OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids

Cang-Zhi Jia, Wen-Ying He & Yu-Hua Yao

To cite this article: Cang-Zhi Jia, Wen-Ying He & Yu-Hua Yao (2016): OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2016.1163294

To link to this article: http://dx.doi.org/10.1080/07391102.2016.1163294

Accepted author version posted online: 09 Mar 2016.

Submit your article to this journal

Article views: 2

View related articles

View Crossmark data
Letter to the editor:

OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids

Cang-Zhi Jia\textsuperscript{1,*}, Wen-Ying He\textsuperscript{1}, Yu-Hua Yao\textsuperscript{2,*}

\textsuperscript{1}Department of Mathematics, Dalian Maritime University, No. 1 Linghai Road, Dalian 116026, China;
\textsuperscript{2}College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

1. Introduction

Hydroxylation of proline and lysine residues in proteins is one of the most abundant protein post-translational modification processes that is catalysed by three enzymes; prolyl 4-hidrolase, prolyl 3-hidrolase and lysyl hidrolase (Mitsuo & Marnisa, 2012).

In recent years, protein hydroxylation has been revealed to play vital physiological roles and its dysfunction leads to many diseases such as metabolic disorder, connective tissue disorder and cancer (Kelly & Ronald, 2010; Xie, 2007; Richards,
Because hydroxylation is a subtle post-translational modification, adding merely 16 atomic mass units to proteins, experimental identification and characterization of protein hydroxylation sites is often time consuming and expensive (Richards, 2006; Cockman, 2009; Webby, 2009). Hence, accurate computational prediction of protein hydroxylation sites represents a valuable and efficient approach to identify novel potential hydroxylation sites.

*Corresponding author. Email: cangzhijia@dlmu.edu.cn; yaoyuhua2288@163.com.

Several bioinformatics tools have been developed to predict protein hydroxylation sites. In 2009, Yang (2009) developed the first tool using a bio-kernel support vector machine (SVM) based on a limited dataset of 37 sequences. In 2010, Hu et al. (2010) developed the second tool using a nearest neighbour algorithm and the impact of physicochemical properties, biochemical properties and evolution information of amino acids on the performance were also considered. Xu et al. (2014) developed iHyd-PseAAC by using the dipeptide position-specific propensity into the general form of pseudo amino acid composition. More recently, Shi et al. (2015) built PredHydroxy by also using a SVM-based approach on the position-weighted amino acids composition. The PredHydroxy tool was found to be the best predictor by reaching an area under the receiver operating characteristic curve (AUC) of 0.827, and a Matthew’s correlation coefficient (MCC) of 0.690 for hydroxyproline and an AUC of 0.874 and a MCC of 0.667 for hydroxylysine in a jackknife cross-validation assessment.
Here, we developed a novel bioinformatics tool, called OH-PRED, to predict hydroxyproline and hydroxylysine sites separately by the adapted normal distribution bi-profile Bayes (ANBPB) feature extraction in combination with the physicochemical property indexes of the amino acids (AAPPI). Based on the results obtained by both jackknife and independent tests, OH-PRED significantly outperforms iHyd-PseAAC and PredHydroxy, and should be useful for the identification of protein hydroxylation sites.

2. Methods

2.1 Datasets

As a comprehensive and unbiased comparison with existing methods, the training datasets recently constructed in (Shi, 2015) were used. The protein sequences containing experimentally verified protein hydroxylation sites were collected from the UniProtKB/Swiss-Prot database (version 2014_1, www.uniprot.org). Total 265 candidate proteins containing hydroxylated prolines and 34 candidate proteins containing hydroxylated lysines were collected, respectively. Homology reductions within the benchmark datasets were performed with similarity threshold 70% between any two protein sequences. Then sequence segments around the hydroxylation sites and non-hydroxylation sites were extracted as positive and negative training datasets, respectively. After removing the identical sequence, the original datasets contain 659 positive sites and 3855 negative sites for hydroxyproline from 112 proteins, and 97 positive sites and 855 negative sites for hydroxylysine from 25 proteins. The size of the negative datasets is much larger (approximate ratio of 1:6) than that of the positive
training datasets, which will result in a bias prediction in favour of negative data.

Many previous approaches have been exploited to solve imbalanced machine learning issues, over-sampling, under-sampling and the voting method used (Chawla, 2011; Chou, 2006; Laurikkala, 2001; Zhang, 1992). We describe the set-up of the negative training datasets below.

Considering a typical protein hydroxylation problem: while the number of possible hydroxylation sites grows quadratically with the number of proteins, the number of positive hydroxylation sites grows typically only linearly (i.e. small fixed number of hydroxylation sites in one protein). So we can select those peptides no definitive hydroxylation information is available. But it is not possible to verify each possible site by experimental method. It has been universally acknowledged that the similarity of protein sequences to determine the function of proteins. Hence, we selected the least similarity peptide with the known hydroxylated peptide in one protein to construct the negative dataset. Firstly, we computed the similarity of one given hydroxylation peptide with other non-hydroxylation peptides within a protein. The BLOSUM62 scoring matrix was used to compute the similarity of protein peptides, and the peptide segment with the lowest score and the lowest three scores were chosen to construct the negative dataset, respectively. Finally, ratios of 1:1 and 1:3 of the number of positive samples and the number of negative samples were used to construct the negative training set, respectively. To save running time, the training dataset with 1:1 ratio was adopted to choose the optimal features. Meanwhile, it is applied to compare with PredHydroxy for in which the same number of positive and
negative samples. The ratio 1:3 was adopted to construct the optimal predictive model to reduce the false positive rate (Shao, 2009).

