mathematics

MDPI

Article

DeepDA-Ace: A Novel Domain Adaptation Method for Species-Specific Acetylation Site Prediction

Yu Liu [1], Qiang Wang [1] and Jianing Xi [2,*]

[1] School of Integrated Circuits, Anhui University, 111 JiuLong Road, Hefei 230601, China; liuyu@ahu.edu.cn (Y.L.); p21301180@stu.ahu.edu.cn (Q.W.)
[2] School of Biomedical Engineering, Guangzhou Medical University, Guangzhou 510182, China
* Correspondence: xjn@gzhmu.edu.cn

Abstract: Protein lysine acetylation is an important type of post-translational modification (PTM), and it plays a crucial role in various cellular processes. Recently, although many researchers have focused on developing tools for acetylation site prediction based on computational methods, most of these tools are based on traditional machine learning algorithms for acetylation site prediction without species specificity, still maintained as a single prediction model. Recent studies have shown that the acetylation sites of distinct species have evident location-specific differences; however, there is currently no integrated prediction model that can effectively predict acetylation sites cross all species. Therefore, to enhance the scope of species-specific level, it is necessary to establish a framework for species-specific acetylation site prediction. In this work, we propose a domain adaptation framework DeepDA-Ace for species-specific acetylation site prediction, including Rattus norvegicus, Schistosoma japonicum, Arabidopsis thaliana, and other types of species. In DeepDA-Ace, an attention based densely connected convolutional neural network is designed to capture sequence features, and the semantic adversarial learning strategy is proposed to align features of different species so as to achieve knowledge transfer. The DeepDA-Ace outperformed both the general prediction model and fine-tuning based species-specific model across most types of species. The experiment results have demonstrated that DeepDA-Ace is superior to the general and fine-tuning methods, and its precision exceeds 0.75 on most species. In addition, our method achieves at least 5% improvement over the existing acetylation prediction tools.

Keywords: domain adaptation; post-translational modification; acetylation; deep learning

MSC: 68T07

check for updates

Citation: Liu, Y.; Wang, Q.; Xi, J. DeepDA-Ace: A Novel Domain Adaptation Method for Species-Specific Acetylation Site Prediction. *Mathematics* **2022**, *10*, 2364. https://doi.org/10.3390/math10142364

Academic Editors: Liang Zou, Liang Zhao, Yonghui Xu and Jakub Nalepa

Received: 1 June 2022
Accepted: 4 July 2022
Published: 6 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Protein lysine acetylation is a reversible and highly regulated post-translational modification (PTM), which plays a vital role in many biological processes, such as protein expression, gene expression and metabolism [1–3]. Recently, studies have reported that it is also associated with some specific human diseases, such as cancer and addiction [4]. The accurate identification of acetylation sites is a key process in the study of protein molecular mechanisms, and has important guiding significance for revealing protein functions. For example, mutations in cancer driver genes can affect the lysine acetylation sites and alter the protein functions to lead to cancer progress [5]. Consequently, the identification of acetylation sites not only can promote the understanding on protein function, but also can locate the attacks in driver gene sequences. So far, many experimental methods have been proposed to identify acetylation sites, such as the radioactive chemical method [6], mass spectrometry analysis [7], and others. However, due to the high cost of experimental-based identifications on both time and money, computational-based acetylation site prediction methods have drawn widespread attention.

Currently, various computation-based methods have been proposed to predict the acetylation site. For instance, Xue et al. [8] developed an acetylation site prediction tool called EnsemblePail using ensemble of support vector machine (SVM) models, where experiment results showed that the ensemble SVM model outperformed the single SVM model. Subsequently, Hou et al. [9] developed a logistic regression-based acetylation site prediction method LaceP, which utilized various biological features, including sequence features and physicochemical property. Meanwhile, Li et al. [10] proposed a species-specific lysine acetylation prediction method called SSPKA through random forest classifiers to combine multiple sequence features. In addition, Chen et al. [11] developed a novel method ProAcePred for species-specific lysine acetylation site prediction, which adopted elastic net to optimize features and considerably improved the prediction performance. The above studies suggest that species-specific models perform better than general models in most cases, and the reason may lie in the differences in the local sequence of acetylation sites between different species [12].

In recent years, as an emerging artificial intelligence technique, deep learning has achieved significant success in different fields [13–15], especially in PTM site prediction [16–18]. Compared with traditional machine learning, deep learning models can automatically mine complex patterns from raw protein sequences, therefore achieving abstract representation of training data and effectively improving prediction performance. For example, Wang et al. [19] successfully applied the Capsule network to lysine acetylation site prediction, which outperformed the traditional machine learning methods and baseline convolutional neural network architecture. Meanwhile, Chen et al. recently proposed a lysine post-translational modification site prediction method called MUscADEL [20] via bidirectional long short-term memory network, which can effectively predict various post-translational modification sites, including acetylation, ubiquitination and methylation. The experimental results showed that MUscADEL can greatly improve the prediction accuracy. Although deep learning algorithms can outperform traditional machine learning algorithms in acetylation site prediction, existing deep learning-based methods are usually designed for general acetylation site prediction, and thus they are difficult to achieve accurate prediction across multiple species.

Despite the efficiency of deep learning, the most crucial challenge in building species-specific acetylation site prediction model is that deep learning algorithms require massive amount of training data [21], which is not applicable on many types of species. For example, there are only 384 and 519 experimentally verified acetylation sites for Arabidopsis thaliana and Oryza sativa in PLMD database. Thus, it is difficult to directly train a dedicated predictive model for these species based on such limited data. A common approach is to adopt a simple transfer learning technique known as fine tuning [22], which builds a pre-trained model on a relevant dataset that has large amount of data firstly, and then fine tunes the model through task-specific datasets. Although the fine-tuning strategy has been widely used in many tasks, it is still limited for a small data size of the task-specific dataset, leading to an inefficiency and even over-fitting issue [23,24].

To address the problem of species-specific acetylation site prediction, here we propose a novel domain adaptation method named DeepDA-Ace, which can effectively enhance the species-specific acetylation site prediction models by transferring useful knowledge from the source domain (human data) to the target domain (other species). In DeepDA-Ace, we design an attention-based densely connected convolutional neural network (DCNN) as sequence feature extraction network, which can efficiently extract discriminative features for acetylation site prediction. In addition, we create four types of sample pairs with both label and domain information for adversarial learning, which are able to not only confuse the domain-class discriminator to reduce the distributional difference between source and target domain, but also maintain a high classification ability for the source and target samples. Extensive experiments illustrate that DeepDA-Ace has advantages over the general prediction model and the species-specific models trained with the fine-tuning strategy. Meanwhile, compared with existing well-known acetylation site prediction tools,

our proposed method also achieves substantial performance improvement. Generally, our domain adaptation based DeepDA-Ace can facilitate the species-specific acetylation site predictions. The relevant code and data of DeepDA-Ace are deposited on GitHub https://github.com/xiaoyuleyuan/DeepDA-Ace/ (accessed on 31 May 2022).

The contributions of this work can be summarized as follows:

(1) We propose a semantic adversarial learning strategy that reduces the differences in acetylation distribution across species, enabling efficient knowledge transfer.
(2) We design an attention-based DCNN model to extract discriminative features from local sequence of potential acetylation sites to improve the prediction accuracy.
(3) We conduct extensive comparison experiments to demonstrate the superiority of the proposed domain adaptation method over the fine-tuning method and state-of-the-art acetylation prediction tools.

