**Supplementary materials:**

**DNNAce：prediction of prokaryote lysine acetylation sites based on the deep neural networks by fusing multiple feature information**

Bin Yu1,2,3,4,\*,†, Zhaomin Yu1,2,†, Cheng Chen1,2, Anjun Ma5, Bingqiang Liu 6, Baoguang Tian 1,2, Qin Ma 5,\*

1College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

2Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

3School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha 410114, China

4School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

5Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, Ohio 43210, USA

6School of Mathematics, Shandong University, Jinan 250100, China

\*To whom correspondence should be addressed.

† Contributed equally to this work.

**Contact:** yubin@qust.edu.cn or qin.ma@osumc.edu.

**Table of contents**

**1. Supplementary Experimental Text**

**Text S1.** Datasets construction.

**Text S2.** Sequence-based features.

**Text S3.** Physicochemical property-based features.

**Text S4.** Evolutionary-derived features.

**2. Supplementary Tables**

**Table S1.** Parameters range and settings of the DNN.

**Table S2.** ACC values corresponding to different  in PseAAC.

**Table S3.** ACC values corresponding to different  in NMBroto.

**Table S4.** ACC values corresponding to different  in EBGW.

**Table S5**. Dimensions corresponding to different feature extraction methods of nine datasets (All represents the initial feature space corresponding to the fusion of eight features).

**Table S6**. Group Lasso different parameters correspond to dimensions and ACC values.

**Table S7.** Comparison of dimensions and ACC values of seven-dimensional reduction methods.

**Table S8.** Comparison of ACC values for different classification methods.

**Table S9.** Comparison of DNNAce and ProAcePred predictions for training datasets and independent test datasets.

**3. Supplementary Figure**

**Fig. S1.** EBGW comparison between acetylation and non-acetylation. The vertical axis represents the log2 ratio of the average EBGW values between acetylation and non-acetylation. The horizontal axis , , and  represent three binary sequences, respectively.

**Fig. S2.** ROC curves of different dimensional reduction methods on nine datasets.

**4. Supplementary References**

**1. Supplementary Experimental Text**

**Text S1.** Datasets construction.

By removing the mistaken sequences with modification sites among the nine datasets, Chen *et al.* collected 5,316 experimentally validated acetylated proteins, containing 8,787 lysine acetylation sites and 87,585 lysine non-acetylation sites. Then the protein sequence was clustered using CD-HIT ([Li and Godzik, 2006](#_ENREF_8)) with a 30% homology threshold to obtain 7,288 lysines acetylated fragments and 41,638 lysine non-acetylated fragments. Afterward, they randomly selected 10% non-homologous lysine acetylated fragments and non-acetylated fragments as independent test datasets for nine species. Finally, because of the number of negative samples more than positive samples for training dataset and independent test dataset. Therefore, negative samples are randomly selected in the negative datasets, for achieving the balance of positive and negative samples. Chen *et al.* determined that the window size of the *Archaea* dataset is 13 (-6~6), the window size of the V. *parahaemolyticus* dataset is 17 (-8~8), and the sample window size of the remaining seven datasets is 21 (-10~10). When the length of the positive and negative samples is insufficient, the virtual amino acid O is defined to achieve the desired window size.

**Text S2.** Sequence-based features.

**BE:** Binary encoding converts each amino acid residue in the protein sample sequence into a 21-dimension numerical vector, which is a feature extraction method based on sequence information. The 20 common amino acids are encoded in the order of 'ACDEFGHIKLMNPQRSTVWY' ([Ju and He, 2017](#_ENREF_6)). For example, aspartic acid D is represented by , tryptophan W is encoded as the vector , and for a virtual amino acid O, which is represented by . Therefore, for a sample with a sequence window length of , the binary encoding dimension is .