After several trials (results listed in Supplementary Table S1), the positive and negative peptides were formatted as 15-mer sequence peptides centred by hydroxylated proline and lysine residues.

2.2. Machine learning features

2.2.1. Bi-profile Bayes profile (BPB).

Given a peptide sequence \( S \), we encoded this sequence into a probability vector \( P = (p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n}) \), where \( p_i \) (\( i = 1, 2, \ldots, n \)) denotes the posterior probability of each amino acid at the \( i \)-th position in the positive samples and \( p_j \) (\( j = n+1, n+2, \ldots, 2n \)) denotes the posterior probability of each amino acid at the \( i \)-th position in the negative samples. The posterior probability of both positive and negative samples was calculated as the occurrence of each amino acid at each position in the training datasets (Shao, 2009).

2.2.2. Adapted normal distribution bi-profile Bayes (ANBPB)

ANBPB is a modified version of classical BPB. In this approach, the frequency of each amino acid at each position was encoded as random variables \( X_{ij} \), where \( i \) (\( i=1, 2, \ldots, 20 \)) represents the \( i \)-th amino acid

\[ \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\} \], and \( j=1, 2, \ldots, 15 \) represents the \( j \)-th position. The random variables \( X_{ij}, (i=1, 2, \ldots, 20; j=1, 2, \ldots, 15) \) are independent and obey the same binomial distribution \( b(n, p) \), where \( n=659 \) is the number of peptide sequences in the positive/negative set, and \( p=1/20 \) is the probability of each amino acid.
acid occurring in each position. According to the de Moivre-Laplace theorem, the
normal form variable \( \frac{X_{ij} - np}{\sqrt{np(1-p)}} \) has a limiting cumulative distribution function that
approximates a normal distribution \( N(0,1) \). Here, we modified the standard variable
normalization to highlight and emphasize the distinct distribution of each amino acid
at one position. We let \( V_j \) denote the standard variance of \( X_{i,j} \) \((i=1,2,\ldots,20)\), i.e., the
deviation of frequencies of each amino acid at the same \( j \)th position. Then we define
\[
X'_{ij} = \frac{X_{ij} - np}{\sqrt{V_j}}
\]
as the new normalization of \( X_{ij} \) and deem it obeys the standard normal
distribution. Thus, the posterior probability \( p_j \) \((j=1,2,\ldots,2n)\) was coded by the adapted
normal distribution as follows: \( p_j = P(X \leq X'_{ij}) = \phi(X'_{ij}) \), where \( \phi(x) \) is the standard
normal distribution function given by \( \phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \). For more details about
the ANBPB method, please refer to the original paper (Jia, 2013).

2.2.3. Physicochemical property indexes of the amino acids (AAPPI)

Thirteen physicochemical features selected from the amino acid index (AAindex,
http://www.genome.ad.jp/aaindex/) database (Kawashima & Kanehisa, 2000;
Kawashima, 2008) were used to encode each amino acid residue in a data instance.

Detail information of the properties, corresponding accession numbers and the
abbreviations are listed in Supplementary Table S2. The values of each amino acid for
each physicochemical property are listed in Supplementary Table S3.

2.2.4. SVM implementation and performance evaluation

The SVM classification method has proven to be powerful in many fields of
bioinformatics (Folkman, 2016; Jia, 2013; Lin, 2014; Liu, 2014; Qiu, 2015; Shao,
2009; Shi, 2015; Xu, 2015). In this work, the SVM was trained with the LIBSVM package (version 3.0) (Chang & Lin, 2011) to build the model and perform the predictions. The radial basis kernel function \( k(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\} \) was selected and the parameters \( c = 4, \gamma = 0.25 \) for the hydroxyproline prediction and \( c = 4, \gamma = 0.125 \) for the hydroxylysine prediction) optimized by the SVMcgForClass program were downloaded from http://www.matlabsky.com.

The jackknife test is deemed as the least arbitrary test that can always yield a unique outcome for a given benchmark dataset (Chou, 2013). Thus, we used the jackknife test to select important features and optimize all parameters. In comparison with other methods, both the jackknife test and independent dataset test were used.

We also assessed the overall prediction performance in terms of the receiver characteristic (ROC) curves. An ROC curve plots the true positive rate (sensitivity) as a function of the false positive rate (1-specificity) at different prediction thresholds. Furthermore, we calculated sensitivity \( (Sn) \), specificity \( (Sp) \), accuracy \( (Acc) \) and \( MCC \), which were defined as follows:

\[
Sn = \frac{TP}{TP + FN} \quad (1)
\]

\[
Sp = \frac{TN}{TN + FP} \quad (2)
\]

\[
Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)
\]

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (4)
\]

3. Results and discussion
3.1. Prediction of protein hydroxylation sites using only BPB and ANBPB

BPB was first proposed by Shao et al. for predicting protein methylation sites (Shao, 2009). One advantage of this method is that the feature vectors are encoded in a bi-profile manner, which contains information from positive and negative samples. ANBPB is a modified version of the classic BPB, which is more powerful than BPB for predicting protein O-GlcNAcylation sites (Jia, 2013) and protein S-nitrosylation sites (Jia, 2014). In this study, the ability of BPB and ANBPB to discriminate between protein hydroxylation sites and non-hydroxylation sites was first compared by the jackknife test (Table 1). For hydroxyproline, the BPB model reached a Sn of 87.56%, a Sp of 91.18%, an Acc of 89.35% and a MCC of 0.788, while the ANBPB model reached a Sn of 88.62%, a Sp of 91.49%, an Acc of 90.04% and a MCC of 0.801. For hydroxylysine, the BPB model reached a Sn of 86.60%, a Sp of 97.94%, an Acc of 92.27% and a MCC of 0.851, while the ANBPB model reached a Sn of 92.78%, a Sp of 98.97%, an Acc of 95.88% and a MCC of 0.919.

