**mAHTPred: a sequence-based meta predictor for improving the prediction of anti-hypertensive peptides using effective feature representation**

*Balachandran Manavalan1*†*, Shaherin Basith1*†*, Tae Hwan Shin1,2, Leyi Wei3,\* and Gwang Lee1,2,\**

*1Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea*

*2Institute of Molecular Science and Technology, Ajou University, Suwon, Republic of Korea*

*3School of Computer Science and Technology, Tianjin University, China*

†Contributed equally.

**\* Corresponding authors**

E-mail: weileyi@tju.edu.cn and glee@ajou.ac.kr

**1 Machine learning algorithms**

Four machine learning algorithms are implemented in mATHPred. A brief description of these methods and how they were used in this study as follows:

**1.1 Random forest (RF)**

Breiman (Breiman, 2001) proposed RF algorithm that performs both classification and regression using ensemble of decision trees, and is regarded as one of the powerful ML algorithm which is widely used in bioinformatics and computational biology (Manavalan, et al., 2017; Manavalan, et al., 2014; Manavalan, et al., 2018; Manavalan, et al., 2018; Song, et al., 2017; Wei, et al., 2017). In RF, three key parameters are the number of trees (*ntree*), number of features randomly selected (*mtry*), and the minimum number of samples required to split and internal node (*nsplit*). Grid search was employed to fine-tune these parameters with the following search space:

(1)

**1.2 Extremely randomized tree (ERT)**

Geurts et al. (Geurts, et al., 2006) proposed the ERT algorithm that perform classification and regression by utilizing hundreds or thousands of independent decision trees, which has been applied in a large number of biological problems (Manavalan, et al., 2018; Manavalan, et al., 2018). ERT aims to further decrease the variance of the prediction model by including stronger randomization techniques. The ERT algorithm is similar to RF, but with the following differences: (*i*) ERT does not apply a bagging procedure for the construction of each tree. Instead, it uses the whole input training set for the construction of each tree. (*ii*) ERT selects a node split very randomly (both a variable index and variable splitting values are chosen randomly), whereas RF finds the best split (optimized by a variable index and a variable splitting value) among a random subset of variables. Furthermore, Grid search was performed to fine-tune the regularization parameters *ntree*, *mtry*, and *nsplit*, whose search space are as follows:

(2)

**1.3 Support vector machine (SVM)**

The objective of SVM is to find the optimal hyperplane with the largest margin to decrease the misclassification rate (Noble, 2006). Basically, it maps the given input features into a high-dimensional space using kernel functions and finds a hyperplane that maximizes the distance between the hyperplane and two classes. For a given test sample that was mapped into the high-dimensional space (as described above), and SVM can predict the test sample based on which side of the hyperplane they fall in. Notably, SVM is one of the most widely used ML algorithms applied to various classification problems (Cao, et al., 2014; Chen, et al., 2018; Feng, et al., 2018; Manavalan and Lee, 2017; Manavalan, et al., 2018; Wei, et al., 2018; Wei, et al., 2018). In this study, we experimented different kernel functions including, linear, polynomial functions, and Gaussian radial-basis function (RBF). Of these, RBF kernel was the most suitable for addressing our problem. Two critical parameters, *C* (controls the trade-off between the training error and margin) and g (controls how peaked Gaussians are centered on the support vectors) requires optimization in RBF-SVM. Therefore, we optimized these parameters using the following range:

(3)

**1.4 Gradient boosting (GB)**

Friedman proposed the GB algorithm (Friedman, 2001), which is a forward learning ensemble method that performs both classification and regression problems. GB produces a final strong prediction model based on the ensemble of weak models (decision trees), which has been widely used in bioinformatics and computational biology (Manavalan, et al., 2018; Rawi, et al., 2018). GB consecutively fits new models to provide a more accurate estimate of the response variables, compared to other ensemble methods such as RF, ERT, and AB. In GB, the three most influential parameters are *ntree*, *mtry*, and *nsplit*, we optimized with the following search space:

(4)

2 10-fold cross-validation

Generally, three CV methods, namely an independent data set test, a sub-sampling (or *k*-fold CV) test, and a leave-one-out CV (LOOCV) test, are often used to evaluate the anticipated success rate of a predictor. Among the three methods, however, the LOOCV test is deemed the least arbitrary and most objective as demonstrated by Eqs.28-32 of (Chou, 2011), and hence it has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (Cao, et al., 2017; Cao, et al., 2016; Cao, et al., 2017; Cao, et al., 2014; Chen, et al., 2018; Chen, et al., 2013; Chen, et al., 2016; Chen, et al., 2017; Kosylo, et al., 2018; Lai, et al., 2017; Lin, et al., 2012; Lin, et al., 2017; Liu, et al., 2018; Tang, et al., 2017; Yang, et al., 2016; Yang, et al., 2018; Zhao, et al., 2017). However, it seems time- and source-consuming. Thus, we used 10-fold CV to examine the proposed models. In 10-fold CV, the benchmarking dataset was randomly partitioned into 10 subsets. One subset is used as a test set and the remaining nine subsets are used as the training sets. This procedure is repeated for 10 times, where each subset being used once as a test set. The performance of the 10 corresponding results are averaged that implies the performance of the classifier.

3 Performance Evaluation

To evaluate the performance of the constructed models, we used four measurements that were commonly used in binary classification tasks (Feng, et al., 2013; Liu, et al., 2018; Liu, et al., 2018; Liu, et al., 2017; Tang, et al., 2018), including sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC). They are calculated as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | |  |  | | --- | --- | |  | (5) | |  | |

where TP is the number of true positives (i.e., AHTPs classified correctly as AHTPs) and TN is the number of true negatives (i.e., non-AHTPs classified correctly as non-AHTPs). FP is the number of false positives (i.e., AHTPs classified incorrectly as non-AHTPs) and FN is the number of false negatives (i.e., non-AHTPs classified incorrectly as AHTPs). Additionally, the receiver operating characteristic (ROC) curve, which is a plot of the true positive rate against the false positive rate under different classification thresholds, is depicted to visually measure the comprehensive performance of different classifiers.