**PseAAC:** Proteins are composed of a series of amino acid residues at different positions. To avoid losing the position and sequence relationship in the amino acid sequence, Chou et al. ([Chou, 2001](#_ENREF_2)) proposed pseudo-amino acid composition to extract the feature vectors in the protein sequence, Sequence  is defined the  dimensional pseudo amino acid composition as follows:

 (1)

Each component in the vector  is defined as follows:

 (2)

where  is the weighting factor, set to 0.05，  is  sequence correlation factor, and  is the -th amino acid occurrence frequency in the protein sequence ( ). According to formula (2), the first 20 dimensions of the feature vector  represent the amino acid composition, and the latter  dimensions reflect the sequence correlation factors of different levels in the amino acid sequence information. The sequence correlation factor is obtained by the physicochemical properties of the amino acids. By setting different parameter values, the optimal value  can be determined from the prediction accuracy.

**Text S3.** Physicochemical property-based features.

**AAindex:** The physical and chemical properties of amino acids represent the most intuitive features of biochemical reactions and have been widely used in bioinformatics research. The AAindex database ([Kawashima, et al., 2008](#_ENREF_7)) collected 544 amino acid indices with a set of values for each amino acid index. If we select the full amino acid index that not only does not adequately reflect the physicochemical properties around the acetylation site but also produces redundancy and noise information. So 12 types of physicochemical properties are selected for reference ([Hasan, et al., 2017](#_ENREF_5)). By utilizing these 12 physicochemical properties, acetylated fragments and non-acetylated fragments are converted into numerical signals, and  dimensional vectors are generated for samples of  window size.

**NMBroto:** Autocorrelation descriptors (AD) are defined by amino acids with different physicochemical properties. In this paper, eight amino acid indexes are selected from the AAindex database ([Kawashima, et al., 2008](#_ENREF_7)), and normalized Moreau-Broto autocorrelation (NMBroto) ([Chen, et al., 2018](#_ENREF_1)) is utilized to transform protein residues sequence into the numerical signals.

For a given protein residue sequence  of length , the values of the eight physicochemical properties corresponding to the 20 common amino acids are normalized by equation (3):

 (3)

where  represents the average of the -th physicochemical properties and  represents the standard deviation of the-th physicochemical properties.

Normalized Moreau-Broto autocorrelation is defined as follows:

 (4)

where , and  represent the normalization physicochemical values of the amino acid at positions  and , respectively.  represents the lag interval of the autocorrelation, and the NMBroto can be used to extract the -dimensional vector of the protein sequence.

**EBGW:** Zhang et al. ([Zhang, et al., 2006](#_ENREF_9)) based on the physicochemical properties of amino acid residues, proposed encoding based on grouped weight (EBGW) to extract the protein sequence. Considering the hydrophobicity and charged character, 20 amino acid residues are divided into four different classes as follows:

neutral and non-polarity residue 

neutral and polarity residue 

acidic residue 

basic residue 

Combining the above four divisions then three combinations are obtained, each of which divides 20 amino acid residues into disjoint parts,  vs. ， vs. ， vs. For a protein sequence , it will be converted into three binary sequences as follows:

 (5)

Each binary sequence is divided into  sub-sequences whose length is increased in order. For example, for, the -th sub-sequence is expressed as , where  represents the number of the number 1 in the sub-sequence of ,  represents the length of the -th subsequence, and  represents the length of the protein sequence. In summary, for a protein sequence  of length , -dimensional vector  can be obtained.

**MMI:** Ding et al. ([Ding, et al., 2017](#_ENREF_3)) proposed the multivariate mutual information (MMI) algorithm to adequately represent the amino acids in the sequence. Based on the dipoles and volume of the side chain of the amino acid residues, the 20 amino acids can be divided into clustered into seven functional groups (Table 1). Any three contiguous amino acids are regarded as a unit, Ding et al. only think about the basic ingredient of the unit and don’t consider the order of three amino acids. The type of 3-gram can be represented by ， …, , the type of 2-gram is represented by ，，…, . The number of 3-gram and 2-gram are counted by the sliding window as shown in Figure 1.