These results demonstrated that the ANBPB model performs better than the BPB model in both hydroxyproline and hydroxylysine predictions. Therefore, further optimization of the predictive model was based on the ANBPB feature extraction.

3.2. Improving predictive performance by incorporating AAPPI

Because the AAPPIs surrounding the candidate hydroxylation sites obviously affect the recognition and catalytic efficiency of protein hydroxylases, incorporating physicochemical information with ANBPB might improve the accuracy of the prediction model. Thirteen representative physicochemical property indexes were
selected from the AAindex database (Kawashima & Kanehisa, 2000; Kawashima, 2008): refractivity (AA1), flexibility (AA2), volume (AA3), transfer free energy to surface (AA4), electron-ion interaction potential values (AA5), hydrophilicity (AA6), polarity (AA7), hydrophobicity (AA8), isoelectric point (AA9), the optimized transfer energy parameter (AA10), the optimized side chain interaction parameter (AA11), residue volume (AA12) and the normalized van der Waals volume (AA13) (Supplementary Table S2).

Initially, we evaluated the predictive performances of ANBPB combined with one of the 13 physicochemical properties on jackknife cross validation (Table S4 for hydroxyproline prediction and Supplementary Table S5 for hydroxylysine prediction). The combinations with improved prediction accuracy are listed in Table 2. For the hydroxyproline prediction, the ANBPB+AA1 model achieved the best prediction accuracy, followed by ANBPB+AA5 and ANBPB+AA11 (Table 2). For the hydroxylysine prediction, the ANBPB+AA12 model achieved the best prediction Acc of 97.42%, followed by the second-best prediction Acc of 96.91% achieved by ANBPB+AA3, ANBPB+AA6 and ANBPB+AA7 (Table2). These results demonstrated that the performance of the prediction model can be increased by combining ANBPB and AAPPI. Moreover, different AAPPIs were found to contribute positively to the hydroxyproline and hydroxylysine predictions. Refractivity, electron-ion interaction potential values and the optimized side chain interaction parameter specifically increased the prediction performance for hydroxyproline, whereas volume, hydrophilicity and polarity specifically increased the prediction
performance for hydroxylysine. Because proline and lysine hydroxylation are catalysed by prolyl hydrolase and lysyl hydrolase (Kelly & Ronald, 2010; Xie, 2007), respectively, these results provide additional information that describes the substrate specificity of these enzymes.

We then evaluated the predictive performances of ANBPB combined with two or three improved-performing AAPPIs, and the results of these combinations are shown in Tables S4 and S5. The combination results that improved the accuracy are also listed in Table 2. For the hydroxyproline prediction, the best-performing combination of ANBPB+AA1+AA5+AA11 model reached a Sn of 91.20%, a Sp of 92.57%, an Acc of 91.88% and a MCC of 0.838. However, no improvements of the prediction accuracy by combining ANBPB with two improved-performing AAPPIs for the hydroxylysine prediction were observed.

Then the contributions of these physicochemical properties were quantified by using the average value of F-score measurement (Ward-Powers, 2011). The high F-score values mean there are significant differences between hydroxylated and non-hydroxylated sites. Supplementary Fig. S1 shows the average value of F-score measurement on AA1, AA5 and AA11 for hydroxyproline. These results demonstrated there are significant differences at positions 3, 6, 8, 11, 14 for AA1, at positions 3, 6, 10, 13 for AA5, and at positions 2, 8, 10, 13 for AA11. Supplementary Fig. S2 shows the average value of F-score measurement on AA12 for hydroxylysine, which reveals that there are significant difference at positions 3, 6, 8, 11 and 14.

To this end, the SVM-based predictor, OH-PRED, was built using the
ANBPB+AA1+AA5+AA11 feature extraction method based on the SVM classifier with the RBF kernel function (cost parameters $c = 4$, $\gamma = 0.25$) to predict hydroxyproline sites and the ANBPB+AA12 feature extraction method based on the SVM classifier with the RBF kernel function (cost parameters $c = 4$, $\gamma = 0.125$) to predict the hydroxylysine sites.

To further evaluate the predictive performance of OH-PRED, the receiver operating characteristic (ROC) curves were plotted and the area under the ROC curve (AUC) were calculated. For hydroxyproline, the ANBPB model reached an AUC of 0.957 and the ANBPB+AA1+AA11+AA5 model reached an AUC of 0.973 (Figure 1). For hydroxylysine, the ANBPB model reached an AUC of 0.991 and the ANBPB+AA12 model reached an AUC of 0.996 (Figure 2). The figure indicates that a multi-feature model is more efficient than a single-feature classification.

3.3. Predictive performance of our model

In order to avoid overestimation of the models, we further conducted the jackknife 4-fold, 6-fold and 8-fold cross-validations on the imbalanced benchmark dataset. The results are listed in Supplementary Table S6. These results achieved by different cross-validations were approximate, and further demonstrated the robust and reliable of the predictor OH-PRED.

We compared the predictive performance of OH-PRED with two available methods, iHyd-PseAAC (Xu, 2014) and PredHydroxy (Shi, 2015). OH-PRED and PredHydroxy were first compared using the jackknife test with an identical training dataset (Table 3). For the hydroxyproline prediction, OH-PRED reached an ACC of
91.88% and a MCC of 0.838, which were 7.37% and 0.148 higher than that of PredHydroxy, respectively. For the hydroxylysine prediction, OH-PRED reached an accuracy of 97.42% and a MCC of 0.949, which were 10.09% and 0.282 higher than that of PredHydroxy, respectively.