**4 Implementation of a webserver**

Generally, user-friendly web servers have been helpful for experimentalists, where they can do the prediction without going through mathematical equations, and also it represent the future direction for developing novel and more useful predictors (Chou and Shen, 2009). Indeed, it has been demonstrated by a series of publications (Chen, et al., 2017; Chen, et al., 2016; Han, et al., 2018; McDermaid, et al., 2018). In this view, we established a user-friendly webserver, mAHTPred, for a wider research community, which is freely accessible via http:/thegleelab.org/mAHTPred. To validate our findings, all data sets utilized in this study can be freely downloaded from our web server. Below, we provide researchers a simple two-step guideline on how to use our webserver in order to obtain the predicted outcomes. In the first step, users need to submit the query sequences into the input box. Note that the input sequences should be in FASTA format. Examples of FASTA-formatted sequences can be seen by clicking on the FASTA format button which is located above the input box. As a final step, clicking on the ‘Submit’ button will provide the predicted results as the output.

**5. Motif analysis**

We combined benchmark dataset and independent datasets and applied MERCI program (Vens, et al., 2011) to identify the motifs present in both AHTPs and non-AHTPs. Results showed that 11 exclusive motifs were present in AHTPs (“PFP”, “LHL”, “YPF”, “LHLP”, “IYP”, “YPFP”, “APFP”, “PFPG”, “VAPF”, “VAPFP”, “YPFPG”), whereas none was found in non-AHTPs.

**Figure S1.** ROC curves of mAHTPred and the existing methods on the independent dataset.

****

**Table S1.** The details of hybrid features generated with different combination and the final input features dimensions.

|  |  |  |
| --- | --- | --- |
| Hybrid feature | Combinations | Input features dimensions |
| H1 | AAC, OF | 25 |
| H2 | DPC, OF | 405 |
| H3 | AAI, OF | 25 |
| H4 | CTD, OF | 152 |
| H5 | AAC, DPC | 420 |
| H6 | AAC, AAI | 40 |
| H7 | AAC, CTD | 167 |
| H8 | AAI, DPC | 420 |
| H9 | DPC, CTD | 547 |
| H10 | AAI, CTD | 167 |
| H11 | AAC, AAI, DPC | 440 |
| H12 | AAC, AAI, CTD | 187 |
| H13 | AAC, CTD, OF | 172 |
| H14 | AAI, DPC, CTD | 567 |
| H15 | DPC, CTD, OF | 552 |
| H16 | AAI, CTD, OF | 172 |
| H17 | AAC, AAI, DPC, CTD | 587 |
| H18 | AAI, DPC, CTD, OF | 572 |
| H19 | AAC, AAI, CTD, OF | 192 |
| H20 | AAC, DPC, AAI, CTD, OF | 592 |

**Table S2.** The details of standard amino acid residue classification based on ten physicochemical properties used in OVP.

|  |  |
| --- | --- |
| **Physicochemical property** | **Amino acid group** |
| Aromatic | F, Y, W, H |
| Negative | D, E |
| Positive | K, R, H |
| Polar | N, Q, S, D, E, C, T, K, R, H, Y, W |
| Hydrophobic | A, G, C, T, I, V, L, K, H, F, Y, W, M |
| Aliphatic | I, V, L |
| Tiny | A, S, G, C |
| Charged | K, H, R, D, E |
| Small | P, N, D, T, C, A, G, S, V |
| Imino acid | P |

**Table S3.** The details of standard amino acid classification based on seven physicochemical properties used in 21-bit.

|  |  |  |  |
| --- | --- | --- | --- |
| **Physicochemical property** | **Group 1** | **Group 2** | **page30image79448Group 3** |
| Hydrophobicity | A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y | D, E | K, R |
| Normalized Van der Waals volume | C, F, I, L, M, V, W | A, G, H, P, S, T, Y | D, E, K, N, Q, R |
| Polarity | A, C, D, G, P, S, T | E, I, L, N, Q, V | F, H, K, M, R, W, Y |
| Polarizability | C, F, I, L, M, V, W, Y | A, G, P, S, T | D, E, H, K, N, Q, R |
| Charge | A, D, G, S, T | C, E, I, L, N, P, Q, V | F, H, K, M, R, W, Y |
| Secondary structures | D, G, N, P, S | A, E, H, K, L, M, Q, R | C, F, I, T, V, W, Y |
| Solvent Accessibility | A, C, F, G, I, L, V, W | H, M, P, S, T, Y | D, E, K, N, R, Q |

**Table S4.** List of all the 51 feature descriptors

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature identifier | Feature descriptor | Dimensions | Feature identifier | Feature descriptor | Dimensions |
| F1 | AAC | 20 | F27 | TOBC4 | 80 |
| F2 | DPC | 400 | F28 | TOBNC4 | 160 |
| F3 | CTD | 147 | F29 | TOBN5 | 100 |
| F4 | AAI | 20 | F30 | TOBC5 | 100 |
| F5 | BPFN3 | 60 | F31 | TOBNC5 | 200 |
| F6 | BPFC3 | 60 | F32 | H1 | 25 |
| F7 | BPFNC3 | 120 | F33 | H2 | 405 |
| F8 | BPFN4 | 80 | F34 | H3 | 25 |
| F9 | BPFC4 | 80 | F35 | H4 | 152 |
| F10 | BPFNC4 | 160 | F36 | H5 | 420 |
| F11 | BPFN5 | 100 | F37 | H6 | 40 |
| F12 | BPFC5 | 100 | F38 | H7 | 167 |
| F13 | BPFNC5 | 200 | F39 | H8 | 420 |
| F14 | OVPN3 | 60 | F40 | H9 | 547 |
| F15 | OVPC3 | 60 | F41 | H10 | 167 |
| F16 | OVPNC3 | 120 | F42 | H11 | 440 |
| F17 | OVPN4 | 80 | F43 | H12 | 187 |
| F18 | OVPC4 | 80 | F44 | H13 | 172 |
| F19 | OVPNC4 | 160 | F45 | H14 | 567 |
| F20 | OVPN5 | 100 | F46 | H15 | 552 |
| F21 | OVPC5 | 100 | F47 | H16 | 172 |
| F22 | OVPNC5 | 200 | F48 | H17 | 587 |
| F23 | TOBN3 | 60 | F49 | H18 | 572 |
| F24 | TOBC3 | 60 | F50 | H19 | 192 |
| F25 | TOBNC3 | 120 | F51 | H20 | 592 |
| F26 | TOBN4 | 80 |  |  |  |