**Table 1.** The division of 20 amino acid types based on side chain polarity and volume.

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Group | Dipole scale | Volume scale |
| *C*0 | *A, G, V* | Dipole  1.0 | Volume  50 |
| *C*1 | *C* | 1.0  Dipole 2.0 (form disulphide bonds) | Volume 50 |
| *C*2 | *D, E* | Dipole  3.0 (opposite orientation) | Volume 50 |
| *C*3 | *F, I, L, P* | Dipole  1.0 | Volume  50 |
| *C*4 | *H, N, Q, W* | 2.0 dipole  3.0 | Volume 50 |
| *C*5 | *K, R* | Dipole 3.0 | Volume 50 |
| *C*6 | *M, S, T, Y* | 1.0 dipole  2.0 | Volume 50 |

 **Figure 1** Characteristic representation of 3-gram and 2-gram.

The entropy and mutual information (MI) refer to the mutual dependence between two amino acids on the sequence. The 3-tuple MI for 3-gram is defined as follows:

 (6)

where ， and  are three conjoint amino acids in a unit, the MI of  and the conditional MI of  are defined as:

 (7)

 (8)

where  is the frequency of  and  arising in 2-gram on one sequence, and  is the frequency of  in the sequence. and  are calculated as follows:

 (9)

 (10)

where  denotes the frequency of ， and  appearing in 3-tuples on one amino acid sequence.

To avoid the infinity of the 3-tuple and 2-tuple mutual information values, the frequency is defined as:

 (11)

where  represents the length of the sequence and  denotes the number of occurrence of category  appearing on the amino acid sequence.  and  are also calculated by a similar method.

The MI of and of the 3-tuple (84-dimensional) and 2-tuple (28-dimensional) is extracted from the amino acid sequence, respectively. A 119-dimension vector is finally generated by calculating the frequency of each category appearing on the amino acid sequence.

**Text S4.** Evolutionary-derived features.

**BLOSUM62:** The BLOSUM62 matrix is used to represent the primary sequence information of the protein. Each residue in the training dataset is represented by the matrix containing  elements, where  is the length and  is the 20 amino acids. Each row of the normalized BLOSUM62 matrix represents one of the 20 common amino acids. The BLOSUM62 descriptor ([Chen, et al., 2018](#_ENREF_1)) can be used to encode peptides of equal length.

**KNN:** The KNN algorithm ([Gao, et al., 2010](#_ENREF_4)) predicts PTM sites by clustering information of local sequences, extracting features from similar sequences of positive and negative datasets to capture local sequence similarity around post-translational modification sites. Specifically, for the two sequence fragments and , the distance  between the sequences  and  is defined as follows:

 (12)

 (13)

represents the protein sequence window size,  is the normalized amino acid substitution matrix. is the substitution matrix, derived from the BLOSUM62 matrix, and  represent two amino acids, and represents the maximum/minimum in the substitution matrix , respectively.

For the query sequence , the corresponding KNN score is calculated in the following three steps. First, calculate the distance between the query sequence  and all comparison datasets that contain the same number of positive and negative datasets. Second, sort by distance and select  nearest neighbors. Finally, the percentage of positive neighbors (samples containing acetylation sites) among the nearest neighbors is calculated as KNN score.

The above steps are repeated for different  values to obtain the acetylation predictor multiple features. In this paper, considering that the dataset E. contains 190 samples, for nine different acetylation site datasets,  is set to 2, 4, 8, 16, 32, 64, 128 in order. So for each protein sequence segment, the KNN encoding produces a 7-dimensional vector.

**2. Supplementary Tables**

**Table S1.** Parameters range and settings of the DNN.

|  |  |  |
| --- | --- | --- |
| Name | Range | Setting |
| Learning rate | 1, 0.1, 0.01, 0.001, 0.0001 | 0.01 |
| Weight initialization | uniform, normal, glorot\_normal, glorot\_uniform,lecun\_uniform,  he\_normal, he\_uniform | glorot\_normal |
| Per-parameter adaptive learning rate methods | SGD, RMSprop, Adagrad, Adadelta,  Adam, Adamax, Nadam | Adam |
| Activation function | relu, tanh, sigmoid, softmax, softplus,  softsign, hard\_sigmoid | relu |
| Dropout rate | 0.1, 0.2, 0.5, 0.8 | 0.5 |

