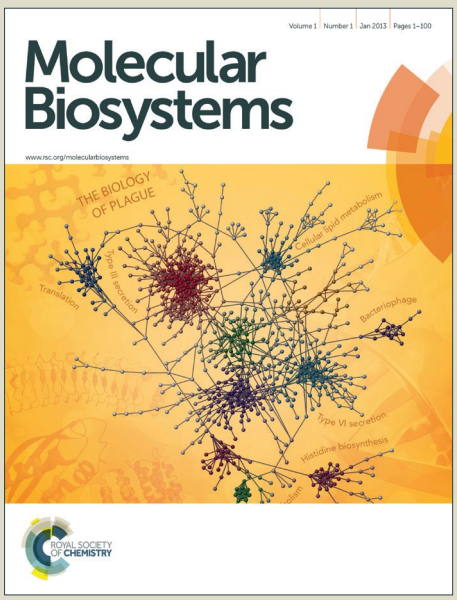
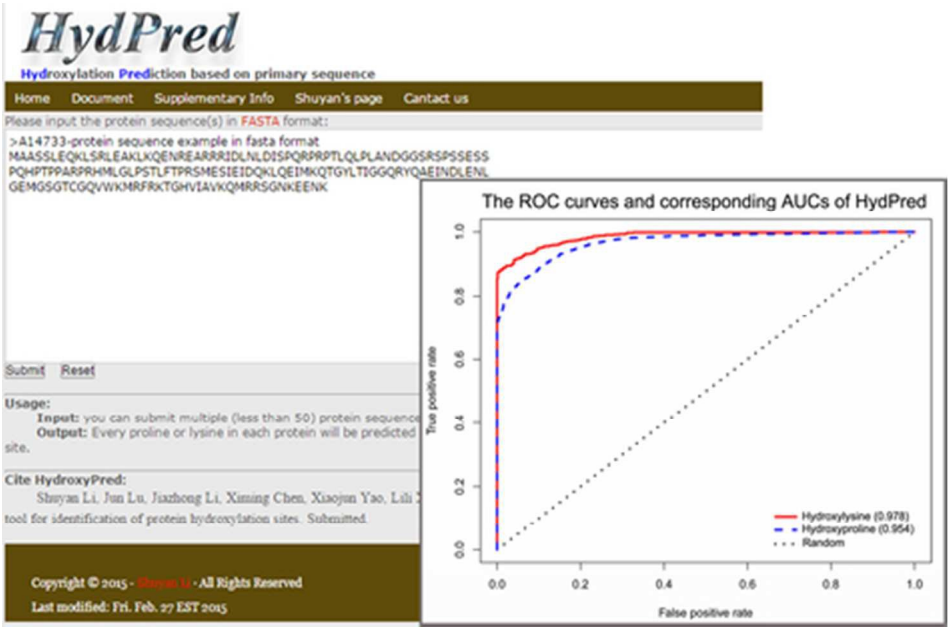
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



39x26mm (300 x 300 DPI)

Molecular BioSystems

ARTICLE

HydPred: A novel method for identification of protein hydroxylation sites that reveals new insights into human inherited disease

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Shuyan Li,[a,#] Jun Lu,[b,#] Jiazhong Li,[c] Ximing Chen,[d] Xiaojun Yao[a] and Lili Xi[e,*]

Disruption of protein hydroxylation is highly associated with several serious diseases and consequently the identification of protein hydroxylation sites has attracted significant attention recently. Here, we report development of an improved method, called HydPred, to identify protein hydroxylation sites (hydroxyproline and hydroxylysine) based on the synthetic minority over-sampling technique (SMOTE), the random forest (RF) algorithm and four blocks of newly composed features that are derived from the protein primary sequence. The HydPred method achieved the best prediction performance reported until now with Matthew's correlation coefficient values of 0.770 and 0.857 for hydroxyproline and hydroxylysine, respectively, according to jack-knife cross-validation. This represents an improvement of 8% for hydroxyproline and 19% for hydroxylysine compared to the best results of available predictors. The prediction performance of HydPred for external validation of hydroxyproline and hydroxylysine was also improved compared with other published methods. We subsequently applied HydPred to study the association of disruption of hydroxylation sites with human inherited disease. The analyses suggested that loss of hydroxylation sites is more likely to cause disease instead of gain of hydroxylation sites and 52 different human inherited diseases were found highly associated with loss of hydroxylation sites. Therefore, HydPred represents a new strategy to discover the molecular basis of pathogenesis associated with abnormal hydroxylation. HydPred is now available online as a user-friendly web server at http://lishuyan.lzu.edu.cn/hydpred/.