**Table S5.** Performance of the 51 individual feature descriptors with RF classifier on the benchmarking dataset during 10-fold CV, where the descriptors are ranked according to the accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptor** | **MCC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| CTD | 0.643 | 0.820 | 0.777 | 0.863 | 0.884 |
| H8 | 0.627 | 0.813 | 0.788 | 0.839 | 0.878 |
| H11 | 0.625 | 0.812 | 0.785 | 0.839 | 0.875 |
| H5 | 0.624 | 0.812 | 0.783 | 0.840 | 0.879 |
| H20 | 0.624 | 0.811 | 0.774 | 0.848 | 0.870 |
| H17 | 0.623 | 0.811 | 0.767 | 0.854 | 0.875 |
| H18 | 0.616 | 0.807 | 0.763 | 0.850 | 0.870 |
| H9 | 0.612 | 0.805 | 0.766 | 0.845 | 0.870 |
| H14 | 0.612 | 0.805 | 0.768 | 0.842 | 0.873 |
| DPC | 0.605 | 0.802 | 0.777 | 0.828 | 0.871 |
| H2 | 0.603 | 0.801 | 0.785 | 0.817 | 0.874 |
| H15 | 0.602 | 0.800 | 0.762 | 0.838 | 0.865 |
| PBFNC5 | 0.600 | 0.800 | 0.780 | 0.819 | 0.872 |
| H12 | 0.601 | 0.799 | 0.744 | 0.853 | 0.863 |
| H19 | 0.600 | 0.799 | 0.747 | 0.850 | 0.861 |
| H10 | 0.588 | 0.793 | 0.751 | 0.835 | 0.857 |
| H7 | 0.585 | 0.791 | 0.740 | 0.841 | 0.858 |
| H13 | 0.585 | 0.791 | 0.739 | 0.842 | 0.856 |
| H16 | 0.583 | 0.790 | 0.742 | 0.839 | 0.860 |
| H3 | 0.578 | 0.789 | 0.761 | 0.816 | 0.852 |
| H1 | 0.576 | 0.788 | 0.760 | 0.815 | 0.850 |
| BPFNC4 | 0.576 | 0.788 | 0.768 | 0.807 | 0.851 |
| OVPC5 | 0.579 | 0.788 | 0.733 | 0.842 | 0.848 |
| AAC | 0.575 | 0.787 | 0.756 | 0.818 | 0.849 |
| H6 | 0.574 | 0.786 | 0.756 | 0.817 | 0.849 |
| BPFC5 | 0.575 | 0.786 | 0.746 | 0.827 | 0.859 |
| AAI | 0.572 | 0.786 | 0.765 | 0.807 | 0.848 |
| OVPNC5 | 0.565 | 0.781 | 0.731 | 0.831 | 0.847 |
| H4 | 0.564 | 0.780 | 0.727 | 0.834 | 0.848 |
| OVPNC4 | 0.551 | 0.774 | 0.720 | 0.828 | 0.828 |
| BPFC4 | 0.548 | 0.773 | 0.736 | 0.811 | 0.838 |
| BPFNC3 | 0.542 | 0.771 | 0.744 | 0.797 | 0.835 |
| OVPC4 | 0.535 | 0.767 | 0.727 | 0.806 | 0.826 |
| OVPNC3 | 0.530 | 0.764 | 0.723 | 0.805 | 0.823 |
| BPFN5 | 0.520 | 0.759 | 0.716 | 0.802 | 0.823 |
| OVPN5 | 0.519 | 0.759 | 0.724 | 0.794 | 0.812 |
| OVPC3 | 0.459 | 0.728 | 0.679 | 0.778 | 0.774 |
| OVPN4 | 0.456 | 0.727 | 0.687 | 0.768 | 0.774 |
| BPFC3 | 0.450 | 0.725 | 0.692 | 0.757 | 0.787 |
| BPFN4 | 0.447 | 0.723 | 0.685 | 0.761 | 0.781 |
| TOBNC5 | 0.429 | 0.714 | 0.699 | 0.730 | 0.776 |
| TOBN5 | 0.422 | 0.711 | 0.698 | 0.724 | 0.770 |
| TOBNC4 | 0.416 | 0.708 | 0.668 | 0.747 | 0.765 |
| TOBN4 | 0.402 | 0.701 | 0.685 | 0.717 | 0.751 |
| BPFN3 | 0.390 | 0.695 | 0.674 | 0.716 | 0.739 |
| TOBNC3 | 0.385 | 0.692 | 0.699 | 0.686 | 0.745 |
| OVPN3 | 0.362 | 0.681 | 0.650 | 0.712 | 0.731 |
| TOBN3 | 0.337 | 0.668 | 0.646 | 0.690 | 0.710 |
| TOBC5 | 0.318 | 0.659 | 0.648 | 0.669 | 0.709 |
| TOBC3 | 0.293 | 0.646 | 0.642 | 0.651 | 0.691 |
| TOBC4 | 0.290 | 0.645 | 0.657 | 0.633 | 0.694 |