2. Materials and Methods

2.1. Data Collection

Step 1.1: To build species-specific acetylation site prediction models, we collected acetylation data for 10 species: Homo sapiens (*H. sapiens*), Schistosoma japonicum (*S. japonicum*), Arabidopsis thaliana (*A. thaliana*), Saccharomyces cerevisiae (*S. cerevisiae*), Bacillus velezensis (*B. velezensis*), Rattus norvegicus (*R. norvegicus*), Escherichia coli (*E. coli*), Plasmodium falciparum (*P. falciparum*), Mus musculus (*M. musculus*), Oryza sativa (*O. sativa*). These data were collected from a comprehensive public database called Protein Lysine Modification Database [25], which contains various types of lysine modification data from multiple species. In total, we collected 23,080 proteins, covering 10 species, of which 3645 are from *M.musculus*, 6078 are from *H. sapiens*, 2960 are from *S. cerevisiae*, 4359 are from *R. norvegicus*, 1251 are from *S. japonicum*, 231 are from *A. thaliana*, 1860 are from *E. coli*, 1146 are from *B. velezensis*, 1214 are from *P. falciparum*, and 336 are from *O. sativa*.

Step 1.2: Subsequently, in order to reduce the evaluation bias caused by protein homology, we clustered the protein sequences using CD-HIT [26] tool to remove homologous proteins with 40% threshold. After that, we obtained 4993 *H. sapiens*, 2759 *M. musculus*, 2729 *S. cerevisiae*, 3372 *R. norvegicus*, 1031 *S. japonicum*, 192 *A. thaliana*, 1766 *E. coli*, 1103 *B. velezensis*, 1177 *P. falciparum* and 287 *O. sativa* distinct proteins.

Step 1.3: For negative samples, we selected lysine residues that were not verified as an acetylation site from the above non-homologous proteins.

Step 1.4: Afterwards, we randomly selected 10% positive and negative samples as an independent testing dataset from each species, and the remaining data were used as the training and validation set. Finally, in order to avoid the over-optimization problem caused by the imbalance of positive and negative samples, we randomly selected the same number of negative samples as the positive samples to form the balance dataset (Table 1). We have provided the original data and preprocessed data on FigShare (https://doi.org/10.6084/m9.figshare.20069816, accessed on 27 May 2022).

Table 1. Statistical summary of acetylation datasets constructed for ten species.

Species	Number of Proteins	Number of Sites
H. sapiens	4993	24,376
M. musculus	2759	9837
S. cerevisiae	2729	12,189
R. norvegicus	3372	10,306
S. japonicum	1031	1921
A. thaliana	192	304
E. coli	1766	8681
B. velezensis	1103	2905
P. falciparum	1177	3060
O. sativa	287	438

2.2. Protein Sequence Coding

Step 2.1: Acetylation site prediction can be regarded as a binary classification problem, namely, each lysine residue can be recognized as either an acetylation site or a non-acetylation site. We truncated 31-residue long (−15~+15) symmetrical windows with lysine residue at the center from the protein fragments as training sample.

Step 2.2: In order to obtain fixed-length sequence fragments, we used a non-existing amino acid 'X' to fill the blank position or non-standard amino acids. Afterward, we used one-hot encoding strategy to transform the protein fragments into a numeric vector for model construction. For instance, lysine (K) was encoded as "00000000000100000000", and serine (S) was encoded as "00000000000000001000".

Step 2.3: Finally, each protein fragment was encoded as 31 × 21 two-dimensional features, where 31 represents the length of the protein fragment and 21 represents the types of amino acids.

2.3. DeepDA-Ace Architecture

DeepDA-Ace is a novel species-specific acetylation site-prediction method, and the architecture is described in Figure 1. It consists of a sequence feature extraction network (g) that maps fragments into an embedding space and a predictor (h) for predicting embedded features. In addition, the DeepDA-Ace also contains a domain-class discriminator to identify sample sources. Among them, the sequence feature extraction network is modeled by three DCNN modules; each module contains six densely connected convolutional layers and a self-attention block. Meanwhile, the predictor and domain-class discriminator are both modeled by a fully connected layer. The structure and parameter information of DeepDA-Ace are shown in Appendix A and Table A1.

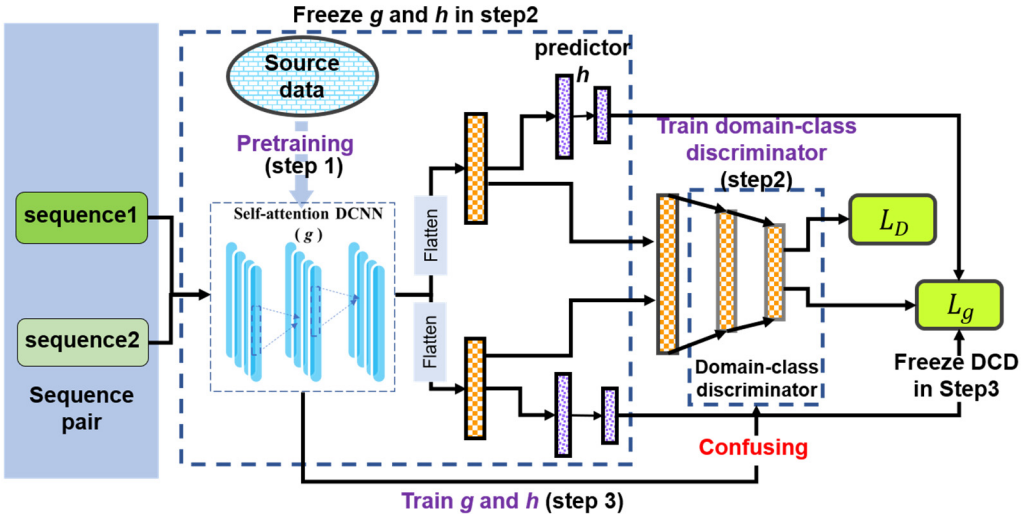


Figure 1. The overall framework of DeepDA-Ace.

In this study, we define human acetylation as the source domain D^s and other species as the target domain D^t. For source domain, we are given training dataset $X^s = \{x_1^s, \ldots x_i^s \ldots, x_N^s\}$ and corresponded label set $Y^s = \{y_1^s, \ldots y_i^s \ldots, y_N^s\}$, where N refers to the number of training samples of *H. sapiens*. For the target domain, we are given training dataset $X^t = \{x_1^t, \ldots x_i^t \ldots, x_M^t\}$ and corresponded label set $Y^t = \{y_1^t, \ldots y_i^t \ldots, y_M^t\}$, where N refers to the number of training samples of other species (e.g., *R. norvegicus*). Due to the large differences in local sequences of acetylation between humans and other species, i.e., the distributions of X^s and X^t are different, a predictive model f_s (here $f_s = h_s \circ g_s$) trained with D^s may not perform well on the other species. Training a new predictive model with D^t could bring substantial performance improvement. However, the size of D^t is usually small for most of species, making it difficult to simply train a new deep model for other species. Therefore, we propose a semantic adversarial

learning strategy to reduce the distribution difference between different species. The main idea of the semantic adversarial is aligning the sequence feature distributions of different species in the embedded feature space. Once the features are aligned, the domain-class discriminator will no longer be able to distinguish whether a feature is coming from human or other species, such that acetylation knowledge can be effectively transferred from human to other species.