We also tested OH-PRED and iHyd-PseAAC on the same imbalance datasets given as Supplementary Information S3 and S4 in (Xu, 2014). In order to reduce the data imbalance effect on prediction performance, we set weight parameter \( w_1 = 3 \) for hydroxyproline and \( w_1 = 1.8 \) for hydroxylysine, respectively. OH-PRED outperformed iHyd-PseAAC with Acc improvements of 2.51% and 12.65% for hydroxyproline and hydroxylysine, respectively (Supplementary Table S7).

To further evaluate our prediction model, we randomly split 10% of the samples from the dataset as an independent test dataset, and the remaining 90% of the samples as a training dataset and then evaluated the performance of the prediction. This approach was repeated three times and the average predictive performance is listed in Table 4. For the hydroxyproline prediction, OH-PRED outperformed PredHydroxy and iHyd-PseAAC by an accuracy improvement of 1.70% and 12.81%, respectively. For the hydroxylysine prediction, OH-PRED outperformed PredHydroxy and iHyd-PseAAC by an accuracy improvement of 3.24% and 7.41%, respectively.

For an independent test, the dataset contains 38 positive sites for hydroxyproline from 20 proteins, and 34 positive sites for hydroxylysine from 6 proteins (Supplementary Materials S1 and S2). The performances of iHyd-PseAAC, PredHydroxy and OH-PRED against this dataset are summarized in Table 5.
OH-PRED was revealed to be the best predictor for hydroxyproline prediction with a Sn of 84.21% and a Sp of 91.42%. As for the hydroxylysine prediction, the OH-PRED could achieve the highest Sn among three models but the lowest Sp of 54.94%. These results demonstrated OH-PRED is a powerful tool for predicting protein hydroxyproline sites but is still waiting to be improved for protein hydroxylysine prediction, especially for the specificity. We think the increments of the limited training data will be very helpful.

4. Conclusions

OH-PRED was developed in this report to predict protein hydroxylation sites by using the ANBPB feature extraction and AAPPI, which outperforms previous methods based on the results obtained by both jackknife and independent tests. OH-PRED should be a powerful tool for in silico identification of protein hydroxylation sites and help to reveal their exact molecular mechanisms in physiological and pathological processes. The MATLAB package of OH-PRED is available as Supplementary files.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities under grant [number 3132014324, 3132015159] and the Scientific Research Plan of the Department of Education of Liaoning Province under grant [L2014200].

References

Chang, C.C. & Lin, C.J. (2011). LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2, 27-27. Available at
http://www.csie.ntu.edu.tw/~cjlin/libsvm. (Accessed: 4th July 2015)

Chawla, N.V. et al. (2011). SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res, 16*, 321-357.

Chou, K.C. & Shen, H.B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J Proteome Res, 5*, 1888-1897.

Chou, K.C. & Shen, H.B. (2013). Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat Sci, 2*, 1090-1103.

Cockman, M.E. et al. (2009). Proteomics-based identification of novel factor inhibiting hypoxia-inducible factor (FIH) substrates indicates widespread asparaginyl hydroxylation of ankyrin repeat domain-containing proteins. *Mol Cell Proteomics, 8*, 535-546.

David, M.H. & David, R.E. (2013). Collagen prolyl3-hydroxylation: a major role for a minor posttranslational modification?. *Connect Tissue Res., 54*, 245-251.

Hu, L.L. et al. (2010). Prediction and analysis of protein hydroxyproline and hydroxylsine. *PLoS One 5*, e15917.

Folkman, L. et al. (2016). EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J Mol Biol*, in press.

Jia, C.Z. et al. (2013). O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Mol BioSyst, 9*, 2909-2913.
Jia, C.Z. et al. (2014). Prediction of protein S-Nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou’s Pseudo amino acid composition. *Int J Mol Sci*, 15, 10410-10423.

Kawashima, S. & Kanehisa, M. (2000). AAindex: amino acid index database. *Nucleic Acids Res*, 28, 374.

Kawashima, S. et al. (2008). AAindex: amino acid index database, progress report. *Nucleic Acids Res*, 36, D202-205.

Kelly, L.G. & Ronald, T.R. (2010). Prolyl 4-hydroxylase. *Crit Rev Biochem Mol Biol.*, 45, 106-124.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution (eds Silvana, Q. et al.) 63-66 (Springer-Verlag).

Lin, H. *et al.* (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, *Nucleic Acids Res*, 42, 12961-12972.

Liu, B. *et al.* (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, *Bioinformatics*, 30, 472-479.

Mitsuo, Y. & Marnisa, S. (2012). Lysine post-translational modifications of collagen. *Essays Biochem.*, 52, 113-133.

Qiu, W. R. *et al.* (2015). iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, *J. Biomol. Struct. Dyn.*, 33, 1731-1742.
Richards, A.A. et al. (2006). Adiponectin multimerization is dependent on conserved lysines in the collagenous domain: evidence for regulation of multimerization by alterations in posttranslational modifications. *Mol Endocrinol*, 20, 1673-1687.

Shao, J. et al. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One*, 4, e4920.

Shi, S.P. et al. (2015). PredHydroxy: computational prediction of protein hydroxylation site locations based on the primary structure. *Mol BioSyst*, 11, 819-825.

Xie, H. et al. (2007). Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res.*, 6, 1917-1932.

Xu, Y. et al. (2014). iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int J Mol Sci.*, 15, 7594-7610.

Xu, Y. et al. (2015). iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position specific propensity. *Scientific Report*, 5, 10184.

Yang, Z.R. (2009). Predict collagen hydroxyproline sites using support vector machines. *J Comput Biol: a journal of computational molecular cell biology*, 16, 691-702.

Zhang, C.T. (1992). Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophys J*, 63, 1523-1529.