**Table S6.** Performance of the 51 individual feature descriptors with ERT classifier on the benchmarking dataset during 10-fold CV, where the descriptors are ranked according to the accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptor** | **MCC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| H11 | 0.646 | 0.823 | 0.793 | 0.852 | 0.883 |
| H20 | 0.647 | 0.823 | 0.788 | 0.858 | 0.882 |
| H17 | 0.641 | 0.820 | 0.786 | 0.853 | 0.881 |
| H18 | 0.641 | 0.820 | 0.785 | 0.854 | 0.878 |
| H14 | 0.639 | 0.819 | 0.780 | 0.858 | 0.877 |
| H9 | 0.637 | 0.818 | 0.777 | 0.859 | 0.878 |
| H5 | 0.636 | 0.817 | 0.782 | 0.852 | 0.880 |
| H8 | 0.635 | 0.817 | 0.783 | 0.850 | 0.879 |
| H15 | 0.635 | 0.817 | 0.779 | 0.854 | 0.875 |
| BPFNC5 | 0.615 | 0.807 | 0.783 | 0.831 | 0.873 |
| H12 | 0.617 | 0.807 | 0.762 | 0.852 | 0.869 |
| H2 | 0.613 | 0.806 | 0.788 | 0.825 | 0.879 |
| H19 | 0.616 | 0.806 | 0.755 | 0.858 | 0.866 |
| H13 | 0.611 | 0.804 | 0.756 | 0.852 | 0.865 |
| DPC | 0.607 | 0.803 | 0.789 | 0.818 | 0.877 |
| H16 | 0.607 | 0.802 | 0.751 | 0.852 | 0.866 |
| H7 | 0.601 | 0.799 | 0.746 | 0.852 | 0.865 |
| H10 | 0.601 | 0.799 | 0.748 | 0.850 | 0.866 |
| AAC | 0.595 | 0.797 | 0.762 | 0.831 | 0.856 |
| BPFNC4 | 0.592 | 0.796 | 0.773 | 0.818 | 0.857 |
| AAI | 0.591 | 0.795 | 0.757 | 0.832 | 0.853 |
| H6 | 0.590 | 0.795 | 0.772 | 0.817 | 0.855 |
| H1 | 0.588 | 0.794 | 0.765 | 0.823 | 0.856 |
| H3 | 0.588 | 0.794 | 0.776 | 0.812 | 0.858 |
| H4 | 0.578 | 0.788 | 0.736 | 0.839 | 0.855 |
| OVPC5 | 0.576 | 0.786 | 0.735 | 0.838 | 0.848 |
| CTD | 0.574 | 0.785 | 0.728 | 0.842 | 0.851 |
| BPFC5 | 0.570 | 0.784 | 0.743 | 0.826 | 0.855 |
| OVPNC5 | 0.573 | 0.784 | 0.725 | 0.843 | 0.843 |
| BPFNC3 | 0.547 | 0.773 | 0.753 | 0.794 | 0.842 |
| OVPNC4 | 0.547 | 0.772 | 0.717 | 0.826 | 0.834 |
| BPFC4 | 0.542 | 0.769 | 0.721 | 0.818 | 0.836 |
| OVPC4 | 0.536 | 0.767 | 0.716 | 0.817 | 0.825 |
| OVPNC3 | 0.529 | 0.764 | 0.734 | 0.794 | 0.824 |
| OVPN5 | 0.520 | 0.759 | 0.701 | 0.816 | 0.812 |
| BPFN5 | 0.517 | 0.757 | 0.703 | 0.811 | 0.822 |
| BPFC3 | 0.455 | 0.727 | 0.690 | 0.763 | 0.780 |
| OVPN4 | 0.451 | 0.725 | 0.693 | 0.757 | 0.776 |
| BPFN4 | 0.442 | 0.720 | 0.682 | 0.758 | 0.777 |
| OVPC3 | 0.437 | 0.718 | 0.689 | 0.747 | 0.771 |
| TOBN5 | 0.431 | 0.715 | 0.693 | 0.737 | 0.769 |
| TOBNC5 | 0.426 | 0.713 | 0.698 | 0.728 | 0.774 |
| TOBNC4 | 0.418 | 0.709 | 0.680 | 0.737 | 0.760 |
| TOBN4 | 0.384 | 0.692 | 0.677 | 0.707 | 0.747 |
| TOBNC3 | 0.380 | 0.690 | 0.690 | 0.690 | 0.738 |
| BPFN3 | 0.369 | 0.685 | 0.664 | 0.705 | 0.735 |
| OVPN3 | 0.356 | 0.678 | 0.653 | 0.703 | 0.727 |
| TOBC4 | 0.316 | 0.658 | 0.650 | 0.667 | 0.717 |
| TOBN3 | 0.311 | 0.656 | 0.646 | 0.665 | 0.706 |
| TOBC5 | 0.306 | 0.653 | 0.636 | 0.669 | 0.698 |
| TOBC3 | 0.283 | 0.641 | 0.624 | 0.658 | 0.671 |