Introduction

Protein hydroxylation is a chemical process that introduces a hydroxyl group (-OH) into residues. The principal hydroxylated residue is proline, forming hydroxyproline and the minor hydroxylated residue is lysine, forming hydroxylysine. Hydroxylation plays several critical roles in biological systems, e.g. it is crucial for the assembly of collagen and other extracellular matrix components[1]. It modulates the function of estrogen[2], regulates cellular oxygen sensing through hypoxia-inducible pathways[3], and hence, creates neuroprotection[4]. Correspondingly, the disruption of protein hydroxylation was discovered to be associated with a number of serious human diseases, such as osteogenesis imperfecta (brittle bone disease)[5], osteoporosis[6], nervous system tumours[7], kidney disease[3], breast cancer[2] and other hormone-related or hypoxia-related cancers[8-10]. Therefore, the identification of protein hydroxylation sites and hence annotation of hydroxylation in proteomes has attracted recent attention. For this goal, mass spectrometry (MS) related methods achieved some success[11, 12], but this approach is always time-consuming and expensive. Moreover, the data produced by MS are also biased toward abundant proteins and prototypic peptides since they were frequently studied. Considering the explosive increase in protein data, computational methods emerged as a promising tool for the first round annotation of hydroxylation sites.

Several predictors were already created for identification of hydroxylation sites. The first predictor was proposed by Yang's group, which directly targeted collagen hydroxyproline sites by support vector machines[13]. Later, this work was expanded on both hydroxyproline and hydroxylysine by Hu et.al. using a more comprehensive dataset that was extracted from UniProt/Swiss-Prot, which is cross-species and contained different protein types[14]. They utilized the nearest neighbor algorithm and 6345 features that were derived from the protein primary sequence. However, the final performance was merely adequate, with Matthew's correlation coefficient (MCC) values of 0.461 and 0.592 for hydroxyproline and hydroxylysine sites, respectively. Then, a predictor called iHyd-PseAAC was presented by Xu et al. based on the dipeptide position-specific propensity of pseudo amino acid composition[15]. Although this work led to the first

[a] College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, China 730000
[b] School of Basic Medical Sciences, Lanzhou University, China 730000
[c] School of Pharmacy, Lanzhou University, Lanzhou, China 730000
[d] Key Laboratory of Desert and Desertification, Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, China 730000
[e] Department of Pharmacy, First Hospital of Lanzhou University, Lanzhou, China 730000
# These two authors contributed equally to this work.
*Corresponding author: L. X.(xill@lzu.edu.cn)

online web server for the identification of hydroxylation sites, the MCC values achieved for hydroxyproline and hydroxylysine of this predictor were both only approximately 0.50. Soon afterwards, Shi et al.[15] described a new hydroxylation prediction method called PredHydroxy, based on the position weight amino acids composition and 8 high-quality amino acid indices. In this method, the MCC values for the first time showed a remarkable increase, up to 0.667 and 0.690 for hydroxyproline and hydroxylysine, respectively. However, to create a reliable hydroxylation sites annotation tool for newly discovered protein data, this prediction performance still needs more improvement.

To achieve this goal, we are proposing a new *in silico* method called HydPred, based on the RF algorithm[16] and the four categories of newly composed features that are derived from protein sequence information. These features were calculated from the amino acid composition, the autocorrelation of amino acid physicochemical properties, the amino acid position weighted matrices and the amino acid binary localization encoding. The SMOTE method[17] was utilized here to fix the data unbalancing problem. Based on the extracted features, the final identification performance of HydPred was greatly improved compared with the best results of previously published predictors. How this method was developed is illustrated in Fig1. Our validated HydPred method was then applied to analyze differences in hydroxylation of protein variants associated with human inherited diseases that could underlie disease pathologies.