**Table S2.** ACC values corresponding to different  in PseAAC.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACC (%) | A. | B. | C. | E. | E. *coil* | G. | M. | S. | V. |
|  | **58.57** | 60.62 | 59.06 | 60.83 | **58.10** | 56.89 | 60.79 | 55.25 | 64.89 |
|  | 54.68 | 60.77 | 59.31 | 60.33 | 57.19 | 52.37 | 61.60 | 57.19 | **66.07** |
|  | 55.70 | **60.96** | 58.97 | 55.67 | 57.56 | 54.53 | 62.41 | **57.48** | 65.55 |
|  | 57.03 | 60.87 | 60.83 | 56.67 | 56.49 | **57.44** | **63.28** | 53.73 | 65.64 |
|  | 58.30 | 59.28 | **60.92** | **62.11** | 56.75 | 54.17 | 61.14 | 55.42 | 65.18 |
|  | 52.59 | 60.24 | 60.63 | 58.67 | 58.03 | 54.72 | 61.83 | 53.46 | 65.41 |

Since the samples of the training dataset contain the virtual amino acid O, the value of  is set 1, 2, 3, 4, 5, and 6, respectively, and the corresponding accuracy values of the nine prokaryotic datasets under different values  are shown in Table S2. It can be seen from Table S2 that for different datasets, the corresponding value is inconsistent when the accuracy rate is the highest. For example, for datasets A. and E. *coil*, when  is 1, the accuracy ACC reaches the highest value of 58.57% and 58.10%, respectively. For datasets C. and E., when the  value is 5, the accuracy ACC reaches the maximum value. Considering comprehensively, we analyze the overall prediction accuracy of the nine datasets under different  values. The overall prediction accuracy reaches a peak value at an interval value of . Therefore, when we use the PseAAC algorithm for extracting protein sequences feature, the optimal parameter  is determined to be 1, and each protein sequence generates a 21-dimensional feature vector.

**Table S3.** ACC values corresponding to different  in NMBroto.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACC (%) | A. | B. | C. | E. | E. *coil* | G. | M. | S. | V. |
|  | 55.22 | 54.33 | 51.52 | 45.61 | 51.09 | 46.84 | 51.44 | 52.63 | 51.22 |
|  | 57.63 | 54.23 | 52.60 | **52.61** | 51.43 | **50.51** | 52.71 | 55.03 | 53.16 |
|  | 58.66 | **56.01** | 53.33 | 49.67 | 50.42 | 46.55 | **54.50** | **55.23** | **54.75** |
|  | 61.74 | 54.66 | 55.93 | 51.39 | 51.38 | 46.02 | 53.23 | 54.12 | 54.51 |
|  | 60.46 | 54.62 | **56.08** | 51.39 | **51.62** | 46.61 | 52.53 | 51.75 | 53.95 |
|  | **65.96** | 53.75 | 54.36 | 49.22 | 50.08 | 43.99 | 54.16 | 52.94 | 54.65 |

NMBroto is defined by the physicochemical properties of the amino acids in the sequence. Since the sequence samples in the dataset with virtual amino acids, the interval  in NMBroto is set to 1, 2, 3, 4, 5, and 6, respectively. The ACC values of different  on nine acetylation sites datasets are shown in Table S3. As can be seen from Table S3, for different datasets, the corresponding NMBroto interval is inconsistent when the accuracy reaches the highest. For example, for dataset A., when the  is 6, the accuracy reaches the highest value of 65.96%; and for the dataset E. *coil*, when the  is 5, the highest prediction accuracy 51.62% is reached. We comprehensively consider the influence of different interval values on the prediction results. By analyzing the prediction accuracy of the nine datasets in NMBroto with different , when the  is 4, the overall prediction accuracy of the model is the highest. Considering that the optimal parameter of the NMBroto is set to 4 when we use NMBroto to extract the feature of the protein sequence, each protein sequence generates a 32-dimensional feature vector.