**Table S7.** Performance of the 51 individual feature descriptors with SVM classifier on the benchmarking dataset during 10-fold CV, where the descriptors are ranked according to the accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptor** | **MCC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| H8 | 0.628 | 0.813 | 0.768 | 0.858 | 0.877 |
| H5 | 0.627 | 0.809 | 0.731 | 0.888 | 0.881 |
| H11 | 0.624 | 0.807 | 0.714 | 0.899 | 0.881 |
| DPC | 0.606 | 0.802 | 0.758 | 0.846 | 0.871 |
| BPFNC5 | 0.599 | 0.797 | 0.735 | 0.860 | 0.870 |
| BPFNC4 | 0.585 | 0.792 | 0.756 | 0.828 | 0.855 |
| H2 | 0.586 | 0.789 | 0.708 | 0.871 | 0.866 |
| BPFC5 | 0.578 | 0.785 | 0.705 | 0.865 | 0.859 |
| H20 | 0.565 | 0.783 | 0.771 | 0.794 | 0.843 |
| H17 | 0.559 | 0.779 | 0.770 | 0.789 | 0.844 |
| H18 | 0.553 | 0.777 | 0.777 | 0.777 | 0.841 |
| H6 | 0.554 | 0.776 | 0.734 | 0.818 | 0.842 |
| H14 | 0.552 | 0.776 | 0.765 | 0.788 | 0.841 |
| H15 | 0.550 | 0.775 | 0.773 | 0.777 | 0.840 |
| H7 | 0.549 | 0.774 | 0.766 | 0.783 | 0.832 |
| H9 | 0.549 | 0.774 | 0.766 | 0.783 | 0.840 |
| H13 | 0.549 | 0.774 | 0.766 | 0.783 | 0.830 |
| H12 | 0.548 | 0.774 | 0.766 | 0.782 | 0.833 |
| BPFNC3 | 0.555 | 0.773 | 0.680 | 0.865 | 0.844 |
| H19 | 0.546 | 0.773 | 0.765 | 0.781 | 0.829 |
| OVPC5 | 0.561 | 0.772 | 0.651 | 0.894 | 0.842 |
| OVPNC5 | 0.556 | 0.772 | 0.671 | 0.873 | 0.839 |
| AAC | 0.544 | 0.771 | 0.730 | 0.813 | 0.838 |
| H1 | 0.536 | 0.767 | 0.725 | 0.809 | 0.833 |
| OVPNC4 | 0.537 | 0.767 | 0.710 | 0.824 | 0.831 |
| H16 | 0.536 | 0.767 | 0.717 | 0.816 | 0.824 |
| BPFC4 | 0.561 | 0.766 | 0.608 | 0.924 | 0.836 |
| H4 | 0.535 | 0.766 | 0.717 | 0.815 | 0.822 |
| AAI | 0.531 | 0.766 | 0.755 | 0.777 | 0.826 |
| H10 | 0.531 | 0.766 | 0.757 | 0.774 | 0.827 |
| CTD | 0.525 | 0.760 | 0.685 | 0.835 | 0.821 |
| H3 | 0.515 | 0.757 | 0.750 | 0.765 | 0.824 |
| OVPNC3 | 0.509 | 0.753 | 0.700 | 0.806 | 0.821 |
| OVPC4 | 0.516 | 0.750 | 0.628 | 0.873 | 0.817 |
| OVPN5 | 0.510 | 0.747 | 0.623 | 0.871 | 0.808 |
| BPFC3 | 0.471 | 0.734 | 0.686 | 0.783 | 0.784 |
| BPFN4 | 0.496 | 0.731 | 0.548 | 0.914 | 0.784 |
| BPFN5 | 0.524 | 0.728 | 0.481 | 0.975 | 0.827 |
| OVPC3 | 0.451 | 0.725 | 0.688 | 0.762 | 0.773 |
| OVPN4 | 0.448 | 0.722 | 0.663 | 0.782 | 0.777 |
| TOBNC5 | 0.429 | 0.713 | 0.659 | 0.767 | 0.765 |
| TOBN5 | 0.425 | 0.711 | 0.646 | 0.776 | 0.760 |
| TOBNC4 | 0.419 | 0.709 | 0.681 | 0.737 | 0.759 |
| BPFN3 | 0.404 | 0.700 | 0.627 | 0.773 | 0.737 |
| TOBNC3 | 0.374 | 0.687 | 0.662 | 0.712 | 0.739 |
| OVPN3 | 0.362 | 0.676 | 0.559 | 0.793 | 0.719 |
| TOBN3 | 0.344 | 0.672 | 0.668 | 0.676 | 0.709 |
| TOBC5 | 0.330 | 0.665 | 0.648 | 0.681 | 0.709 |
| TOBN4 | 0.394 | 0.664 | 0.389 | 0.940 | 0.669 |
| TOBC4 | 0.314 | 0.654 | 0.555 | 0.753 | 0.692 |
| TOBC3 | 0.294 | 0.646 | 0.578 | 0.713 | 0.701 |

**Table S8.** Performance of the 51 individual feature descriptors with GB classifier on the benchmarking dataset during 10-fold CV, where the descriptors are ranked according to the accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptor** | **MCC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| H11 | 0.619 | 0.809 | 0.795 | 0.824 | 0.861 |
| H8 | 0.615 | 0.807 | 0.789 | 0.826 | 0.867 |
| H17 | 0.612 | 0.805 | 0.762 | 0.848 | 0.867 |
| H5 | 0.609 | 0.805 | 0.790 | 0.819 | 0.863 |
| H14 | 0.607 | 0.802 | 0.760 | 0.845 | 0.864 |
| H20 | 0.603 | 0.801 | 0.759 | 0.842 | 0.867 |
| H9 | 0.595 | 0.796 | 0.750 | 0.842 | 0.853 |
| H18 | 0.594 | 0.796 | 0.757 | 0.836 | 0.860 |
| H2 | 0.592 | 0.796 | 0.785 | 0.806 | 0.856 |
| DPC | 0.590 | 0.795 | 0.797 | 0.793 | 0.851 |
| H12 | 0.591 | 0.795 | 0.760 | 0.829 | 0.851 |
| BPFNC5 | 0.588 | 0.794 | 0.780 | 0.808 | 0.852 |
| H15 | 0.590 | 0.794 | 0.754 | 0.835 | 0.862 |
| H19 | 0.589 | 0.793 | 0.743 | 0.843 | 0.852 |
| H7 | 0.583 | 0.790 | 0.748 | 0.832 | 0.846 |
| H10 | 0.581 | 0.790 | 0.754 | 0.826 | 0.855 |
| H16 | 0.579 | 0.789 | 0.751 | 0.826 | 0.847 |
| H13 | 0.576 | 0.786 | 0.738 | 0.835 | 0.850 |
| BPFNC4 | 0.566 | 0.783 | 0.765 | 0.801 | 0.846 |
| H1 | 0.566 | 0.783 | 0.753 | 0.813 | 0.833 |
| AAC | 0.563 | 0.782 | 0.762 | 0.801 | 0.828 |
| AAI | 0.558 | 0.779 | 0.761 | 0.796 | 0.827 |
| OVPC5 | 0.558 | 0.778 | 0.738 | 0.818 | 0.831 |
| H3 | 0.557 | 0.778 | 0.756 | 0.801 | 0.833 |
| BPFC5 | 0.556 | 0.778 | 0.746 | 0.809 | 0.828 |
| OVPNC5 | 0.558 | 0.778 | 0.733 | 0.823 | 0.837 |
| H6 | 0.550 | 0.775 | 0.757 | 0.793 | 0.825 |
| CTD | 0.549 | 0.774 | 0.742 | 0.806 | 0.839 |
| H4 | 0.550 | 0.774 | 0.728 | 0.819 | 0.836 |
| OVPNC4 | 0.538 | 0.768 | 0.737 | 0.800 | 0.825 |
| BPFNC3 | 0.531 | 0.766 | 0.753 | 0.779 | 0.817 |
| BPFC4 | 0.525 | 0.762 | 0.723 | 0.801 | 0.821 |
| OVPC4 | 0.517 | 0.758 | 0.731 | 0.785 | 0.790 |
| OVPNC3 | 0.518 | 0.758 | 0.719 | 0.797 | 0.811 |
| BPFN5 | 0.503 | 0.751 | 0.725 | 0.778 | 0.796 |
| OVPN5 | 0.502 | 0.750 | 0.717 | 0.783 | 0.779 |
| BPFC3 | 0.427 | 0.713 | 0.691 | 0.735 | 0.735 |
| OVPN4 | 0.426 | 0.713 | 0.701 | 0.725 | 0.745 |
| BPFN4 | 0.422 | 0.711 | 0.696 | 0.726 | 0.748 |
| TOBNC5 | 0.415 | 0.708 | 0.690 | 0.725 | 0.754 |
| OVPC3 | 0.413 | 0.707 | 0.682 | 0.731 | 0.717 |
| TOBN5 | 0.408 | 0.704 | 0.707 | 0.701 | 0.730 |
| TOBNC4 | 0.398 | 0.699 | 0.679 | 0.719 | 0.743 |
| TOBN4 | 0.374 | 0.687 | 0.680 | 0.693 | 0.721 |
| TOBNC3 | 0.356 | 0.678 | 0.679 | 0.677 | 0.713 |
| BPFN3 | 0.351 | 0.675 | 0.655 | 0.696 | 0.707 |
| OVPN3 | 0.328 | 0.664 | 0.634 | 0.693 | 0.688 |
| TOBC5 | 0.300 | 0.650 | 0.632 | 0.668 | 0.685 |
| TOBN3 | 0.276 | 0.638 | 0.623 | 0.653 | 0.660 |
| TOBC4 | 0.270 | 0.635 | 0.641 | 0.629 | 0.658 |
| TOBC3 | 0.266 | 0.633 | 0.599 | 0.666 | 0.644 |