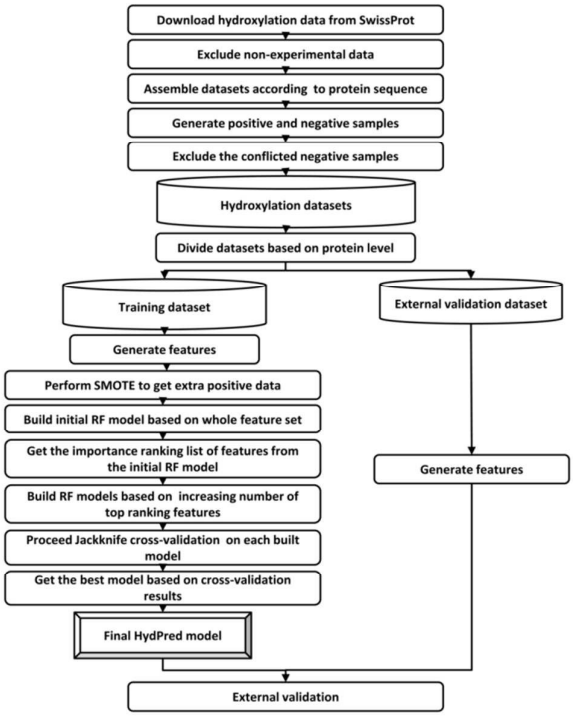


Fig1. Flowchart of the proposed HydPred method for the identification of hydroxylation sites.

Materials and methods

Hydroxylation Dataset

The hydroxylation data were assembled from Swiss-Prot(Release:2014-09), which is currently the most comprehensive dataset of hydroxylation data. Here, only the experimentally verified information was extracted by searching 'hydroxyproline' or 'hydroxylysine' in the field 'modified residue' in the text format of the SwissProt dataset. The data with an annotation confidence of 'probable', 'potential', or 'by similarity' were excluded. If the proteins from different species have the same sequence and different hydroxylation entries, these entries were assembled to a unique protein sequence as non-redundant entries. Therefore, following these steps, 1031 hydroxyproline sites from 206 unique proteins and 127 hydroxylysine sites from 33 unique proteins were extracted.

A peptide fragment of 13 amino acids in length, which was centered on proline or lysine with 6 upstream and 6 downstream amino acids, was represented as a study sample. Because there is physical evidence of kinase-substrate binding within neighborhood residues of the modification site [16], we believe that the prolyl-hydoxylases/lysyl-hydorxoylases will likely recognize residues surrounding the modification site as well. Taking hydroxyproline as an example, the fragment that was centered by a proline that had a hydroxylation entry was treated as positive sample, and the fragments that were centered by non-hydroxylated prolines were treated as negative samples. The length of 13 is commonly used in previous predictors for hydroxylation sites identification[15, 18], therefore, we also used this length. All of the redundancy entries within the same sequence were integrated into a unique sample. Since most of the positive samples of hydroxylation shared high sequence identities, we believed that there should be several conserved sites exist inside the sample sequence which are very important for the binding of prolyl-hydoxylases/lysyl-hydorxoylases. Therefore, instead of roughly removing the redundant samples according to a certain percent of sequence identity, only the identical hydroxylation samples were removed from the database to avoid the pre-introduced bias.

Meanwhile, the negative samples that either had the same sequence with positive samples or had a non-experimental annotation confidence of 'probable', 'potential', or 'by similarity' for hydroxylation sites were also eliminated from the negative sets in order to obtain a rigorous dataset. Then, 825 positive and 2600 negative samples were obtained for hydroxyproline, while 121 positive and 937 negative samples remained for hydroxylysine.