**Table S4.** ACC values corresponding to different  in EBGW.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACC (%) | A. | B. | C. | E. | E. *coil* | G. | M. | S. | V. |
|  | 51.05 | 53.08 | 50.78 | 52.72 | 50.65 | 51.32 | 53.07 | 52.06 | 52.62 |
|  | 52.11 | 58.75 | 56.32 | 53.39 | 54.30 | 54.23 | 60.10 | 53.17 | 61.64 |
|  | 53.66 | 61.01 | 58.42 | 59.39 | 58.23 | 52.11 | 62.07 | 52.27 | 63.77 |
|  | 51.03 | **62.26** | 59.16 | 60.39 | 57.04 | 49.49 | 62.82 | 55.41 | 64.09 |
|  | 57.05 | 62.12 | 58.57 | 60.22 | 56.57 | 53.70 | 62.82 | 56.91 | 65.83 |
|  | 56.18 | 60.53 | 58.47 | 61.78 | 58.29 | 54.80 | 62.99 | 58.45 | 65.23 |
|  | 58.36 | 62.07 | 58.87 | 62.56 | 57.58 | **56.64** | **64.09** | 58.97 | **67.14** |
|  | 56.21 | 60.29 | 60.09 | 64.17 | 57.92 | 53.77 | 61.95 | 58.38 | 66.21 |
|  | 58.07 | 61.25 | 60.43 | **65.72** | **59.25** | 55.32 | 62.94 | 56.18 | 64.38 |
|  | 57.01 | 60.34 | **60.73** | 63.11 | 56.72 | 55.85 | 63.04 | 56.73 | 66.16 |
|  | 57.50 | 61.35 | 59.01 | 63.50 | 58.10 | 54.01 | 63.34 | **61.24** | 66.21 |
|  | 60.33 | 61.30 | 59.26 | 64.61 | 57.61 | 52.60 | 63.05 | 60.15 | 65.98 |
|  | **63.74** | 60.72 | 59.55 | 64.50 | 57.77 | 53.23 | 63.51 | 58.68 | 58.68 |

For the dataset A., the sample window size is 13, so the number of sub-sequences in the EBGW is set to 1, 2, 3, ... , 12, respectively. The ACC values corresponding to the different subsequence numbers for the nine datasets are shown in Table S4. It is known from Table S4 that for the prediction of nine prokaryote acetylation sites datasets, the corresponding subsequence number of EBGW is inconsistent when the accuracy rate is the highest. For example, for dataset A., when the  is 13, the accuracy reaches the highest value. For dataset G., dataset M., and dataset V., when the  is 7, the ACC reached the maximum, 56.64%, 64.09%, and 67.14%, respectively. Overall, by analyzing the overall prediction accuracy under the different subsequence number of 9 datasets, we find when the number of the subsequence is 7, the overall prediction accuracy of the nine datasets achieve the maximum ACC value. Therefore, when the ENGW algorithm is selected to extract the feature of the protein sequence, a  dimension feature vector is obtained for each protein sequence.

**Table S5.** Dimensions corresponding to different feature extraction methods of nine datasets (All represents the initial feature space corresponding to the fusion of eight features).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dimensions | A. | B. | C. | E. | E. *coil* | G. | M. | S. | V. |
| BE | 273 | 441 | 441 | 441 | 441 | 441 | 441 | 441 | 357 |
| PseAAC | 21 | 21 | 21 | 21 | 21 | 21 | 21 | 21 | 21 |
| AAindex | 156 | 252 | 252 | 252 | 252 | 252 | 252 | 252 | 204 |
| NMBroto | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 | 32 |
| EBGW | 21 | 21 | 21 | 21 | 21 | 21 | 21 | 21 | 21 |
| MMI | 119 | 119 | 119 | 119 | 119 | 119 | 119 | 119 | 119 |
| BLOSUM62 | 260 | 420 | 420 | 420 | 420 | 420 | 420 | 420 | 340 |
| KNN | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| All | 889 | 1313 | 1313 | 1313 | 1313 | 1313 | 1313 | 1313 | 1101 |