Step 3.1: To achieve semantic adversarial learning, we create four types of pairs as follows: the first type $\mathcal{P}_1$ contains two samples with the same class label and both from the *H. sapiens* dataset; the second type $\mathcal{P}_2$ also have the same class label but one of the samples is from *H. sapiens* and the other is from another species. Similarly, the third type $\mathcal{P}_3$ contains two samples with different class labels and both come from the *H. sapiens* dataset, while the fourth type $\mathcal{P}_4$ contains two samples with the different class label and come from different species (one from *H. sapiens* and one from other species). In this way, we encode both species and label information of training data pairs, which may facilitate the semantic alignment of *H. sapiens* and other species.

Step 3.2: Classical adversarial learning is only interested in learning domain discriminator to identify sample sources. However, in our work, we need to consider not only domain information, but also semantic information. Therefore, we learn a domain-class discriminator to achieve domain and semantic alignment of human and other species. Specifically, domain-class discriminator is a multi-label classifier that can determine the type to which a given sample pair belongs. We model the domain-class discriminator with 3 fully connected layers, and the last layer is activated with a softmax function to output the classification result. To obtain the optimal domain discrimination performance, we train the domain-class discriminator with the standard categorical cross-entropy loss:

$$L_D = \frac{1}{Q} \sum_{\mathcal{P}_i^j \in P} \sum_{i=1}^{4} y_{\mathcal{P}_i^j} \log\Big(D\Big(\phi\Big(\mathcal{P}_i^j\Big)\Big)\Big) \tag{1}$$

where P denotes sets of all data pairs, Q denotes the number of pairs, $y_{\mathcal{P}_i^j}$ denotes the label of $\mathcal{P}_i^j$ and D denotes the domain-class discriminator function. ϕ denotes a symbolic function that connects the features extracted from the sequence feature extraction network for two samples in a sample pair. The output of ϕ is further fed into the domain class discriminator for pair type determination.

Step 3.3: Next, we focus on updating g_t to confuse the domain-class discriminator so that the domain-class discriminator is no longer able to distinguish between sample pairs of type 1 and type 2, and between sample pairs of type 3 and type 4. Specifically, we use the following loss function to optimize g_t:

$$L_g = \frac{1}{Q_2} \sum_{\mathcal{P}_2^i \in \mathcal{P}_2} \Big[y_{\mathcal{P}_1} \log\Big(D\Big(\phi\Big(\mathcal{P}_2^i\Big)\Big)\Big)\Big] + \frac{1}{Q_4} \sum_{\mathcal{P}_4^j \in \mathcal{P}_4} \Big[y_{\mathcal{P}_3} \log\Big(D\Big(\phi\Big(\mathcal{P}_4^j\Big)\Big)\Big)\Big] \tag{2}$$

where Q_2 denotes the number of pairs of $\mathcal{P}_2$, and Q_4 denotes the number of pairs of $\mathcal{P}_4$. Furthermore, to obtain high prediction accuracy simultaneously, we further minimize the loss (2) with the acetylation classification loss:

$$\begin{aligned} L_g = -\lambda \tfrac{1}{Q_2} \sum_{\mathcal{P}_2^i \in \mathcal{P}_2} & \big[y_{\mathcal{P}_1} \log(D(\phi(\mathcal{P}_2^i)))\big] + \tfrac{1}{Q_4} \sum_{\mathcal{P}_4^j \in \mathcal{P}_4} \Big[y_{\mathcal{P}_3} \log\Big(D\Big(\phi\Big(\mathcal{P}_4^j\Big)\Big)\Big)\Big] \\ & - \tfrac{1}{N} \sum_{k=1}^{N} y_k^s ln s_k^s + (1 - y_k^s) ln(1 - s_k^s) \; - \tfrac{1}{M} \sum_{l=1}^{M} y_l^t ln s_l^t \\ & + (1 - y_l^t) ln(1 - s_l^t) \end{aligned} \tag{3}$$

where λ strikes the balance between classification and confusion, s_k^s is the classification score of kth samples of source domain, and s_l^t is the classification score of lth samples of

target domain. Through the above iterative training, domain class discriminator would not be able to distinguish human and other species, which indicates that the difference in feature distributions of human and other species is reduced. Meanwhile, the domain class discriminator is still able to discriminate the pairs that consist of different class sample or same class sample, which indicate that different class samples are mapped separability in the embedded feature space.

2.4. Performance Evaluation

To evaluate the performance of DeepDA-Ace, we use several common indicators to measure prediction performance, including area under the ROC curves (AUC), sensitivity (Sn), specificity (Sp), precision (Pre), accuracy (Acc) and F1 score based on the independent test dataset. The calculation formula of these indicators is as follows:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{6}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Pre} \times \mathrm{Sn}}{\mathrm{Pre} + \mathrm{Sn}} \tag{8}$$

where FP, TP, FN and TN represent false positives, true positives, false negatives and true negatives, respectively.

3. Results

3.1. Sequence Analysis of Acetylation Sites in Different Species

The patterns of acetylation can be visualized using a two-sample logo tool [27], which identifies significant differences in amino acid compositions between acetylation sites versus non-acetylation (Figure 2). The upper section of each logo graph represents amino acids that are enriched around acetylation sites; the lower section represents amino acids that depleted around acetylation sites. From Figure 2, we find that sequence preferences differ across species. For example, compared with *H. sapiens* and *R. norvegicus*, Tyrosine/Y and Phenylalanine/F prefer to enrich in position +2 and -1, -2 around the *B. velezensis* acetylation sites. Alanine/A and Valine/V are found to be enriched in upstream for *H. sapiens* but not enriched for *E. coli*. Although the acetylated sequence of *H. sapiens* has a difference from other species, there are still similarities between them. For most species, the Aspartic acid/D residue is observed to be enriched in the -1 position of acetylation sites. Meanwhile, Leucine/L residue prefers to occur in the +1 position of the acetylation site on *H. sapiens*, *R. norvegicus* and *E. coli*. In addition, the lysine/K is significantly depleted at varying positions upstream and downstream of the acetylation site on most species. These observations show that there are certain similarities of the acetylation sequence between human and other species, so that acetylation-related knowledge of *H. sapiens* could be transferred to prediction tasks in other species to improve the prediction performance. Furthermore, given the contrasting difference in the sequence preference of different species, the domain adaptation technique is necessary to reduce the distribution differences between different species and achieve efficient knowledge transfer.

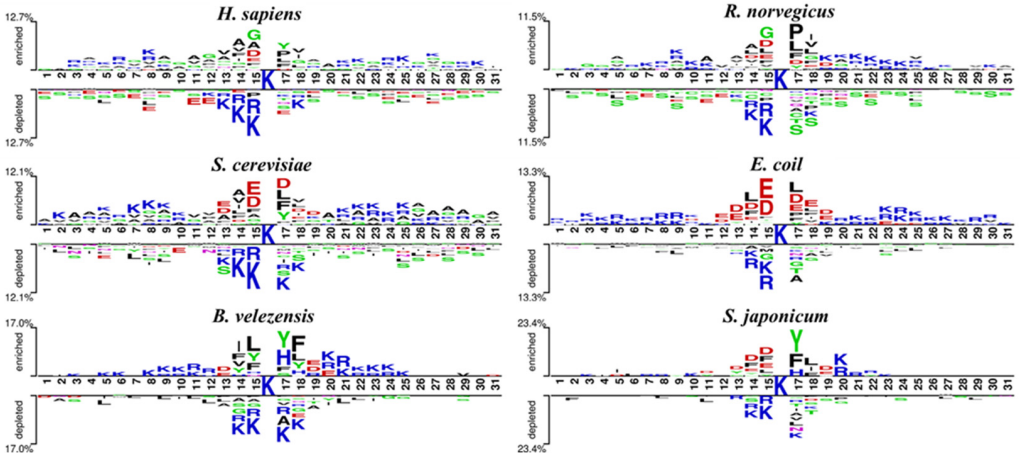


Figure 2. The amino acid frequencies of different species of *H. sapiens*, *R. norvegicus*, *S. cerevisiae*, *E. coli*, *B. velezensis* and *S. japonicum* dataset. Only residues significantly enriched or depleted (*t*-test, $p < 0.05$) flanking the centered acetylation sites (upstream 15 residues and downstream 15 residues) are shown.