Ward-Powers, D.M. (2011). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *J Mach Learn Res*, 2, 37-63.
Webby, C.J. et al. (2009). Jmd6 catalyses lysyl-hydroxylation of U2AF65, a protein associated with RNA splicing. *Science* 325, 90-93.
Table 1. Jackknife test performances of BPB and ANBPB.

| Methods | Residue type | Sn (%) | Sp (%) | Acc (%) | MCC  |
|---------|--------------|--------|--------|---------|------|
| BPB     | Proline      | 87.56  | 91.18  | 89.35   | 0.788|
|         | Lysine       | 87.63  | 95.88  | 91.75   | 0.838|
| ANBPB   | Proline      | 88.62  | 91.49  | 90.04   | 0.801|
|         | Lysine       | 92.78  | 98.97  | 95.88   | 0.919|
Table 2. Jackknife test of different sequences encoding combinations with improved performances.

| Residue type | Features        | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------|-----------------|--------|--------|---------|------|
| Proline      | ANBPB+AA1       | 89.98  | 92.26  | 91.11   | 0.823|
|              | ANBPB+AA2       | 88.77  | 91.49  | 90.11   | 0.803|
|              | ANBPB+AA5       | 89.38  | 91.80  | 90.57   | 0.812|
|              | ANBPB+AA9       | 89.07  | 91.49  | 90.27   | 0.806|
|              | ANBPB+AA10      | 89.23  | 91.80  | 90.50   | 0.810|
|              | ANBPB+AA11      | 89.98  | 92.11  | 91.03   | 0.821|
|              | ANBPB+AA1+AA5   | 90.74  | 92.11  | 91.42   | 0.829|
|              | ANBPB+AA1+AA11  | 90.59  | 92.41  | 91.49   | 0.830|
|              | ANBPB+AA1+AA5+AA11 | 91.20 | 92.57 | 91.88   | 0.838|
| Lysine       | ANBPB+AA1       | 93.81  | 98.97  | 96.39   | 0.929|
|              | ANBPB+AA3       | 94.85  | 98.97  | 96.91   | 0.939|
|              | ANBPB+AA4       | 93.81  | 98.97  | 96.39   | 0.929|
|              | ANBPB+AA6       | 94.85  | 98.97  | 96.91   | 0.939|
|              | ANBPB+AA7       | 94.85  | 98.97  | 96.91   | 0.939|
|              | ANBPB+AA12      | 95.88  | 98.97  | 97.42   | 0.949|
|              | ANBPB+AA13      | 94.85  | 97.94  | 96.39   | 0.928|
Table 3. Performance of our method and PredHydroxy on the jackknife test.

| Residue type | Method      | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------|-------------|--------|--------|---------|------|
| Proline      | PredHydroxy | 83.78  | 85.24  | 84.51   | 0.690|
|              | OH-PRED     | 91.20  | 92.57  | 91.88   | 0.838|
| Lysine       | PredHydroxy | 84.21  | 82.46  | 83.33   | 0.667|
|              | OH-PRED     | 95.88  | 98.97  | 97.42   | 0.949|

Table 4. Performance of our method, iHyd-PseAAC and PredHydroxy for hydroxylation.

| Residue type | Method        | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------|---------------|--------|--------|---------|------|
| Proline      | iHyd-PseAAC*  | 81.48  | 79.63  | 80.56   | 0.611|
|              | PredHydroxy*  | 96.30  | 87.04  | 91.67   | 0.837|
|              | OH-PRED       | 92.99  | 93.94  | 93.37   | 0.866|
| Lysine       | iHyd-PseAAC*  | 66.67  | 100    | 87.50   | 0.776|
|              | PredHydroxy*  | 83.33  | 100    | 91.67   | 0.845|
|              | OH-PRED       | 93.27  | 96.67  | 94.91   | 0.902|

*The data is directly obtained from (Shi, 2015).
Table 5. Performances of OH-PRED, iHyd-PseAAC and PredHydroxy on 20 hydroxyproline proteins and 6 hydroxylysine proteins which are not included in our training dataset

| Residue type | Method        | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------|---------------|--------|--------|---------|------|
| Proline      | iHyd-PseAAC*  | 78.57  | 91.81  | 90.61   | 0.573|
|              | PredHydroxy   | 80.65  | 91.51  | 90.40   | 0.599|
|              | OH-PRED       | 84.21  | 91.42  | 90.52   | 0.650|
| Lysine       | iHyd-PseAAC*  | 93.75  | 61.17  | 65.91   | 0.388|
|              | PredHydroxy*  | 92.86  | 60.11  | 64.45   | 0.360|
|              | OH-PRED       | 94.12  | 54.94  | 61.73   | 0.373|

*It should be noted the 20 hydroxyproline proteins and 6 hydroxylysine proteins may be included in the training dataset of iHyd-PseAAC.
Figures legends:

**Figure 1.** The ROC curves to assess the predictive performance based on different sequences encoding schemes for hydroxyproline.