Lastly, to rigorously validate the prediction performance, the protein set was randomly divided into a cross-validation training set and an external validation set by a ratio of approximately 9 to 1 on the protein level to avoid intra-protein bias. It was rounded towards plus infinity for the protein number in the external validation set for both hydroxyproline and hydroxylysine. Thus, 21 proteins with 55 positive samples and 165 negative samples were split into the

external validation set of hydroxyproline, and 4 proteins with 12 positive samples and 169 negative samples were split into the external validation set for hydroxylysine. The detailed information including protein ID, site location, fragment sequence, etc. are listed in the supplementary Table S1.

Feature generation

Each site was represented by a 13-residue long peptide centered at the proline or lysine of interest as described above. If the upstream or downstream residue number was less than 6, symbol 'X' was assigned to represent the missing amino acid in the sequence fragment. Afterwards, each fragment was transformed into four categories of features.

I. Amino acid composition. Both the single amino acid composition and the K-spaced amino acid pair composition was calculated here. Single amino acid composition is the fraction of each amino acid type in a sequence fragment. The K-spaced amino acid pair composition was first proposed by Chen et al.[15, 18] for the prediction of protein flexible/rigid regions, and proved to be useful for the prediction of O-glycosylation sites19 and palmitoylation sites20. The K-spaced amino acid pair compositions were calculated by considering the fraction of amino acid pairs that are separated by k amino acids within a protein sequence fragment (there are 441 possible pairs, e.g., AA, AC, AD, ..., XX). We refer to such a feature vector as $(C_{AkA}C_{AkC}C_{AkD}...C_{XkX})_{441}$. For instance, $C_{A3C} = N_{A3C}/(N-1)$, where N_{A3C} is the number of occurrences of the AC amino acid pair separated by 3 amino acids and N is the residue length of the peptide. In this work, N set to 13 and $k = 0,1,...,9,10$ were jointly considered. In total, 4872 (21+441*11) compositional features were generated in this category.

II. Autocorrelation of amino acid physicochemical properties. Thirteen amino acid properties were used for this feature set, including (1) the hydrophobicity scale[21], (2) the average flexibility index[22], (3) the polarizability parameter[23], (4)the free energy of solution in water, (5) the residue accessible surface area for a tripeptide[24], (6) the average volumes of residues[25], (7) the steric parameter[26], (8) the relative mutability[27], (9) the polarity factor, (10) the secondary structure factor, (11) the molecular volume factor, (12) the codon diversity factor and (13) the electrostatic charge factor. The latter five properties were derived from the literature[28-30] and assembled by Hu et al.[31] as transformed attributes. All of the properties were centralized and standardized before the calculation, and the properties for residues represented by X were set to zero.

Two autocorrelation calculating algorithms, the Geary autocorrelation and the normalized Moreau-Broto autocorrelation, were adopted here. In this way, the Geary autocorrelation features[32] are defined as:

$$G(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \overline{P})^2} \quad (1)$$

where $d = 1,2,3,...,30$ is the lag of the autocorrelation. P_i and P_{i+d} are the particular property values of the amino acids at position *i* and *i*+*d*, respectively. $\overline{P}$ is the average value of P_i.

The normalized Moreau-Broto autocorrelation features are defined as:

$$NMB(d) = \frac{\sum_{i=1}^{N-d} P_i P_{i+d}}{N-d} \quad (2)$$

where d, P_i and P_{i+d} are previously defined and N is the length of peptide.

In summary, this block of features was inspired by the PROFEAT method[33]. There are 780 (13*30*2) features in this block.

III. Amino acid position weighted matrices (PWMs). For a given dataset of fixed-length sequence fragments, each amino acid at each position is associated with its frequency of occurrence. Using this frequency in both hydroxylation and non-hydroxylation sets of fragments in the training set, two position-weighted matrices (PWMs)[34, 35] are calculated for both hydroxyproline and hydroxylysine data. Each row of the matrix corresponds to one type of amino acid, and every column corresponds to a position in the peptide. The element in the i-th row and j-th column of the matrix is defined as:

$$f_{ij} = N_{ij} / N_{pep} \quad (3)$$

where N_{ij} is the number of occurrences of i-th amino acid in the j-th position of the peptide and N_{pep} is the number of peptides in the whole dataset, $i = 1,2,3,...,21$, $j = 1,2,3,...,13$. Because two PWMs were built, 26 (13*2) features were extracted in this group.