3.2. Effectiveness of Domain Adaptation

To evaluate the ability of our proposed domain adaptation method in acetylation site prediction, we compare DeepDA-Ace with other three baseline models, including both general model and species-specific models. (1) Combined method: we build general acetylation site prediction model using the data by combining the training datasets of all species. (2) Simple training: we train the species-specific acetylation site prediction model for each species directly using only dataset of target species. (3) Fine-tune: this method uses a simple transfer learning technique. Specifically, we first train a prediction model on *H. sapiens* dataset, and then fine-tune the parameters using the dataset of other species to form a species-specific prediction model.

We compare the DeepDA-Ace with baseline models on the independent testing set; the ROC curves are plotted in Figure A1, and the AUC values are list in Table 2. Clearly, the species-specific models are observed to outperform general prediction models on most species, but the simple training method is inferior to the combined method on few species. For instance, the simple training strategy only attains 0.696, 0.707 and 0.659 AUC value on *A. thaliana*, *B. velezensis* and *O. sativa*, which indicate that building predictive models based on limited species-specific datasets cannot achieve satisfying results. Meanwhile, we find that using the transfer learning technique could achieve a valuable performance increase over the simple training strategy. For instance, the fine-tuning method obtains 0.742, 0.722 and 0.756 AUC value on *A. thaliana*, *B. velezensis* and *O. sativa*, which has an improvement of 4.6%, 1.5% and 9.7% compared with the simple training method.

Table 2. The AUC value of DeepDA-Ace and baseline methods.

Species	Species-Specific			General
	DeepDA-Ace	Fine-Tune	Simple Training	Combined
M. musculus	0.758	0.735	0.702	0.724
S. cerevisiae	0.780	0.732	0.713	0.732
R. norvegicus	0.732	0.699	0.691	0.703
S. japonicum	0.795	0.680	0.617	0.722
A. thaliana	0.798	0.742	0.696	0.704
E. coli	0.749	0.701	0.705	0.712
B. velezensis	0.794	0.722	0.707	0.684
P. falciparum	0.688	0.594	0.636	0.593
O. sativa	0.836	0.756	0.659	0.670
Average	0.770	0.707	0.680	0.694

Although the fine-tuning strategy has a certain improvement in the predictive performance of acetylation sites, the performance is still inferior to our proposed method. By using the domain adaptation technique, DeepDA-Ace performs knowledge transfer more efficiently and improves the performance on all species. For example, DeepDA-Ace achieves 4.8%, 3.4% and 11.5% AUC value improvements on *E. coli*, *R. norvegicus* and *S. japonicum* over the fine-tuning method. Additionally, DeepDA-Ace also shows superior advantages on species that only have few training samples, such as *A. thaliana*, *B. velezensis* and *O. sativa*. Meanwhile, our proposed method performs consistently better than the general prediction model on almost species. For instance, DeepDA-Ace obtains 0.758, 0.749 and 0.798 AUC value on *M. musculus*, *E. coli* and *A. thaliana*, which has an improvement of 3.4%, 3.7% and 9.4% compared with the combined method. In addition, we find an interesting phenomenon that, benefited from the large amount of training data, the general prediction model obtains better performance than the simple training method, but it has unsatisfactory performance on *B. velezensis* and *O. sativa*; the reason probably lies in the fact that these two species have larger differences with the sequence pattern of other species. This result indicates that the general prediction model cannot effectively extract species-specific features and limit the performance of species-specific acetylation sites prediction. Our proposed method could effectively transfer acetylation-related knowledge from human data to other species prediction tasks by learning the embedding feature space, resulting in significantly improved prediction performance on other species. Taken together, the domain adaptation technique can solve the small sample learning problem and achieve superior performance on multiple species.

In addition to the ROC curve and AUC value, we also adopt F1 score, Sn, Sp, Acc and Pre to evaluate our proposed domain adaptation method. Specifically, we inspect these indicators at medium stringency level (90%) and high stringency level (95%) specificity threshold. Correspondingly, we calculate the above indicators for each species, displayed in Table 3. Obviously, species-specific models consistently yield higher measurements than general prediction model. Taking *O. sativa* as an example, at medium stringency level, the F1 score, Acc, Pre and Sn values of the fine-tune method are 0.535, 0.629, 0.826 and 0.396, respectively, while the general prediction method only achieves 0.514, 0.618, 0.818 and 0.375 on the above indicators. Meanwhile, we observe that our proposed DeepDA-Ace has better prediction performance compared with the fine-tuning method. For instance, on the *O. sativa* dataset, compared with the simple training method, DeepDA-Ace obtains 21.3% improvement for F1 score, 8.7% for Pre, 11.2% for Acc, and 20.8% for Sn. Furthermore, DeepDA-Ace also achieves better results at a high-stringency level. For instance, on *S. cerevisiae*, the performance of the fine-tuning method are Pre of 0.758, F1 score of 0.252, Acc of 0.543 and Sn score of 0.151. In comparison, DeepDA-Ace further improves the prediction performance, and the corresponding Pre, F1 score, Acc and Sn reach to 0.809, 0.326, 0.57 and 0.204, respectively. In summary, DeepDA-Ace effectively alleviates the small sample problem and significantly improves the acetylation site prediction performance.

Table 3. The predictive performance of DeepDA-Ace and baseline methods.

			Sp = 95%				*Sp* = 90%			
Species	**Method**		**Pre**	**F1**	**Acc**	**Sn**	**Pre**	**F1**	**Acc**	**Sn**
R. norvegicus	species-specific	DeepDA-Ace	0.783	0.287	0.557	0.176	0.746	0.411	0.587	0.284
		Fine-tune	0.720	0.213	0.531	0.125	0.700	0.341	0.558	0.225
		Simple training	0.759	0.256	0.546	0.154	0.734	0.390	0.578	0.266
	general	Combined	0.749	0.243	0.541	0.145	0.743	0.406	0.585	0.279
S. japonicum	species-specific	DeepDA-Ace	0.804	0.339	0.584	0.215	0.804	0.542	0.657	0.408
		Fine-tune	0.706	0.213	0.54	0.126	0.703	0.353	0.571	0.236
		Simple training	0.697	0.205	0.538	0.12	0.708	0.359	0.574	0.241
	general	Combined	0.778	0.297	0.569	0.183	0.736	0.403	0.592	0.277
S. cerevisiae	species-specific	DeepDA-Ace	0.809	0.326	0.57	0.204	0.791	0.50	0.627	0.365
		Fine-tune	0.758	0.252	0.543	0.151	0.752	0.421	0.59	0.292
		Simple training	0.780	0.281	0.553	0.171	0.752	0.421	0.59	0.292
	general	Combined	0.771	0.268	0.549	0.163	0.748	0.415	0.588	0.287
M. musculus	species-specific	DeepDA-Ace	0.793	0.298	0.56	0.184	0.760	0.438	0.599	0.308
		Fine-tune	0.735	0.225	0.535	0.133	0.716	0.365	0.567	0.245
		Simple training	0.743	0.234	0.538	0.139	0.731	0.386	0.576	0.263
	general	Combined	0.776	0.273	0.551	0.166	0.736	0.395	0.579	0.27
E. coli	species-specific	DeepDA-Ace	0.805	0.325	0.573	0.203	0.771	0.462	0.613	0.33
		Fine-tune	0.750	0.246	0.545	0.147	0.723	0.379	0.576	0.257
		Simple training	0.788	0.297	0.563	0.183	0.749	0.422	0.594	0.293
	general	Combined	0.788	0.297	0.563	0.183	0.746	0.415	0.591	0.288
B. velezensis	species-specific	DeepDA-Ace	0.829	0.352	0.570	0.224	0.815	0.541	0.64	0.405
		Fine-tune	0.725	0.208	0.516	0.122	0.723	0.360	0.554	0.240
		Simple training	0.750	0.233	0.525	0.138	0.733	0.377	0.561	0.253
	general	Combined	0.797	0.295	0.547	0.181	0.741	0.388	0.566	0.263