**Table S6.** Group Lasso different parameters correspond to dimensions and ACC values.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Parameters | Indexes | A. | B. | C. | E. | E.coil | G. | M. | S. | V. |
| 0.01 | dim | 237 | 384 | 397 | 184 | 364 | 289 | 395 | 278 | 316 |
| ACC | 78.26 | 73.37 | 73.81 | 93.61 | 60.53 | 85.48 | 74.94 | 87.63 | 73.63 |
| 0.02 | dim | 188 | 307 | 323 | 173 | 279 | 238 | 329 | 229 | 259 |
| ACC | 83.17 | 73.37 | 75.03 | 94.11 | 62.01 | 84.87 | 74.61 | 88.40 | 74.38 |
| 0.03 | dim | 178 | 233 | 242 | 170 | 205 | 209 | 231 | 205 | 197 |
| ACC | **84.47** | 73.89 | 75.38 | **96.89** | **63.08** | 89.15 | **76.62** | 90.51 | **75.46** |
| 0.04 | dim | 151 | 173 | 180 | 126 | 136 | 203 | 187 | 189 | 167 |
| ACC | 82.58 | **74.76** | 75.28 | 95.78 | 60.68 | 87.82 | 76.28 | 88.19 | **75.46** |
| 0.05 | dim | 150 | 128 | 139 | 144 | 89 | 197 | 138 | 187 | 117 |
| ACC | 80.79 | 73.03 | **75.42** | 96.33 | 60.14 | 88.35 | 76.34 | 87.35 | 73.91 |
| 0.06 | dim | 127 | 103 | 103 | 133 | 57 | 176 | 104 | 182 | 90 |
| ACC | 81.86 | 73.75 | 73.46 | 95.39 | 61.52 | 88.36 | 75.76 | **91.34** | 73.95 |
| 0.07 | dim | 115 | 79 | 78 | 124 | 46 | 157 | 79 | 172 | 64 |
| ACC | 82.12 | 73.51 | 72.93 | 95.17 | 61.96 | 89.14 | 75.06 | 87.57 | 74.00 |
| 0.08 | dim | 105 | 54 | 62 | 123 | 27 | 133 | 58 | 150 | 50 |
| ACC | 83.14 | 73.51 | 73.85 | 94.22 | 62.56 | 88.10 | 74.89 | 88.74 | 74.71 |
| 0.09 | dim | 99 | 44 | 49 | 119 | 23 | 125 | 47 | 142 | 42 |
| ACC | 82.16 | 73.46 | 74.10 | 96.33 | 62.53 | 89.17 | 76.22 | 86.72 | 73.86 |
| 0.1 | dim | 83 | 38 | 43 | 108 | 18 | 109 | 34 | 131 | 33 |
| ACC | 80.59 | 73.37 | 73.71 | 92.56 | 61.88 | **90.23** | 74.84 | 88.46 | 74.01 |

*Note*: dim denotes dimension.

As can be seen from Table S6, When the value of  is set too large, more "sparse" solutions are generated, and the variables with lower correlation are removed. When the value of  is small, all feature attributes will be selected. As  increases, the corresponding dimension of each dataset decreases, and the optimal parameters in Group Lasso are selected according to the prediction accuracy. The ACC values of the datasets A., E., E. *coil*, and M. reach the maximum when the parameter value is 0.03, which are 84.47%, 96.89%, 63.08%, and 76.62%, respectively. When the parameter is 0.04, the ACC value of the dataset B. reaches the maximum, and the ACC value of the dataset V. reaches the maximum at the same time when the parameters are 0.03 and 0.04. The peak ACC values of datasets C., G., and S. when parameters are 0.05, 0.1, and 0.06, respectively. Therefore, 0.03 is the optimal parameter of the Group Lasso algorithm.

**Table S7.** Comparison of dimensions and ACC values of seven-dimensional reduction methods.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dimensional  reduction  methods | Indexes | A. | B. | C. | E. | E. *coil* | G. | M. | S. | V. |
| SVD | dim | 164 | 174 | 186 | 158 | 134 | 210 | 183 | 217 | 167 |
| ACC | 62.13 | 65.48 | 65.38 | 58.67 | 56.44 | 62.70 | 66.81 | 63.25 | 66.54 |
| MI | dim | 178 | 233 | 242 | 170 | 205 | 209 | 231 | 205 | 197 |
| ACC | 68.71 | 70.10 | 68.72 | 66.72 | 57.11 | 60.32 | 68.65 | 58.42 | 68.79 |
| IG | dim | 178 | 233 | 242 | 170 | 205 | 209 | 231 | 205 | 197 |
| ACC | 69.78 | 70.62 | 70.38 | 67.61 | 59.22 | 70.37 | 70.91 | 60.92 | 69.25 |
| ET | dim | 395 | 569 | 591 | 548 | 796 | 587 | 565 | 616 | 475 |
| ACC | 73.41 | 71.11 | 72.73 | 72.94 | 59.38 | 68.04 | 70.21 | 65.60 | 70.81 |
| Elastic net | dim | 379 | 544 | 548 | 295 | 504 | 469 | 539 | 448 | 441 |
| ACC | 69.22 | 70.43 | 71.36 | 95.28 | 60.68 | 70.66 | 71.07 | 74.36 | 72.31 |
| LR | dim | 195 | 507 | 537 | 129 | 541 | 226 | 510 | 212 | 451 |
| ACC | 79.05 | 70.14 | 71.02 | 94.78 | 60.76 | 85.20 | 71.94 | 87.89 | 71.28 |
| Group Lasso | dim | 178 | 233 | 242 | 170 | 205 | 209 | 231 | 205 | 197 |
| ACC | **84.47** | **73.89** | **75.38** | **96.89** | **63.08** | **89.15** | **76.62** | **90.51** | **75.46** |