IV. Amino acid binary localization encoding. Each amino acid of a peptide is encoded into a 21-bit vector by a one and twenty zeros. This vector represents the localization of 20 normal amino acids and a non-existent residue (represented by X) as "ACDEFGHIKLMNPQRSTVWXY". A (Alanine) is encoded as "10000000000000000000", the one in the vector alphabetically shifts to the right localization based on the letter abbreviation of the amino acids. Therefore, a 13-length peptide is encoded as 273 (21*13) features.

In total, 5951 features were generated. To reduce redundant information, constant features and highly correlated features were excluded in the feature space. In addition, if any two features shared a correlation coefficient greater than 0.85, one of them was removed randomly. Finally, 4959 features remained for the next procedure.

Balanced positive sample generation

Because, until now, the unbalanced problem of the positive and negative sample numbers hindered the MCC value for hydroxylation site prediction performance, the SMOTE[17] method was utilized to fix this problem.

With the generated features of positive samples in the training set, the SMOTE method used a *k*-nearest neighbors (KNN)[35] based algorithm to create extra positive samples from the current positive dataset to fix the unbalanced problem.

The SMOTE method was performed by the DMwR package in the R program[36].

Feature ranking and modeling

Aiming to further reduce the dimension of feature space and to find a restricted number of features yielding good classification performance, feature ranking and modeling by the RF method was performed here. The RF method was first introduced by Leo Breiman[16] and then proved to be a very powerful classification method in the fields of chemometrics and bioinformatics[37-40]. RF is a classifier consisting of collection of tree-structured classifiers with two major advantages being (1) using an out-of-bag method[41] to monitor error, strength and correlation, and (2) measuring variable importance through permutation.

With a whole set of features to build an initial RF classification model, the results of RF model are able to show the importance for each feature by its association with the prediction target. In this process, all features will first be trained to fit in a random forest. During the fitting process, the out-of-bag error for each data point is recorded and averaged over the forest. To measure the importance of the i-th feature after training, the values of thei-th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the i-th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences. Features which produce large values for this score are ranked as more important than features which produce small values. Then, the importance ranking list of features was generated in this work. After that, with an increased number of top ranking features, different RF models will be built to select the best model that had the least feature space and had an equivalent prediction performance that compared with using the whole feature space.

There are two important parameters in RF method that need to be tuned to get a better performance. The first one is 'ntree', which is the number of trees to grow. It should not be set to too small a number, to ensure that every input row gets predicted at least a few times. The second one is 'mtry0', which is the number of variables to split on at each node. The other parameters of RF model were all set to default value.

The feature ranking and modeling processes were solely based on the training dataset without any involvement of the external validation set. This performed a reliable supervised feature selection and modeling to avoid the over-fitting problem as Smialowski et al. emphasized[42]. The RF method was implemented using the R package randomForest v 4.6-7.

Model evaluation

Model evaluation was performed by not only jack-knife cross-validation on the training dataset, but also on an external validation set in order to prove the generalization ability of HydPred. For jack-knife cross-validation, each sample of the internal cross-validation set was treated as a test set once, while the remaining data, including N-1 samples, were gathered as training data, where N is the total sample number of the internal cross-validation set. After N turns of calculation, the averaged prediction result of test sets was taken as the final jack-knife cross-validation result. For the external validations set, it was only used to test the generalization ability of the built model and not involved in any procedure of modeling building process.

Five frequently used indicators were utilized here, sensitivity (Sens), specificity (Spec), accuracy (ACC), MCC and the area under receiver operating characteristic curve (AUC). A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier system while its decision threshold is varied. Each point in the ROC curve is created by plotting the true positive rate vs. the false positive rate at a particular decision threshold. The AUC, the area under the ROC curve, can give a complete evaluation of a prediction method. Some other parameters of the RF model, such as the number of trees and number of leaves on each tree, were also optimized to maximize the AUC.