Table 3. *Cont.*

			Sp = 95%				*Sp* = 90%			
Species	**Method**		**Pre**	**F1**	**Acc**	**Sn**	**Pre**	**F1**	**Acc**	**Sn**
P. falciparum	species-specific	DeepDA-Ace	0.717	0.203	0.513	0.118	0.746	0.391	0.567	0.265
		Fine-tune	0.706	0.194	0.510	0.112	0.681	0.301	0.529	0.193
		Simple training	0.758	0.245	0.528	0.146	0.713	0.341	0.546	0.224
	general	Combined	0.741	0.227	0.521	0.134	0.678	0.297	0.528	0.190
O. sativa	species-specific	DeepDA-Ace	0.882	0.462	0.607	0.313	0.852	0.613	0.674	0.479
		Fine-tune	0.714	0.182	0.494	0.104	0.826	0.535	0.629	0.396
		Simple training	0.867	0.413	0.584	0.271	0.765	0.400	0.562	0.271
	general	Combined	0.818	0.305	0.539	0.188	0.818	0.514	0.618	0.375
A. thaliana	species-specific	DeepDA-Ace	0.933	0.491	0.532	0.333	0.917	0.667	0.645	0.524
		Fine-tune	0.500	0.045	0.323	0.024	0.895	0.557	0.565	0.405
		Simple training	0.889	0.314	0.435	0.19	0.882	0.508	0.532	0.357
	general	Combined	0.875	0.28	0.419	0.167	0.778	0.275	0.403	0.167

3.3. Comparison with Existing Acetylation Site Prediction Tools

We compare DeepDA-Ace with two well-known acetylation site prediction tools CapsNet [19] and PAIL [28], based on independent test data. For each species, we plot the ROC curves and calculate the AUC values of different tools as shown in Figure 3. The results show that our proposed DeepDA-Ace obtains the best prediction performance on all species. Taking *R. norvegicus* as an example, the ROC curve of DeepDA-Ace is significantly higher than other tools, and the AUC value of DeepDA-Ace is 6.6% and 20.5% higher than CapsNet and PAIL, respectively. Additionally, on the *S. japonicum* dataset, DeepDA-Ace obtains 0.789 AUC value, while CapsNet and PAIL only has 0.778 and 0.590 AUC value, respectively. In addition, we find that CapsNet is superior to PAIL on most species. For example, the AUC values of CapsNet are 0.659, 0.778, 0.739 and 0.711 for *R. norvegicus*, *S. japonicum*, *S. cerevisiae* and *M. musculus*, and the obtained performance improvements are 15.5%, 18.8%, 19.6% and 17.7% compared with PAIL. These results show that, compared to traditional machine learning, deep learning can learn protein sequence patterns more effectively and has more advantages in acetylation site prediction.

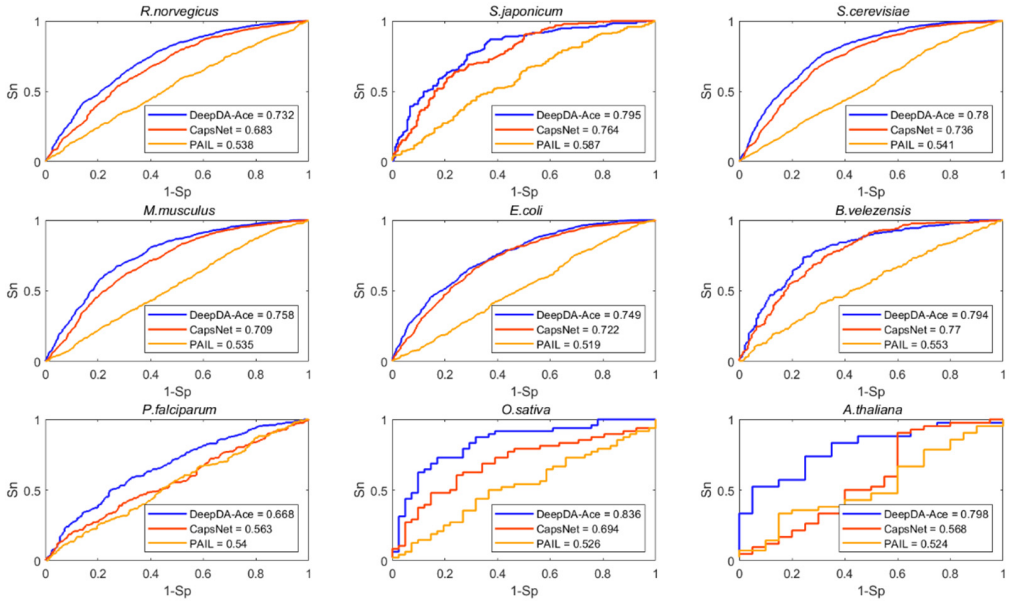


Figure 3. ROC curves on *R. norvegicus*, *S. japonicum*, *S. cerevisiae*, *M. musculus*, *E. coil*, *R. norvegicus*, *B. velezensis*, *F. palciparum*, *O. sativa* and *A. nidulans* with different methods. The blue lines represent DeepDA-Ace, the red and orange lines represent CapsNet and PAIL, respectively.

Additionally, we also calculate indicators including F1 score, Sn, Acc and Pre to evaluate different methods, and the results are listed in Table A2. Correspondingly, we also draw a histogram (Figure 4) to show the performance intuitively. Clearly, DeepDA-Ace achieves higher indicators than other tools on most species. For example, at the high stringency level, the performance of DeepDA-Ace is increased by 2.8% for Acc, 3.3% for Sn, 2.3% for Pre and 4.3% for F1 score, compared with CapsNet on *B. velezensis*. Additionally, on the *E. coli* dataset, the indicators of Acc, Sn, Pre and F1 score are also improved 3.2%, 6.5%, 6.7%, and 9.2%, respectively. In addition, DeepDA-Ace also achieves the best performance at a medium-stringency level. Taking *B. velezensis* as an example, the F1 score, Pre, Acc, Sn values of DeepDA-Ace are increased 11.8%, 5.4%, 5.8% and 11.2% compared with CapsNet. Meanwhile, compared to PAIL, DeepDA-Ace also achieves 33.1%, 23.3%, 14.4% and 27.7% improvement. In conclusion, our proposed DeepDA-Ace is superior to the existing, well-known acetylation tools.