**Figure 2.** The ROC curves to assess the predictive performance based on different sequences encoding schemes for hydroxyllysine.
OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids

Cang-Zhi Jia\textsuperscript{1,\ast}, Wen-Ying He\textsuperscript{1}, Yu-Hua Yao\textsuperscript{2,\ast}

\textsuperscript{1}Department of Mathematics, Dalian Maritime University, No. 1 Linghai Road, Dalian 116026, China;
\textsuperscript{2}College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

\ast To whom correspondence should be addressed (email:cangzhijia@dlmu.edu.cn;yaoyuhua2288@163.com).
The BLOSUM62 scoring matrix used in the study

|   | C   | S   | I   | P   | A   | G   | N   | D   | E   | Q   | H   | R   | K   | M   | I   | L   | V   | F   | Y   | W   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| C | 9   | -1  | -1  | -3  | 0   | -3  | -3  | -4  | -3  | -3  | -3  | -1  | -1  | -1  | -2  | -2  | -2  | -2  | -2  | -2  |
| S | -1  | 4   | 1   | -1  | 1   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | -1  | -2  | -2  | -2  | -3  | -2  | -2  | -2  | -2  | -3  |
| I | -1  | 1   | 5   | -1  | 0   | -2  | 0   | -1  | -1  | -1  | -2  | -1  | -1  | -1  | -1  | 0   | -2  | -2  | -2  | -2  | -2  | -3  |
| P | -3  | -1  | -1  | 7   | -1  | -2  | -2  | -1  | -1  | -1  | -2  | -2  | -1  | -2  | -3  | -3  | -2  | -3  | -2  | -4  | -3  | -4  |
| A | 0   | 1   | 0   | 1   | 4   | 0   | -2  | -2  | -1  | -1  | -2  | -1  | -1  | -1  | -1  | 0   | -2  | -2  | -3  | -3  | -3  | -3  |
| G | -3  | 0   | -2  | 2   | 0   | 6   | 0   | -1  | -2  | -2  | -2  | -2  | -3  | -4  | -3  | -3  | -3  | -2  | -3  | -3  | -2  | -4  |
| N | -3  | 1   | 0   | -2  | -2  | 0   | 6   | 1   | 0   | 0   | 1   | 0   | 0   | -2  | -3  | -3  | -3  | -3  | -3  | -3  | -3  | -4  |
| D | -3  | 0   | -1  | -1  | -2  | 1   | 6   | 2   | 0   | -1  | -1  | -3  | -3  | -4  | -3  | -3  | -3  | -3  | -3  | -3  | -3  | -4  |
| E | -4  | 0   | -1  | -1  | -2  | 0   | 2   | 5   | 2   | 0   | 0   | 1   | -2  | -3  | -3  | -3  | -3  | -2  | -3  | -2  | -3  | -3  |
| Q | -3  | 0   | -1  | -1  | -2  | 0   | 2   | 5   | 0   | 1   | 1   | 0   | -3  | -2  | -2  | -3  | -1  | -2  | -2  | -3  | -1  | -2  |
| H | -3  | -1  | -2  | -2  | -2  | 1   | 1   | 0   | 0   | 0   | 0   | 1   | -2  | -3  | -3  | -3  | -1  | 2   | 2   | -2  | -3  | -2  |
| R | -3  | -1  | -1  | -1  | -2  | 0   | -2  | 0   | 1   | 0   | 5   | 2   | -1  | -3  | -2  | -3  | -3  | 2   | 2   | -3  | -3  | -2  |
| K | -3  | 0   | -1  | -1  | -1  | -2  | 0   | -1  | 1   | 1   | 1   | 1   | 1   | 2   | 5   | -3  | -2  | -2  | -3  | -2  | -3  | -3  |
| M | -1  | -1  | -1  | -2  | -1  | -3  | -2  | -3  | -2  | 0   | -2  | -1  | -1  | 5   | 1   | 2   | 1   | 0   | -1  | -1  | -1  | -3  |
| I | -1  | -2  | -1  | -3  | -4  | -3  | -3  | -3  | -3  | -3  | -3  | -3  | -2  | -3  | -4  | 2   | 3   | 0   | -1  | -3  | -3  | -3  |
| L | -1  | -2  | -1  | -3  | -4  | -3  | -4  | -3  | -2  | -2  | -2  | -2  | -2  | -2  | -2  | -2  | 2   | 4   | 1   | 0   | -1  | -2  |
| V | -1  | -2  | 0   | -2  | 0   | -3  | -3  | -3  | -2  | -2  | -3  | -2  | -2  | -3  | -2  | 1   | 3   | 1   | 4   | -1  | -1  | -3  |
| F | -2  | -2  | -2  | -4  | -2  | -3  | -3  | -3  | -3  | -3  | -3  | -1  | -3  | -3  | 0   | 0   | 0   | -1  | 6   | 3   | 1   |
| Y | -2  | -2  | -2  | -3  | -2  | -3  | -2  | -3  | -2  | -1  | -1  | -1  | -1  | 1   | 3   | 7   | 2   | -2  | -2  | -2  | -3  |
| W | -2  | -3  | -2  | -4  | -3  | -2  | -3  | -3  | -3  | -2  | -3  | -3  | -1  | -3  | -2  | -3  | 1   | 2   | 11  | -3  | -3  | -3  |
**Table S1.** The jackknife results of models using ANBPB on different window sizes.

| Residue type | Length | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------|--------|--------|--------|---------|------|
| Proline      | 17     | 88.47  | 91.18  | 89.81   | 0.797|
|              | 15     | **88.62** | **91.49** | **90.04** | **0.801** |
|              | 13     | 89.07  | 89.47  | 89.27   | 0.785|
|              | 17     | 91.75  | 92.78  | 92.27   | 0.865|
| Lysine       | 15     | **92.78** | **98.97** | **95.88** | **0.919** |
|              | 13     | 90.72  | 1      | 95.36   | 0.911|