Identification of loss or gain of hydroxylation sites in human inherited disease

Because protein hydroxylation is associated with several serious diseases[2, 3, 5, 6, 8-10, 43, 44], HydPred was then applied to analyze human inherited disease associated proteins to assess the loss and gain of hydroxylation sites that could underlie disease pathologies.

Firstly, HydPred estimates the probability that a residue s_i can be hydroxylated, as $P(s_i = s_i^p \mid S)$, given the protein sequence S.

Then, we can express the probability of loss of hydroxylation at residue s_i as below:

$$P_l(s_i) = P(s_i = s_i^p \mid S) \times (1 - P(s_i = s_i^p \mid S_{xjy})) \qquad (5)$$

where S is the wild type protein sequence and S_{xjy} is the same sequence with a mutation from residue x to residue y at position j, where x, y $\in$ (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,U,W,Y). Note that only the variation that exists in the neighbourhood (±6 residues) of prolines or lysines is considered to have influence on gain or loss of hydroxylation sites. With the same principle, the probability of a gain of hydroxylation site at residue s_i is defined as:

$$P_g(s_i) = (1 - P(s_i = s_i^p \mid S)) \times P(s_i = s_i^p \mid S_{xjy}) \qquad (6)$$

The threshold that distinguishes a gain or loss of hydroxylation sites was set to the same cut-off value of HydPred.

Results and discussion

Feature selection and prediction performance

At first, the whole feature space was used to build the initial RF models for hydroxyproline and hydroxylysine, respectively. The value of ntree was tuned from 200-1000 with a step of 100 while the value of mtry0 was tuned from 2 to 50 with a step of 1 in each tuning step of ntree. Then, ntree=500 and mtry0=20 was selected as the optimal parameters with the best AUC performance as 0.952 for hydroxyproline and provided the ranking of feature importance accordingly. Likewise, for hydroxylysine, the initial RF model was built with the best AUC value as 0.977 (ntree=500, mtry0=10). Then, different models were built on the increased number of top ranked features for both hydroxyproline and hydroxyproline by RF algorithm.

After then, approximate highest peak MCC, ACC and AUC scores were obtained when the top ranked 49 features were selected for hydroxyproline and the top 44 features were selected for hydroxylysine. The AUC for hydroxyproline reached 0.954, and for hydroxylysine, it reached 0.978 based on jack-knife cross-validation (Fig2). Therefore, the 49 top ranked features were selected for hydroxyproline and the 44 top features were selected for hydroxylysine to build the HydPred predictor. Information regarding these selected key features are listed in supplementary Table S2.

The detailed prediction performance of HydPred for jack-knife cross-validation and external validation is listed in Table 1 and Table 2, where the results of the external validation set are fairly stable compared with cross validation results on the training set, which is a sign of favorable generalization ability for a classifier[45].

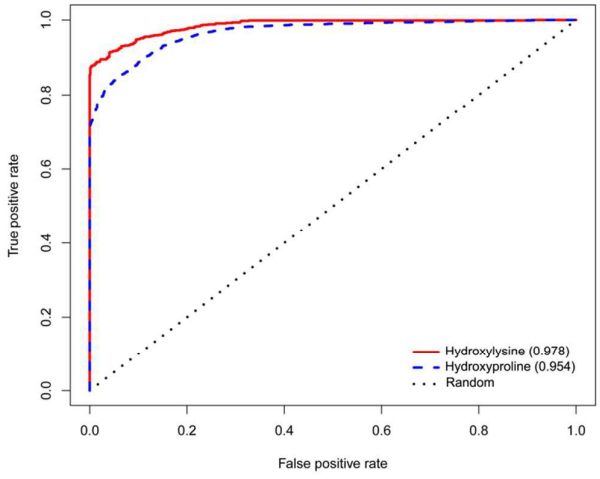


Fig2. ROC curves and the corresponding AUCs of HydPred for hydroxyproline and hydroxylysine.