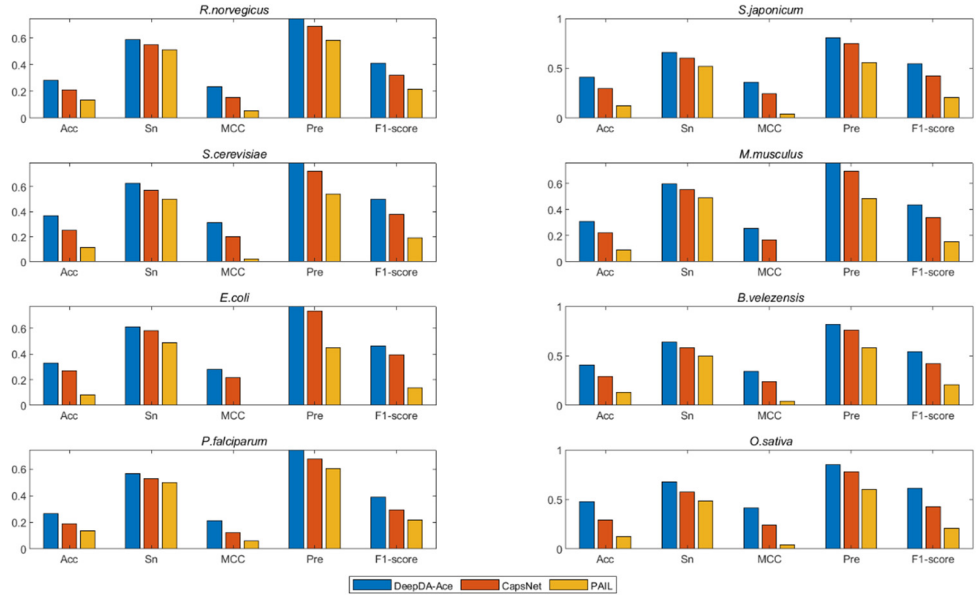


Figure 4. The F1, Acc, Pre, Sn value comparison with different methods at specificity of 90.0%.

4. Discussion and Conclusions

Protein lysine acetylation is an important type of PTMs, and it plays a key role in many cellular processes. In this work, we develop a novel domain adaptation method DeepDA-Ace for species-specific acetylation site prediction. DeepDA-Ace takes raw protein sequence as input and obtains excellent prediction performance on several species. The experiment results show that our proposed DeepDA-Ace surpass other well-known acetylation site prediction tools, achieving AUC value improvement of more than 5% on most species. Additionally, by comparison with other training strategies, the domain adaptation technique shows more powerful strength for predicting species-specific acetylation sites. Especially on species that have few training samples (e.g., *O. sativa* and *A. thaliana*), DeepDA-Ace achieves AUC values of 0.863 and 0.798, respectively, which are 17.7% and 10.2% higher than the simple training method. In general, the superiority of DeepDA-Ace is mainly due to the following three reasons: Firstly, our proposed sequence feature extraction network adopts densely connected ideal and attention mechanism that has stronger representation ability of sequence features compared with traditional machine learning methods. Secondly, compared with the general prediction model of the acetylation site, our method builds a model for various species, which could extract more species-specific features to improve the prediction performance. Finally, our proposed semantic adversarial learning strategy not only reduces the distribution difference between human and other species, but also achieves semantic alignment in the embedded feature space so that the model can transfer related knowledge between different species more effectively.

Although DeepDA-Ace has substantial improvements in species-specific acetylation site prediction, it can still be enhanced by integrating more advanced computing techniques and biological features. Firstly, some other biological information such as protein–protein interaction [29,30] and post-translational modification cross talk [31] have also been reported to be helpful for acetylation site prediction, so we will further integrate these biological features in future work. Secondly, in addition to the acetylation modification sites reported in existing databases, there are still a large number of acetylation sites that have not been discovered. If we could introduce these potential acetylation data when model training, it could further alleviate the small sample problem in species-specific acetylation site prediction and improve the prediction performance. Therefore, we will extend our method with the semi-supervised learning technique to enhance the model generalization ability on other species. Additionally, for the application on cancer, our prediction model can measure the changes of lysine acetylation on the mutation sites in

driver gene sequence [32], facilitating the understanding of protein function in cancer. The architecture of DeepDA-Ace is based on deep neural network, which still lacks interpretability. In the future work, we will explore some visualization methods of the neural network to enhance model interpretability for further research on the mechanism of acetylation. In summary, DeepDA-Ace shows powerful capability of reducing the differences in the sequence distribution of different species, which suggests that our proposed DeepDA-Ace could be extended to another small sample learning task. In conclusion, DeepDA-Ace not only has outstanding performance on small sample species but also has the potential to be extended to other prediction tasks.

Author Contributions: Conceptualization, Y.L. and J.X.; methodology, Y.L.; validation, Y.L., J.X. and Q.W.; investigation, Y.L.; resources, J.X.; data curation, Y.L.; writing—original draft preparation, Y.L. and Q.W.; writing—review and editing, J.X.; project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 61901322.

Institutional Review Board Statement: Not applicable. All the data used in this study are downloaded from public online database.

Informed Consent Statement: Not applicable. All the data used in this study are downloaded from public online database.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank Lei Su for her helpful suggestion.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Details of DeepDA-Ace Architectures

In this section, we will introduce the details of deep learning architectures of DeepDA-Ace that are all implemented by Pytorch. The model consists with different layers, including the convolutional neuron network (CNN) layer, the flatten layer, full connection layer and the softmax layer, the optimizer used here is Adam. Specifically, the sequence feature extraction network g is modeled by three densely connected modules, each module contains six densely connected convolutional layers and an attention block, and the parameters of each convolution layer are shown in Table A1. The domain class discriminator is modeled by three fully connected layers, and the label classifier is modeled by two fully connected layers.

1. CNN layer: in the convolutional layer, it has a different number of feature maps in a different CNN layer, where the kernel size is 1 in the first CNN layer, 3 in the other layers, strides are 1, and the activation function is ReLU.
2. Flatten layer: in the flatten layer, the feature maps are flattened as one-dimensional features.
3. Fully connected layer: in the fully connected layer, the activation function is ReLU.
4. Softmax layer: the output layer, which has two neurons corresponding to acetylation and non-acetylation. The activation function is softmax.

Appendix A.2. Implementation Details of Architectures

Our DeepDA-Ace framework is implemented in Pytorch and trained on 1 NVIDIA GTX2080Ti GPU. We first pretrain sequence feature extraction network (g) and predictor (h) using human acetylation data. The learning rate of this stage is 1×10^{-4}, and the batch size is 256. Next, we train the domain-class discriminator with a 0.01 learning rate and 40 batch size. Finally, we train the sequence feature extraction network and predictor using both human and other species data. The learning rate of this stage is 1×10^{-5}, and the batch size is 20.

In the Section 3, the ROC curves are plotted by MATLAB. The performance indicators used in our paper, including F1 score, AUC value, etc., are calculated by MATLAB, and the corresponding MATLAB code is published in GitHub.