**Table S9.** Performance of the 51 individual feature descriptors with AB classifier on the benchmarking dataset during 10-fold CV, where the descriptors are ranked according to the accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptor** | **MCC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| H15 | 0.537 | 0.768 | 0.778 | 0.759 | 0.770 |
| H9 | 0.524 | 0.762 | 0.771 | 0.753 | 0.763 |
| H20 | 0.519 | 0.760 | 0.759 | 0.760 | 0.760 |
| H18 | 0.518 | 0.759 | 0.758 | 0.760 | 0.760 |
| H14 | 0.509 | 0.755 | 0.757 | 0.753 | 0.755 |
| H17 | 0.509 | 0.755 | 0.757 | 0.753 | 0.755 |
| H7 | 0.508 | 0.754 | 0.732 | 0.776 | 0.755 |
| H10 | 0.508 | 0.754 | 0.732 | 0.776 | 0.755 |
| H12 | 0.508 | 0.754 | 0.732 | 0.776 | 0.755 |
| BPFNC5 | 0.495 | 0.748 | 0.733 | 0.762 | 0.748 |
| CTD | 0.490 | 0.745 | 0.742 | 0.748 | 0.745 |
| H13 | 0.490 | 0.745 | 0.745 | 0.745 | 0.745 |
| H16 | 0.490 | 0.745 | 0.745 | 0.745 | 0.745 |
| H19 | 0.490 | 0.745 | 0.745 | 0.745 | 0.745 |
| OVPNC5 | 0.487 | 0.743 | 0.704 | 0.781 | 0.743 |
| H4 | 0.482 | 0.741 | 0.743 | 0.739 | 0.742 |
| H5 | 0.479 | 0.739 | 0.727 | 0.751 | 0.743 |
| H8 | 0.479 | 0.739 | 0.727 | 0.751 | 0.743 |
| H11 | 0.479 | 0.739 | 0.727 | 0.751 | 0.743 |
| H1 | 0.478 | 0.739 | 0.716 | 0.761 | 0.737 |
| H3 | 0.478 | 0.739 | 0.716 | 0.761 | 0.737 |
| DPC | 0.477 | 0.734 | 0.638 | 0.830 | 0.741 |
| BPFC5 | 0.465 | 0.732 | 0.725 | 0.739 | 0.732 |
| H2 | 0.473 | 0.731 | 0.621 | 0.840 | 0.738 |
| AAC | 0.460 | 0.730 | 0.702 | 0.757 | 0.730 |
| AAI | 0.460 | 0.730 | 0.702 | 0.757 | 0.730 |
| H6 | 0.460 | 0.730 | 0.702 | 0.757 | 0.730 |
| BPFNC4 | 0.458 | 0.729 | 0.719 | 0.739 | 0.729 |
| BPFC4 | 0.454 | 0.727 | 0.703 | 0.750 | 0.727 |
| OVPNC4 | 0.451 | 0.725 | 0.681 | 0.768 | 0.725 |
| BPFNC3 | 0.442 | 0.721 | 0.694 | 0.747 | 0.721 |
| OVPC5 | 0.441 | 0.720 | 0.686 | 0.755 | 0.720 |
| OVPNC3 | 0.433 | 0.716 | 0.676 | 0.756 | 0.716 |
| BPFN5 | 0.416 | 0.708 | 0.679 | 0.736 | 0.708 |
| OVPC4 | 0.403 | 0.700 | 0.648 | 0.753 | 0.700 |
| OVPN5 | 0.391 | 0.694 | 0.648 | 0.740 | 0.694 |
| BPFC3 | 0.387 | 0.693 | 0.655 | 0.731 | 0.693 |
| OVPC3 | 0.361 | 0.678 | 0.596 | 0.760 | 0.678 |
| OVPN4 | 0.354 | 0.677 | 0.653 | 0.701 | 0.677 |
| TOBNC4 | 0.324 | 0.662 | 0.622 | 0.701 | 0.662 |
| BPFN4 | 0.323 | 0.661 | 0.627 | 0.696 | 0.661 |
| TOBNC5 | 0.316 | 0.658 | 0.631 | 0.685 | 0.657 |
| OVPN3 | 0.307 | 0.653 | 0.631 | 0.676 | 0.653 |
| TOBNC3 | 0.307 | 0.653 | 0.609 | 0.697 | 0.653 |
| TOBC5 | 0.300 | 0.650 | 0.613 | 0.686 | 0.648 |
| TOBN5 | 0.293 | 0.646 | 0.620 | 0.673 | 0.645 |
| TOBC4 | 0.292 | 0.645 | 0.599 | 0.691 | 0.645 |
| BPFN3 | 0.287 | 0.644 | 0.617 | 0.670 | 0.644 |
| TOBC3 | 0.273 | 0.636 | 0.607 | 0.666 | 0.636 |
| TOBN4 | 0.272 | 0.636 | 0.599 | 0.673 | 0.636 |
| TOBN3 | 0.257 | 0.628 | 0.586 | 0.670 | 0.628 |