Performance comparison

Because prior methods to predict protein hydroxylation sites used several different datasets, a more rigorous performance comparison between HydPred and previous predictors was performed. First, jack-knife cross-validation results reported for established methods were compared with those for HydPred (Table 1). Then, the well-established methods for which the software was either available online or ready to be downloaded were tested on the same external validation set.

By this criterion, the PredHydroxy [18] and iHyd-PseAAC [15] methods were utilized to perform the comparative assessment on the same external validation set (Table 2).

Compared with the best score of prior published predictors, our HydPred method showed a remarkable improvement on the MCC score in both jack-knife cross-validation (Table 1) and external validation (Table 2). For the first time, the MCC score achieved up to 0.770 for hydroxyproline and 0.857 for hydroxylysine based on jack-knife evaluation. However, a prediction method with only cross-validation results is still far from being a reliable predictor for the over-fitting problem that is occasionally found in such a situation[46]. Therefore, an external validation on a dataset that is not involved in the model building process is crucial for estimating the real performance and generalization ability for a predictor. In other words, only

Table 1. Prediction performance of hydroxyproline and hydroxylysine by different predictors based on jack-knife cross-validation.

Target	Predictor	Sens	Spec	ACC	MCC
Hydroxyproline	**HydPred(Cut=0.408)**	**0.796**	**0.961**	**0.880**	**0.770**
	PredHydroxy	0.838	0.852	0.845	0.690
	Hu's method	0.648	0.816	0.760	0.461
	iHyd-PseAAC	0.807	0.805	0.806	0.510
Hydroxylysine	**HydPred(Cut=0.303)**	**0.859**	**0.991**	**0.925**	**0.857**
	PredHydroxy	0.842	0.825	0.833	0.667
	Hu's method	0.704	0.880	0.821	0.592
	iHyd-PseAAC	0.879	0.830	0.836	0.500

Table 2. Prediction performance of hydroxyproline and hydroxylysine by different predictors based on the same external validation set.

Target	Predictor	Sens	Spec	ACC	MCC
Hydroxyproline	**HydPred**	**0.709**	**0.891**	**0.846**	**0.593**
	PredHydroxy(*60%)	0.618	0.885	0.818	0.509
	PredHydroxy(*70%)	0.491	0.976	0.873	0.581
	PredHydroxy(*80%)	0.364	0.976	0.823	0.471
	iHyd-PseAAC	0.545	0.909	0.818	0.488
Hydroxylysine	**HydPred**	**0.917**	**0.970**	**0.967**	**0.778**
	PredHydroxy(*60%)	0.833	0.947	0.939	0.633
	PredHydroxy(*70%)	0.750	0.964	0.950	0.645
	PredHydroxy(*80%)	0.667	0.982	0.961	0.676
	iHyd-PseAAC	0.833	0.947	0.939	0.633

* The threshold value of specificity of the PredHydroxy method.

predictors with outstanding prediction performance on both cross-validation and external validation can be proved to be reliable predictors.

In this way, HydPred is proved to be a reliable tool, as it showed a good prediction performance both on cross-validation and external validation.

The SMOTE method also has favorable performance by solving the unbalanced problem of positive and negative samples. As observed from Table 1, the unbalanced problem did not usually affect the cross-validation results, as nearly all of the methods had a comparative sensitivity and specificity whether they accounted for the unbalanced problem. However, once they were tested by an external validation set, the unbalanced problem occurred (Table 2). This was particularly evident for prediction of hydroxyproline. Although the PredHydroxy (70%) method showed comparative MCC results with our HydPred method, it only showed a sensitivity of 0.491, which means that more than half of the positive samples were mis-predicted to be negative samples. Although it also showed good ACC results, this lack of sensitivity of the PredHyrdroxy method is not acceptable. Consequently, HydPred is currently the most promising predictor of hydroxylation sites due to its ability to solve the problem of unbalanced data.