Appendix B

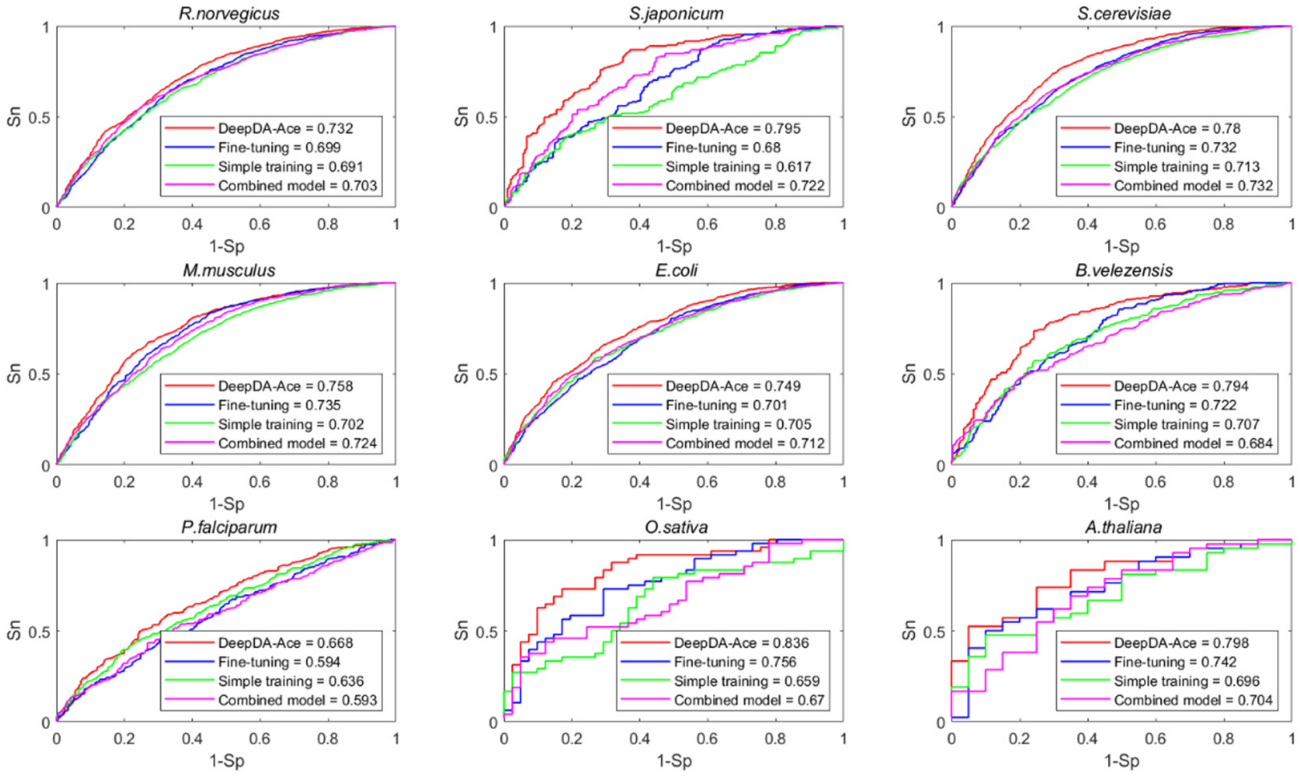


Figure A1. ROC curves on *R. norvegicus*, *S. japonicum*, *S. cerevisiae*, *M. musculus*, *E. coil*, *R. norvegicus*, *B. velezensis*, *F. palciparum*, *O. sativa* and *A. nidulans* with different methods. The red lines represent DeepDA-Ace, the blue, green and purple lines represent the fine tuning, simple training and combined model, respectively.

Table A1. The architecture of DeepDA-Ace.

Layers	Channels	Kernel	Output Size
Conv1(ReLU)	32	1×1	31×21
Conv2(ReLU)	48	3×3	31×21
Conv3(ReLU)	64	3×3	31×21
Conv4(ReLU)	80	3×3	31×21
Conv5(ReLU)	96	3×3	31×21
Conv6(ReLU)	112	3×3	31×21
Conv7(ReLU)	128	3×3	31×21
Transition1	64	1×1	31×21
Max-pooling1	64	2×2	15×10
Self-attention1			15×10
Conv8(ReLU)	80	3×3	15×10
Conv9(ReLU)	96	3×3	15×10
Conv10(ReLU)	112	3×3	15×10
Conv11(ReLU)	128	3×3	15×10
Conv12(ReLU)	144	3×3	15×10
Conv13(ReLU)	160	3×3	15×10
Transition2	80	1×1	15×10

Table A1. *Cont.*

Layers	Channels	Kernel	Output Size
Max-pooling2	80	2×2	7×5
Self-attention2			7×5
Conv14(ReLU)	96	3×3	7×5
Conv15(ReLU)	112	3×3	7×5
Conv16(ReLU)	128	3×3	7×5
Conv17(ReLU)	144	3×3	7×5
Conv18(ReLU)	160	3×3	7×5
Conv19(ReLU)	176	3×3	7×5
FC1(ReLU)	—	—	6160×1
FC2(Softmax)	—	—	2×1

Table A2. The predictive performance of DeepDA-Ace and existing tools.

		***Sp* = 95%**				***Sp* = 90%**			
Species	**Method**	**Pre**	**F1**	**Acc**	**Sn**	**Pre**	**F1**	**Acc**	**Sn**
R. norvegicus	DeepDA-Ace	0.783	0.287	0.557	0.176	0.746	0.411	0.587	0.284
	CapsNe	0.691	0.188	0.523	0.109	0.686	0.323	0.550	0.211
	PAIL	0.545	0.105	0.497	0.058	0.581	0.217	0.511	0.134
S. japonicum	DeepDA-Ace	0.804	0.339	0.584	0.215	0.804	0.542	0.657	0.408
	CapsNet	0.744	0.252	0.553	0.152	0.747	0.421	0.600	0.293
	PAIL	0.583	0.130	0.514	0.073	0.558	0.205	0.517	0.126
S. cerevisiae	DeepDA-Ace	0.809	0.326	0.57	0.204	0.791	0.500	0.627	0.365
	CapsNet	0.714	0.206	0.527	0.121	0.724	0.375	0.570	0.253
	PAIL	0.538	0.102	0.494	0.056	0.542	0.189	0.499	0.114
M. musculus	DeepDA-Ace	0.793	0.298	0.56	0.184	0.760	0.438	0.599	0.308
	CapsNet	0.682	0.179	0.519	0.103	0.697	0.338	0.555	0.223
	PAIL	0.467	0.077	0.488	0.042	0.484	0.153	0.488	0.091
E. coli	DeepDA-Ace	0.805	0.325	0.573	0.203	0.771	0.462	0.613	0.33
	CapsNet	0.738	0.233	0.541	0.138	0.731	0.391	0.581	0.267
	PAIL	0.488	0.085	0.495	0.047	0.449	0.136	0.486	0.08
B. velezensis	DeepDA-Ace	0.829	0.352	0.570	0.224	0.815	0.541	0.640	0.405
	CapsNet	0.806	0.309	0.552	0.191	0.761	0.423	0.582	0.293
	PAIL	0.611	0.129	0.491	0.072	0.582	0.210	0.496	0.128
P. falciparum	DeepDA-Ace	0.717	0.203	0.513	0.118	0.746	0.391	0.567	0.265
	CapsNet	0.727	0.213	0.516	0.125	0.674	0.293	0.526	0.187
	PAIL	0.605	0.128	0.489	0.072	0.606	0.219	0.500	0.134
O. sativa	DeepDA-Ace	0.882	0.462	0.607	0.313	0.852	0.613	0.674	0.479
	CapsNet	0.714	0.182	0.494	0.104	0.778	0.424	0.573	0.292
	PAIL	0.500	0.077	0.461	0.042	0.600	0.207	0.483	0.125
A. thaliana	DeepDA-Ace	0.933	0.491	0.532	0.333	0.917	0.667	0.645	0.524
	CapsNet	0.667	0.089	0.339	0.048	0.667	0.167	0.355	0.095
	PAIL	0.750	0.130	0.355	0.071	0.600	0.128	0.339	0.071