**Table S10.** Performance of the 51 individual feature descriptors with *k*-NN classifier on the benchmarking dataset during 10-fold CV, where the descriptors are ranked according to the accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptor** | **MCC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| H6 | 0.517 | 0.757 | 0.698 | 0.816 | 0.817 |
| AAC | 0.498 | 0.749 | 0.715 | 0.782 | 0.819 |
| H5 | 0.503 | 0.744 | 0.627 | 0.862 | 0.815 |
| H1 | 0.486 | 0.742 | 0.682 | 0.801 | 0.807 |
| H11 | 0.486 | 0.740 | 0.671 | 0.809 | 0.813 |
| OVPNC5 | 0.474 | 0.736 | 0.693 | 0.779 | 0.808 |
| H8 | 0.486 | 0.734 | 0.602 | 0.866 | 0.808 |
| DPC | 0.482 | 0.727 | 0.562 | 0.893 | 0.808 |
| OVPC5 | 0.454 | 0.726 | 0.690 | 0.762 | 0.791 |
| OVPNC3 | 0.446 | 0.721 | 0.658 | 0.784 | 0.783 |
| OVPNC4 | 0.445 | 0.721 | 0.670 | 0.772 | 0.794 |
| H2 | 0.446 | 0.721 | 0.664 | 0.779 | 0.780 |
| BPFNC5 | 0.446 | 0.717 | 0.830 | 0.605 | 0.824 |
| H7 | 0.427 | 0.713 | 0.747 | 0.679 | 0.781 |
| H12 | 0.427 | 0.713 | 0.748 | 0.678 | 0.782 |
| H9 | 0.426 | 0.713 | 0.680 | 0.745 | 0.778 |
| H14 | 0.426 | 0.713 | 0.678 | 0.747 | 0.778 |
| CTD | 0.425 | 0.712 | 0.676 | 0.748 | 0.778 |
| H17 | 0.425 | 0.712 | 0.676 | 0.748 | 0.781 |
| BPFNC4 | 0.432 | 0.711 | 0.816 | 0.607 | 0.807 |
| H10 | 0.423 | 0.711 | 0.675 | 0.747 | 0.779 |
| BPFC5 | 0.420 | 0.708 | 0.780 | 0.635 | 0.790 |
| BPFNC3 | 0.420 | 0.708 | 0.780 | 0.635 | 0.789 |
| AAI | 0.414 | 0.707 | 0.720 | 0.694 | 0.788 |
| H20 | 0.414 | 0.706 | 0.755 | 0.657 | 0.779 |
| BPFC4 | 0.409 | 0.704 | 0.736 | 0.673 | 0.778 |
| OVPC4 | 0.411 | 0.703 | 0.619 | 0.786 | 0.774 |
| H19 | 0.407 | 0.703 | 0.753 | 0.653 | 0.778 |
| H4 | 0.405 | 0.702 | 0.748 | 0.655 | 0.778 |
| H16 | 0.405 | 0.702 | 0.753 | 0.651 | 0.779 |
| H18 | 0.402 | 0.700 | 0.750 | 0.650 | 0.777 |
| H15 | 0.401 | 0.699 | 0.749 | 0.650 | 0.777 |
| H13 | 0.397 | 0.697 | 0.749 | 0.645 | 0.777 |
| OVPN5 | 0.372 | 0.686 | 0.671 | 0.700 | 0.758 |
| OVPC3 | 0.359 | 0.680 | 0.665 | 0.694 | 0.743 |
| H3 | 0.359 | 0.679 | 0.711 | 0.647 | 0.744 |
| TOBNC5 | 0.352 | 0.675 | 0.634 | 0.716 | 0.725 |
| BPFC3 | 0.349 | 0.675 | 0.667 | 0.682 | 0.744 |
| OVPN4 | 0.347 | 0.674 | 0.688 | 0.659 | 0.722 |
| BPFN4 | 0.340 | 0.669 | 0.715 | 0.623 | 0.723 |
| TOBNC4 | 0.337 | 0.668 | 0.633 | 0.703 | 0.719 |
| BPFN5 | 0.337 | 0.667 | 0.737 | 0.597 | 0.757 |
| TOBN5 | 0.326 | 0.662 | 0.711 | 0.613 | 0.710 |
| OVPN3 | 0.308 | 0.654 | 0.661 | 0.647 | 0.707 |
| TOBC5 | 0.300 | 0.650 | 0.646 | 0.654 | 0.701 |
| TOBN4 | 0.290 | 0.645 | 0.678 | 0.611 | 0.690 |
| TOBNC3 | 0.288 | 0.644 | 0.658 | 0.630 | 0.699 |
| TOBN3 | 0.285 | 0.642 | 0.671 | 0.613 | 0.675 |
| TOBC4 | 0.276 | 0.638 | 0.664 | 0.612 | 0.686 |
| TOBC3 | 0.270 | 0.635 | 0.616 | 0.654 | 0.695 |
| BPFN3 | 0.255 | 0.625 | 0.730 | 0.520 | 0.691 |