**Table S2.** The 13 relevant physicochemical properties selected in this paper.

| Property                                      | AAindex    | Abbrevation |
|-----------------------------------------------|------------|-------------|
| Refractivity                                  | MCMT640101 | AA1         |
| Flexibility                                   | BHAR880101 | AA2         |
| Volume                                        | CHOC750101 | AA3         |
| Transfer free energy to surface               | BULH740101 | AA4         |
| Electron-ion interaction potential values     | COSI940101 | AA5         |
| Hydrophility                                  | HOPT810101 | AA6         |
| Polarity                                      | CHAM820101 | AA7         |
| Hydrophobicity                                | EISD840101 | AA8         |
| Isoelectric point                             | ZIMJ680104 | AA9         |
| Optimized transfer energy parameter           | OOBM850103 | AA10        |
| Optimized side chain interaction parameter    | OOBM850105 | AA11        |
| Residue volume                                | BIGC670101 | AA12        |
| Normalized van der Waals volume               | FAUJ880103 | AA13        |
Table S3. The values corresponding to each amino acid for 13 physicochemical properties.

| Name of the indices AA1 | AA2 | AA3 | AA4 | AA5 | AA6 | AA7 | AA8 | AA9 | AA10 | AA11 | AA12 | AA13 |
|-------------------------|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|------|
| Amino acids             |     |     |     |     |     |     |     |     |      |      |      |      |
| AA1                     | 4.34| 0.357| 91.5|-0.2| 0.0373| -0.5| 0.046| 0.25| 6.046| 1.16 | 52.6 | 1     |
| AA2                     | 26.66| 0.529| 202|-0.12| 0.0959| 3.0 | 0.291| -1.76| 10.76| -1.54| 1.72 | 109.1 |
| AA3                     | 13.28| 0.463| 135.2|0.08| 0.0036| 0.2 | 0.134| -0.64| 5.41 | 1.31 | 1.97 | 75.7 |
| AA4                     | 12.0 | 0.511| 124.5|-0.2| 0.1263| 3.0 | 0.105| -0.72| 2.77 | -0.33| 2.66 | 68.4 |
| AA5                     | 35.77| 0.346| 117.7|-0.45| 0.0829| -1.0 | 0.128| 0.04 | 5.05 | 0.2  | 0.5  | 68.3 |
| AA6                     | 17.56| 0.493| 161.1|0.16| 0.0761| 0.2 | 0.18 | -0.69| 5.65 | -1.12| 3.87 | 89.7 |
| AA7                     | 17.26| 0.497| 155.1|-0.3| 0.0058| 3.0 | 0.151| -0.62| 3.22 | 0.48 | 2.4  | 84.7 |
| AA8                     | 0.0 | 0.544| 66.4| 0  | 0.005| 0    | 0   | 0.16| 5.97 | 0.64 | 1.63 | 36.3 |
| AA9                     | 0.0 | 0.323| 167.3|-0.12| 0.0242| -0.5| 0.23 | -0.4| 7.59 | -1.31| 0.86 | 91.9 |
| AA10                    | 19.06| 0.462| 168.8|-2.26| 0     | -1.8| 0.186| 0.73| 6.02 | 3.28 | 0.57 | 102  |
| AA11                    | 18.78| 0.365| 167.9|-2.46| 0     | -1.8| 0.186| 0.53| 5.98 | 0.43 | 0.51 | 102  |
| AA12                    | 21.29| 0.466| 171.3|-0.35| 0.0371| 3.0 | 0.219| -1.1| 9.74 | -1.71| 3.9  | 105.1|
| AA13                    | 21.64| 0.295| 170.8|-1.47| 0.0823| -1.3| 0.221| 0.26| 5.74 | 0.15 | 0.4  | 97.7 |
| Amino acids             |     |     |     |     |     |     |     |     |      |      |      |      |
| and values              |     |     |     |     |     |     |     |     |      |      |      |      |
| AA1                     | 29.4 | 0.314| 203.4|-2.33| 0.0946| -2.5| 0.29 | 0.61| 5.48 | 0.52 | 0.43 | 113.9|
| AA2                     | 10.93| 0.509| 129.3|-0.98| 0.0198| 0   | 0.131| -0.07| 6.3  | -0.58| 2.04 | 73.6 |
| AA3                     | 6.35 | 0.507| 99.1|-0.39| 0.0829| 0.3 | 0.062| -0.26| 5.68 | -0.83| 1.61 | 54.9 |
| AA4                     | 11.01| 0.444| 122.1|-0.52| 0.0941| -0.4| 0.108| -0.18| 5.66 | -1.52| 1.48 | 71.2 |
| AA5                     | 42.53| 0.305| 237.6|-2.01| 0.0548| -3.4| 0.409| 0.37| 5.89 | 1.25 | 0.75 | 135.4|
| AA6                     | 31.53| 0.42 | 203.6|-2.24| 0.0516| -2.3| 0.298| 0.02| 5.66 | -2.21| 1.72 | 116.2|
| AA7                     | 13.92| 0.386| 141.7|-1.56| 0.0057| -1.5| 0.14 | 0.54| 5.96 | 0.54 | 0.59 | 85.1 |
| AA8                     | 0    | 0    | 0    |     |     |     |     |     |      |      |      |      |
| AA9                     | 0    | 0    | 0    |     |     |     |     |     |      |      |      |      |
| AA10                    | 0    | 0    | 0    |     |     |     |     |     |      |      |      |      |
| AA11                    | 0    | 0    | 0    |     |     |     |     |     |      |      |      |      |
| AA12                    | 0    | 0    | 0    |     |     |     |     |     |      |      |      |      |
| AA13                    | 0    | 0    | 0    |     |     |     |     |     |      |      |      |      |
Table S4. The performance of models trained with various features combination for hydroxyproline in jackknife test.

