Mapping and analysis of chromatin state dynamics in nine human cell types, author=Ernst, Jason and Kheradpour, Pouya and Mikkelsen, Tarjei S and Shoresh, Noam and Ward, Lucas D and Epstein, Charles B and Zhang, Xiaolan and Wang, Li

and Issner, Robbyn and Coyne, Michael and others, journal=Nature, volume=473, number=7345, pages=43–49, year=2011, publisher=Nature Publishing Group.
11. title=Integrating diverse datasets improves developmental enhancer prediction, author=Erwin, Genevieve D and Oksenberg, Nir and Truty, Rebecca M and Kostka, Dennis and Murphy, Karl K and Ahituv, Nadav and Pollard, Katherine S and Capra, John A, journal=PLoS Comput Biol, volume=10, number=6, pages=e1003677, year=2014, publisher=Public Library of Science.
12. title=Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines, author=Fernandez, Michael and Miranda-Saavedra, Diego, journal=Nucleic acids research, volume=40, number=10, pages=e77–e77, year=2012, publisher=Oxford University Press.
13. title=Discover regulatory DNA elements using chromatin signatures and artificial neural network, author=Firpi, Hiram A and Ucar, Duygu and Tan, Kai, journal=Bioinformatics, volume=26, number=13, pages=1579–1586, year=2010, publisher=Oxford University Press.
14. title=Clustering by passing messages between data points, author=Frey, Brendan J and Dueck, Delbert, journal=science, volume=315, number=5814, pages=972–976, year=2007, publisher=American Association for the Advancement of Science.
15. title=DEEP: a general computational framework for predicting enhancers, author=Kleftogiannis, Dimitrios and Kalnis, Panos and Bajic, Vladimir B, journal=Nucleic acids research, volume=43, number=1, pages=e6–e6, year=2015, publisher=Oxford University Press.
16. title=RFECS: a random-forest based algorithm for enhancer identification from chromatin state, author=Rajagopal, Nisha and Xie, Wei and Li, Yan and Wagner, Uli and Wang, Wei and Stamatoyannopoulos, John and Ernst, Jason and Kellis, Manolis and Ren, Bing, journal=PLoS Comput Biol, volume=9, number=3, pages=e1002968, year=2013, publisher=Public Library of Science.
17. title=A new method for enhancer prediction based on deep belief network, author=Bu, Hongda and Gan, Yanglan and Wang, Yang and Zhou, Shuigeng and Guan, Jihong, journal=BMC bioinformatics, volume=18, number=12, pages=418, year=2017, publisher=Springer.
18. title=A new method for enhancer prediction based on deep belief network, author=Bu, Hongda and Gan, Yanglan and Wang, Yang and Zhou, Shuigeng and Guan, Jihong, journal=BMC bioinformatics, volume=18, number=12, pages=418, year=2017, publisher=Springer.
19. title=BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone, author=Yang, Bite and Liu, Feng and Ren, Chao and Ouyang, Zhangyi and Xie, Ziwei and Bo, Xiaochen and Shu, Wenjie, journal=Bioinformatics, volume=33, number=13, pages=1930–1936, year=2017, publisher=Oxford University Press.
20. title=Random forests, author=Breiman, Leo, journal=Machine learning, volume=45, number=1, pages=5–32, year=2001, publisher=Springer.
21. title=EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features, author=Jia, Cangzhi and He, Wenying, journal=Scientific reports, volume=6, pages=38741, year=2016, publisher=Nature Publishing Group.
22. title=dna2vec: Consistent vector representations of variable-length k-mers, author=Ng, Patrick, journal=arXiv preprint arXiv:1701.06279, year=2017.
23. title=Word embeddings: A survey, author=Almeida, Felipe and Xexéo, Geraldo, journal=arXiv preprint arXiv:1901.09069, year=2019.

24. title=iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach, author=Liu, Bin and Li, Kai and Huang, De-Shuang and Chou, Kuo-Chen, journal=Bioinformatics, volume=34, number=22, pages=3835–3842, year=2018, publisher=Oxford University Press.
25. title=Ensemble of Deep Recurrent Neural Networks for Identifying Enhancers via Dinucleotide Physicochemical Properties, author=Tan, Kok Keng and Le, Nguyen Quoc Khanh and Yeh, Hui-Yuan and Chua, Matthew Chin Heng, journal=Cells, volume=8, number=7, pages=767, year=2019, publisher=Multidisciplinary Digital Publishing Institute.
26. title=iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding, author=Le, Nguyen Quoc Khanh and Yapp, Edward Kien Yee and Ho, Quang-Thai and Nagasundaram, N and Ou, Yu-Yen and Yeh, Hui-Yuan, journal=Analytical biochemistry, volume=571, pages=53–61, year=2019, publisher=Elsevier.
27. title=EnhancerPred2. 0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection, author=He, Wenying and Jia, Cangzhi, journal=Molecular Biosystems, volume=13, number=4, pages=767–774, year=2017, publisher=Royal Society of Chemistry.
28. title=A comparison of word embeddings for the biomedical natural language processing, author=Wang, Yanshan and Liu, Sijia and Afzal, Naveed and Rastegar-Mojarad, Majid and Wang, Liwei and Shen, Feichen and Kingsbury, Paul and Liu, Hongfang, journal=Journal of biomedical informatics, volume=87, pages=12–20, year=2018, publisher=Elsevier.
29. title=Learning protein sequence embeddings using information from structure, author=Bepler, Tristan and Berger, Bonnie, journal=arXiv preprint arXiv:1902.08661, year=2019.