**Table S11.** List of feature importance score (FIS) by RF method

|  |  |  |
| --- | --- | --- |
| **Rank** | **Feature identifier** | **FIS** |
| 1 | F28 | 0.763 |
| 2 | F47 | 0.757 |
| 3 | F39 | 0.577 |
| 4 | F18 | 0.525 |
| 5 | F15 | 0.487 |
| 6 | F41 | 0.380 |
| 7 | F3 | 0.317 |
| 8 | F14 | 0.280 |
| 9 | F10 | 0.269 |
| 10 | F7 | 0.263 |
| 11 | F4 | 0.248 |
| 12 | F49 | 0.240 |
| 13 | F6 | 0.227 |
| 14 | F13 | 0.185 |
| 15 | F0 | 0.157 |
| 16 | F36 | 0.155 |
| 17 | F31 | 0.155 |
| 18 | F44 | 0.155 |
| 19 | F9 | 0.152 |
| 20 | F38 | 0.152 |
| 21 | F25 | 0.149 |
| 22 | F23 | 0.146 |
| 23 | F11 | 0.144 |
| 24 | F34 | 0.141 |
| 25 | F40 | 0.132 |
| 26 | F43 | 0.129 |
| 27 | F45 | 0.126 |
| 28 | F26 | 0.125 |
| 29 | F50 | 0.122 |
| 30 | F30 | 0.121 |
| 31 | F27 | 0.118 |
| 32 | F5 | 0.116 |
| 33 | F24 | 0.116 |
| 34 | F19 | 0.115 |
| 35 | F46 | 0.114 |
| 36 | F33 | 0.114 |
| 37 | F48 | 0.113 |
| 38 | F42 | 0.113 |
| 39 | F1 | 0.111 |
| 40 | F20 | 0.108 |
| 41 | F2 | 0.107 |
| 42 | F21 | 0.107 |
| 43 | F16 | 0.105 |
| 44 | F37 | 0.105 |
| 45 | F22 | 0.103 |
| 46 | F8 | 0.098 |
| 47 | F17 | 0.094 |
| 48 | F35 | 0.092 |
| 49 | F12 | 0.092 |
| 50 | F32 | 0.092 |
| 51 | F29 | 0.090 |

**References**

Breiman, L. Random forests. *Machine learning* 2001;45(1):5-32.

Cao, R.*, et al.* QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* 2017;33(4):586-588.

Cao, R.*, et al.* DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics* 2016;17(1):495.

Cao, R.*, et al.* ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* 2017;22(10).

Cao, R., Wang, Z. and Cheng, J. Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. *BMC Struct Biol* 2014;14:13.

Cao, R.*, et al.* SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *BMC Bioinformatics* 2014;15:120.

Chen, J.*, et al.* ProtDec-LTR2.0: an improved method for protein remote homology detection by combining pseudo protein and supervised Learning to Rank. *Bioinformatics* 2017;33(21):3473-3476.

Chen, J., Wang, X. and Liu, B. iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Sci Rep* 2016;6:19062.

Chen, W.*, et al.* iRNA-3typeA: Identifying Three Types of Modification at RNA's Adenosine Sites. *Mol Ther Nucleic Acids* 2018;11:468-474.

Chen, W.*, et al.* iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* 2013;41(6):e68.

Chen, W.*, et al.* iRNA-PseU: Identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids* 2016;5:e332.

Chen, W.*, et al.* iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 2017;33(22):3518-3523.

Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology* 2011;273(1):236-247.

Chou, K.-C. and Shen, H.-B. Recent advances in developing web-servers for predicting protein attributes. *Natural Science* 2009;1(02):63.

Feng, P.*, et al.* iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 2018.

Feng, P.M.*, et al.* iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal Biochem* 2013;442(1):118-125.

Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 2001:1189-1232.

Geurts, P., Ernst, D. and Wehenkel, L. Extremely randomized trees. *Machine learning* 2006;63(1):3-42.

Han, S.*, et al.* LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief Bioinform* 2018.

Kosylo, N.*, et al.* Artificial Intelligence on Job-Hopping Forecasting: AI on Job-Hopping. In, *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE; 2018. p. 1-5.

Lai, H.Y.*, et al.* Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* 2017;8(17):28169-28175.

Lin, H.*, et al.* The prediction of protein structural class using averaged chemical shifts. *J Biomol Struct Dyn* 2012;29(6):643-649.

Lin, H.*, et al.* Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans Comput Biol Bioinform* 2017.

Liu, B.*, et al.* iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 2018.

Liu, B.*, et al.* iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. *Bioinformatics* 2018;34(18):3086-3093.

Liu, B.*, et al.* iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 2018;34(1):33-40.

Liu, Y., Wang, X. and Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics* 2017.

Manavalan, B.*, et al.* MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 2017;8(44):77121-77136.

Manavalan, B.*, et al.* iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. *Front Immunol* 2018;9:1695.

Manavalan, B.*, et al.* iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. *Frontiers in Immunology* 2018;9:1695.

Manavalan, B. and Lee, J. SVMQA: support–vector-machine-based protein single-model quality assessment. *Bioinformatics* 2017;33(16):2496-2503.

Manavalan, B., Lee, J. and Lee, J. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PloS one* 2014;9(9):e106542.

Manavalan, B.*, et al.* PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. *Front Immunol* 2018;9:1783.

Manavalan, B., Shin, T.H. and Lee, G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 2018;9(2):1944-1956.

Manavalan, B.*, et al.* Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J Proteome Res* 2018;17(8):2715-2726.

Manavalan, B.*, et al.* Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J Proteome Res* 2018.

McDermaid, A.*, et al.* A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation. *Front Genet* 2018;9:313.

Noble, W.S. What is a support vector machine? *Nature biotechnology* 2006;24(12):1565.

Rawi, R.*, et al.* PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics* 2018;34(7):1092-1098.

Song, J.*, et al.* PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci Rep* 2017;7(1):6862.

Tang, H.*, et al.* A two-step discriminated method to identify thermophilic proteins. *International Journal of Biomathematics* 2017;10(04):1750050.

Tang, H.*, et al.* HBPred: a tool to identify growth hormone-binding proteins. *International Journal of Biological Sciences* 2018;14(8):957-964.

Vens, C., Rosso, M.N. and Danchin, E.G. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* 2011;27(9):1231-1238.

Wei, L., Chen, H. and Su, R. M6APred-EL: A Sequence-Based Predictor for Identifying N6-methyladenosine Sites Using Ensemble Learning. *Mol Ther Nucleic Acids* 2018;12:635-644.

Wei, L.*, et al.* CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. *J Proteome Res* 2017;16(5):2044-2053.

Wei, L.*, et al.* ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 2018.

Yang, H.*, et al.* Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. *Biomed Res Int* 2016;2016:5413903.

Yang, J., Cao, R. and Si, D. EMNets: A Convolutional Autoencoder for Protein Surface Retrieval Based on Cryo-Electron Microscopy Imaging. In, *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM; 2018. p. 639-644.

Zhao, Y.W.*, et al.* IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types. *Int J Mol Sci* 2017;18(9).