| Features               | Sn (%)  | Sp (%)  | Acc (%) | MCC    |
|------------------------|---------|---------|---------|--------|
| ANBPB                  | 88.62   | 91.49   | 90.04   | 0.801  |
| ANBPB+AA1              | 89.98   | 92.26   | 91.11   | 0.823  |
| ANBPB+AA2              | 88.77   | 91.49   | 90.11   | 0.803  |
| ANBPB+AA3              | 88.62   | 90.87   | 89.73   | 0.795  |
| ANBPB+AA4              | 88.77   | 90.71   | 89.73   | 0.795  |
| ANBPB+AA5              | 89.38   | 91.80   | 90.57   | 0.812  |
| ANBPB+AA6              | 88.47   | 90.71   | 89.58   | 0.792  |
| ANBPB+AA7              | 86.80   | 89.63   | 88.20   | 0.764  |
| ANBPB+AA8              | 89.07   | 91.02   | 90.04   | 0.801  |
| ANBPB+AA9              | 89.07   | 91.49   | 90.27   | 0.806  |
| ANBPB+AA10             | 89.23   | 91.80   | 90.50   | 0.810  |
| ANBPB+AA11             | 89.98   | 92.11   | 91.03   | 0.821  |
| ANBPB+AA12             | 89.23   | 90.56   | 89.89   | 0.798  |
| ANBPB+AA13             | 88.16   | 90.56   | 89.35   | 0.787  |
| ANBPB +AA1+AA2         | 87.71   | 91.33   | 89.50   | 0.791  |
| ANBPB +AA1+AA5         | 90.74   | 92.11   | 91.42   | 0.829  |
| ANBPB +AA1+AA9         | 88.16   | 92.11   | 90.11   | 0.803  |
| ANBPB +AA1+AA10        | 88.77   | 91.64   | 90.19   | 0.804  |
| ANBPB +AA1+AA11        | 90.59   | 92.41   | 91.49   | 0.830  |
| ANBPB +AA1+AA5+AA11    | 91.20   | 92.57   | 91.88   | 0.838  |
| Features           | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------------|--------|--------|---------|------|
| ANBPB             | 92.78  | 98.97  | 95.88   | 0.919 |
| ANBPB+AA1         | 93.81  | 98.97  | 96.39   | 0.929 |
| ANBPB+AA2         | 94.85  | 96.91  | 95.88   | 0.918 |
| ANBPB+AA3         | 94.85  | 98.97  | 96.91   | 0.939 |
| ANBPB+AA4         | 93.81  | 98.97  | 96.39   | 0.929 |
| ANBPB+AA5         | 92.78  | 96.91  | 94.85   | 0.898 |
| ANBPB+AA6         | 94.85  | 98.97  | 96.91   | 0.939 |
| ANBPB+AA7         | 94.85  | 98.97  | 96.91   | 0.939 |
| ANBPB+AA8         | 92.78  | 97.94  | 95.36   | 0.908 |
| ANBPB+AA9         | 91.75  | 95.88  | 93.81   | 0.877 |
| ANBPB+AA10        | 94.85  | 96.91  | 95.88   | 0.918 |
| ANBPB+AA11        | 94.85  | 95.88  | 95.36   | 0.907 |
| ANBPB+AA12        | 95.88  | 98.97  | 97.42   | 0.949 |
| ANBPB+AA13        | 94.85  | 97.94  | 96.39   | 0.928 |
| ANBPB+AA12+AA1    | 93.81  | 98.97  | 96.39   | 0.929 |
| ANBPB+AA12+AA3    | 93.81  | 98.97  | 96.39   | 0.929 |
| ANBPB+AA12+AA4    | 9278   | 9897   | 9588    | 0.919 |
| ANBPB+AA12+AA6    | 93.81  | 98.97  | 96.39   | 0.929 |
| ANBPB+AA12+AA7    | 9485   | 98.97  | 96.91   | 0.939 |
| ANBPB+AA12+AA13   | 9381   | 98.97  | 96.39   | 0.929 |
Table S6. The jackknife, 4-fold, 6-fold, 8-fold and 10-fold cross-validation results by the predictor on the unbalanced dataset.

| Residue type | Cross-validation | Sn (%) | Sp (%) | Acc (%) | MCC  |
|--------------|------------------|--------|--------|---------|------|
| Proline      | 4-fold           | 80.95  | 96.73  | 92.56   | 0.804|
|              | 6-fold           | 80.40  | 96.58  | 92.41   | 0.798|
|              | 8-fold           | 80.80  | 96.98  | 92.72   | 0.808|
|              | 10-fold          | 81.23  | 97.32  | 93.04   | 0.817|
|              | jackknife        | 83.46  | 95.26  | 92.17   | 0.796|
| Lysine       | 4-fold           | 90.04  | 98.58  | 96.34   | 0.905|
|              | 6-fold           | 92.03  | 99.29  | 97.39   | 0.931|
|              | 8-fold           | 90.05  | 98.21  | 96.08   | 0.901|
|              | 10-fold          | 90.48  | 98.32  | 95.81   | 0.897|
|              | jackknife        | 90.72  | 98.60  | 96.60   | 0.909|
Table S7. The jackknife test performances of OH-PRED and iHyd-PseAAC on the benchmark datasets in (Xu, 2014).

| Residue type | Method       | Sn(%)  | Sp(%)  | Acc(%)  | MCC  |
|--------------|--------------|--------|--------|---------|------|
| Proline      | iHyd-PseAAC* | 70.68  | 89.03  | 78.42   | 0.52 |
|              | OH-PRED      | 75.82  | 82.32  | 80.93   | 0.52 |
| Lysine       | iHyd-PseAAC* | 79.04  | 86.37  | 83.12   | 0.51 |
|              | OH-PRED      | 90.70  | 96.19  | 95.77   | 0.75 |

*The data is directly obtained from (Xu, 2014).
Figure S1. The average values of F-score measurement on AA1, AA5 and AA11 properties for hydroxyproline.
Figure S2. The average value of F-score measurement on AA12 property for hydroxylysine