References

1. Kim, S.C.; Sprung, R.; Chen, Y.; Xu, Y.; Ball, H.; Pei, J.; Cheng, T.; Kho, Y.; Xiao, H.; Xiao, L.; et al. Substrate and Functional Diversity of Lysine Acetylation Revealed by a Proteomics Survey. *Mol. Cell* **2006**, *23*, 607–618. [CrossRef] [PubMed]
2. Kamita, M.; Kimura, Y.; Ino, Y.; Kamp, R.M.; Polevoda, B.; Sherman, F.; Hirano, H. N(α)-Acetylation of yeast ribosomal proteins and its effect on protein synthesis. *J. Proteom.* **2011**, *74*, 431–441. [CrossRef] [PubMed]
3. Glozak, M.A.; Sengupta, N.; Zhang, X.; Seto, E. Acetylation and deacetylation of non-histone proteins. *Gene* **2005**, *363*, 15–23. [CrossRef] [PubMed]
4. Gil, J.; Ramírez-Torres, A.; Encarnación-Guevara, S. Lysine acetylation and cancer: A proteomics perspective. *J. Proteom.* **2016**, *150*, 297–309. [CrossRef]
5. Xi, J.; Yuan, X.; Wang, M.; Li, A.; Li, X.; Huang, Q. Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* **2020**, *36*, 1855–1863. [CrossRef] [PubMed]
6. Medzihradszky, K.F. Peptide sequence analysis. *Methods Enzymol.* **2005**, *402*, 209–244.
7. Zhou, H.; Boyle, R.; Aebersold, R. Quantitative Protein Analysis by Solid Phase Isotope Tagging and Mass Spectrometry. *Methods Mol. Biol.* **2004**, *261*, 511–518. [CrossRef] [PubMed]
8. Xu, Y.; Wang, X.-B.; Ding, J.; Wu, L.-Y.; Deng, N.-Y. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *J. Theor. Biol.* **2010**, *264*, 130–135. [CrossRef]
9. Hou, T.; Zheng, G.; Zhang, P.; Jia, J.; Li, J.; Xie, L.; Wei, C.; Li, Y. LAceP: Lysine Acetylation Site Prediction Using Logistic Regression Classifiers. *PLoS ONE* **2014**, *9*, e89575. [CrossRef]
10. Li, Y.; Wang, M.; Wang, H.; Tan, H.; Zhang, Z.; Webb, G.I.; Song, J. Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci. Rep.* **2014**, *4*, 5765. [CrossRef]
11. Chen, G.; Cao, M.; Luo, K.; Wang, L.; Wen, P.; Shi, S. ProAcePred: Prokaryote lysine acetylation sites prediction based on elastic net feature optimization. *Bioinformatics* **2018**, *34*, 3999–4006. [CrossRef] [PubMed]
12. Chen, Z.; Zhou, Y.; Zhang, Z.; Song, J. Towards more accurate prediction of ubiquitination sites: A comprehensive review of current methods, tools and features. *Brief. Bioinform.* **2015**, *16*, 640–657. [CrossRef] [PubMed]
13. Zou, L.; Liu, W.; Lei, M.; Yu, X. An Improved Residual Network for Pork Freshness Detection Using Near-Infrared Spectroscopy. *Entropy* **2021**, *23*, 1293. [CrossRef] [PubMed]
14. Singh, G.; Sharma, S.; Kumar, V.; Kaur, M.; Baz, M.; Masud, M. Spoken Language Identification Using Deep Learning. *Comput. Intell. Neurosci.* **2021**, *2021*, 5123671. [CrossRef]
15. Lei, M.; Li, J.; Li, M.; Zou, L.; Yu, H. An Improved UNet++ Model for Congestive Heart Failure Diagnosis Using Short-Term RR Intervals. *Diagnostics* **2021**, *11*, 534. [CrossRef]
16. Yang, H.; Wang, M.; Liu, X.; Zhao, X.-M.; Li, A. PhosIDN: An integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. *Bioinformatics* **2021**, *37*, 4668–4676. [CrossRef]
17. Luo, F.; Wang, M.; Liu, Y.; Zhao, X.-M.; Li, A. DeepPhos: Prediction of protein phosphorylation sites with deep learning. *Bioinformatics* **2019**, *35*, 2766–2773. [CrossRef]
18. Liu, Y.; Li, A.; Zhao, X.M.; Wang, M. DeepTL-Ubi: A novel deep transfer learning method for effectively predicting ubiquitination sites of multiple species. *Methods* **2021**, *192*, 103–111. [CrossRef]
19. Wang, D.; Liang, Y.; Xu, D. Capsule network for protein post-translational modification site prediction. *Bioinformatics* **2018**, *35*, 2386–2394. [CrossRef]
20. Chen, Z.; Liu, X.; Li, F.; Li, C.; Marquez-Lago, T.; Leier, A.; Akutsu, T.; Webb, G.I.; Xu, D.; Smith, A.I.; et al. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief. Bioinform.* **2018**, *20*, 2267–2290. [CrossRef]
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *NIPS* **2012**, *60*, 84–90. [CrossRef]
22. Ng, H.-W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, DC, USA, 9 November 2015; pp. 443–449.
23. Tzeng, E.; Hoffman, J.; Darrell, T.; Saenko, K. Simultaneous Deep Transfer Across Domains and Tasks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
24. Rozantsev, A.; Salzmann, M.; Fua, P. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 801–814. [CrossRef] [PubMed]
25. Xu, H.; Zhou, J.; Lin, S.; Deng, W.; Zhang, Y.; Xue, Y. PLMD: An updated data resource of protein lysine modifications. *J. Genet. Genom.* **2017**, *44*, 243–250. [CrossRef] [PubMed]
26. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
27. Vacic, V.; Iakoucheva, L.; Radivojac, P. Two Sample Logo: A graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **2006**, *22*, 1536–1537. [CrossRef]
28. Deng, W.; Wang, C.; Zhang, Y.; Xu, Y.; Zhang, S.; Liu, Z.; Xue, Y. GPS-PAIL: Prediction of lysine acetyltransferase-specific modification sites from protein sequences. *Sci. Rep.* **2016**, *6*, 39787. [CrossRef]

29. Linding, R.; Jensen, L.J.; Ostheimer, G.J.; van Vugt, M.A.; Jørgensen, C.; Miron, I.M.; Diella, F.; Colwill, K.; Taylor, L.; Elder, K.; et al. Systematic Discovery of In Vivo Phosphorylation Networks. *Cell* **2007**, *129*, 1415–1426. [CrossRef]
30. Song, C.; Ye, M.; Liu, Z.; Cheng, H.; Jiang, X.; Han, G.; Songyang, Z.; Tan, Y.; Wang, H.; Ren, J.; et al. Systematic Analysis of Protein Phosphorylation Networks from Phosphoproteomic Data. *Mol. Cell. Proteom.* **2012**, *11*, 1070–1083. [CrossRef]
31. Liu, Y.; Wang, M.; Xi, J.; Luo, F.; Li, A. PTM-ssMP: A Web Server for Predicting Different Types of Post-translational Modification Sites Using Novel Site-specific Modification Profile. *Int. J. Biol. Sci.* **2018**, *14*, 946–956. [CrossRef]
32. Xi, J.; Wang, M.; Li, A. Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. *BMC Bioinform.* **2018**, *19*, 214. [CrossRef]