*Note*: dim denotes dimension.

**Table S8.** Comparison of ACC values for different classification methods.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACC (%) | A. | B. | C. | E. | E. *coil* | G. | M. | S. | V. |
| AdaBoost | 61.21 | 69.38 | 69.75 | 71.28 | 59.67 | 61.92 | 69.24 | 60.77 | 70.76 |
| NB | 64.47 | 72.21 | 70.19 | 69.50 | 57.87 | 61.40 | 71.94 | 69.18 | 58.21 |
| XGBoost | 67.39 | 71.88 | 73.71 | 69.28 | 60.92 | 60.58 | 72.98 | 62.99 | 73.35 |
| KNN | 70.03 | 68.51 | 66.56 | 85.39 | 57.38 | 69.85 | 66.75 | 65.31 | 67.23 |
| RF | 70.79 | 71.88 | 72.04 | 74.11 | 62.01 | 66.71 | 73.80 | 66.24 | 74.66 |
| SVM | 76.74 | 72.93 | 74.44 | 86.78 | **63.37** | 73.35 | 74.03 | 73.73 | 73.34 |
| DNN | **84.47** | **73.89** | **75.38** | **96.89** | 63.08 | **89.15** | **76.62** | **90.51** | **75.46** |

**Table S9.** Comparison of DNNAce and ProAcePred predictions for training datasets and independent test datasets.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Datasets | Models | ACC(%) | Sn(%) | Sp(%) | MCC | AUC | AUPR |
|  | A. | ProAcePred | 90.00 | 90.00 | 90.00 | 0.8030 | 0.8970 | N/A |
|  | DNNAce | 84.47 | 82.92 | 86.03 | 0.6938 | 0.9098 | 0.8942 |
|  | B. | ProAcePred | 79.60 | 79.90 | 79.20 | 0.5920 | 0.8000 | N/A |
|  | DNNAce | 73.89 | 74.81 | 72.98 | 0.4799 | 0.8163 | 0.8049 |
|  | C. | ProAcePred | 80.00 | 80.70 | 79.40 | 0.6020 | 0.8070 | N/A |
|  | DNNAce | 75.38 | 76.90 | 73.86 | 0.5087 | 0.8228 | 0.8071 |
|  | E. | ProAcePred | 98.30 | 97.80 | 98.90 | 0.9670 | 0.9880 | N/A |
|  | DNNAce | 96.89 | 96.89 | 96.89 | 0.9398 | 0.9968 | 0.9949 |
| training datasets | E. *coil* | ProAcePred | 69.00 | 69.10 | 68.90 | 0.3810 | 0.6870 | N/A |
| DNNAce | 63.08 | 64.26 | 61.91 | 0.2621 | 0.6747 | 0.6617 |
|  | G. | ProAcePred | 89.70 | 91.70 | 87.80 | 0.8010 | 0.9020 | N/A |
|  | DNNAce | 89.15 | 90.47 | 87.84 | 0.7858 | 0.9624 | 0.9573 |
|  | M. | ProAcePred | 83.40 | 83.00 | 83.80 | 0.6710 | 0.8300 | N/A |
|  | DNNAce | 76.62 | 76.33 | 76.92 | 0.5334 | 0.8409 | 0.8205 |
|  | S. | ProAcePred | 81.60 | 89.50 | 73.70 | 0.6400 | 0.7780 | N/A |
|  | DNNAce | 90.51 | 90.23 | 90.78 | 0.8163 | 0.9623 | 0.9551 |
|  | V. | ProAcePred | 80.20 | 0.81 | 0.80 | 0.6050 | 0.7990 | N/A |
|  | DNNAce | 75.46 | 74.57 | 76.34 | 0.5101 | 0.8315 | 0.8224 |
|  | A. | ProAcePred | 81.00 | 81.00 | 81.00 | 0.6190 | 0.8550 | N/A |
|  | DNNAce | 90.00 | 85.00 | 95.00 | 0.8309 | 0.9750 | 0.9825 |
|  | B. | ProAcePred | 95.20 | 95.70 | 94.80 | 0.9040 | 0.9420 | N/A |
|  | DNNAce | 98.26 | 98.26 | 98.26 | 0.9665 | 1.0000 | 0.9991 |
|  | C. | ProAcePred | 87.20 | 85.00 | 89.40 | 0.7440 | 0.8610 | N/A |
|  | DNNAce | 92.88 | 93.79 | 91.97 | 0.8657 | 0.9742 | 0.9790 |
|  | E. | ProAcePred | 90.00 | 80.00 | 100.00 | 0.8160 | 0.8960 | N/A |
|  | DNNAce | 90.00 | 100.00 | 80.00 | 0.8000 | 0.9000 | 0.9667 |
| Independent  test datasets | E. *coil* | ProAcePred | 89.90 | 88.70 | 91.10 | 0.7980 | 0.8870 | N/A |
| DNNAce | 86.18 | 85.95 | 86.41 | 0.7259 | 0.9373 | 0.9308 |
|  | G. | ProAcePred | 88.10 | 95.20 | 81.00 | 0.7700 | 0.8540 | N/A |
|  | DNNAce | 97.50 | 95.00 | 100.00 | 0.9577 | 1.0000 | 0.9409 |
|  | M. | ProAcePred | 88.00 | 88.50 | 87.50 | 0.7600 | 0.8630 | N/A |
|  | DNNAce | 96.44 | 96.89 | 96.00 | 0.9336 | 1.0000 | 0.9972 |
|  | S. | ProAcePred | 81.60 | 89.50 | 73.70 | 0.6400 | 0.7780 | N/A |
|  | DNNAce | 95.00 | 95.00 | 95.00 | 0.9155 | 1.0000 | 0.9796 |
|  | V. | ProAcePred | 86.90 | 89.00 | 84.70 | 0.7380 | 0.8550 | N/A |
|  | DNNAce | 94.02 | 94.02 | 94.02 | 0.8855 | 0.9907 | 0.9919 |

*Note:* N/A means not available.

**3. Supplementary Figure**

**C:\Users\Administrator\Desktop\图片4.emf**

**Fig. S1.** EBGW comparison between acetylation and non-acetylation. The vertical axis represents the log2 ratio of the average EBGW values between acetylation and non-acetylation. The horizontal axis , , and  represent three binary sequences, respectively.

F:\项目2_乙酰化位点_Acetylation\ProAcePred_DataSet相关程序以及结果\以ACC确定参数预测_2\ROC+PR曲线—selection\feature_ROC.emf

**Fig. S2.** ROC curves of different dimensional reduction methods on nine datasets.

**4. Supplementary References**

Chen,Z. et al. (2018) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, **34,** 2499-2502.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins,* **43,** 246-255.

Ding,Y. et al. (2017) Identification of drug-target interactions via multiple information integration. *Inform. Sciences*, **418-419,** 546-560

Gao,J. et al. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell. Proteomics*, **9,** 2586-2600.

Hasan,M.M. et al. (2017) Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. *Mol. Biosyst.*, **13,** 2545-2550.

Ju,Z. and He,J.J. (2017) Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. *J. Mol. Graph. Model.*, **76,** 356-363.

Kawashima,S. et al. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36,** D202-D205.

Li,W. and Godzik,A. (2006) CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22,** 1658-1659.

Zhang,Z.H. et al. (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *Febs Lett.,* **580,** 6169-6174.