Analyses of key features

There are 49 and 44 key features were selected in the HydPred predictor for hydroxyproline and hydroxylysine, respectively. There are lots of underlying information along with these features that may provide interesting insights into the understanding of hydroxylation mechanisms.

Among these features, the majority of them were classified into category I as component features (30/49 for hydroxyproline and 24/44 for hydroxylysine) which suggests that the residue component is one of the keys that helps prolyl-hydoxylases/lysyl-hydorxoylases to recognize their substrate. Here, Glysine (G) plays a rather important role since it is involved in more than half of these component features (16/30 for hydroxyproline and 17/24 in hydroxylysine).

The Binary localization encoding strategy as category IV implies the same deduction with more details. It suggests that the binary localization encoding of G at position 2, 3, 5, 6, 8, 9 and 11 are all important for hydroxyproline. It is slightly different for hydroxylysine since the important binary localization of G are only at position 2, 5, 8 and 11.

The above analyses was further complemented by the key features from PWMs as category III. For hydroxyproline, it emphasizes that positive PWMs at position 2, 5, 8 and 11 and the negative PWMs at position 3, 6, 8, 9, 11 and 12. For hydroxylysine, it selected the key features of positive PWMs at position 2, 3, 6, 8, 9 and 11 and negative PWMs at position 2, 3, 5, 8 and 11.

In order to illustrate these results intuitively, the Two Sample Logo (TSL)[47] tool was utilized here (Fig3). TSL method can calculate and visualize the differences between the

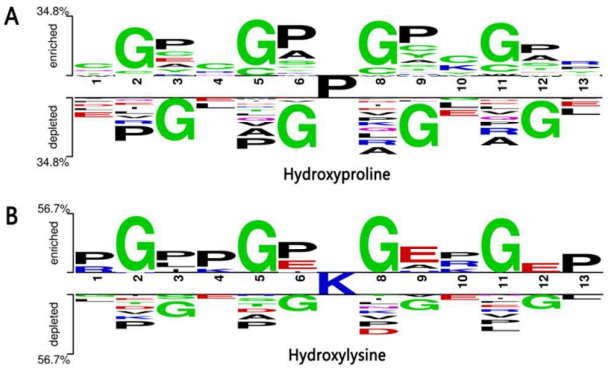


Fig3. Enrichment and depletion of residues at different positions of hydroxylation samples.

positive and negative sets of aligned peptides by T-test and then present the enriched and depleted residue types in each position. The enrichment of residues can partially but not entirely reflect the positive PWMs features and the depleted residues can partially reflect the negative ones. In Fig3, it is obvious that G is enriched at position 2, 5, 8 and 11 in both hydroxyproline and hydroxylysine samples, meanwhile it is also depleted at position 3, 6, 8 and 12 for both of them as well. Besides, P is also enriched in multiple positions of hydroxylation samples. Here, it is also slightly different for hydroyproline from hydroxylysine. The enriched positions for P in hydroxyproline are 3, 6, 9 and 12, but in hydroxylysine, the positions are 1, 3, 4, 6, 10 and 13. From Table 3, residue P is also involved in many key component features in HydPred, as 6 for hydroxyproline and 7 for hydroxylysine. Residue P has a distinctive five-member cyclic group compared with other residues which can only be a H-bond acceptor but not a donor. It is interesting that residues G and P are the only two residues that do not follow along with the typical Ramachandran plot[48] and they are all enriched in hydroxylation samples. The combination of G and P in a peptide would deduce two completely different consequence in their 2D structures, which is either rather ordered because of the distinct structure of residue P, such as in collagen (basically as a repeated GPP format) or tend to be a disordered coil since residue P is rarely found in α-helix or β-sheet structures expect in fibre proteins like collagen. It implies that the binding site of prolyl-hydoxylases/lysyl-hydorxoylases is either in a rather ordered structure like in the helix in collagen or tend to be in a disordered coil. Residue C is also enriched in most positions of hydroxyproline samples, such as position 1, 3, 4, 5, 8, 9, 10 and 11. Coordinate with this phenomenon, the single amino acid component of C is also selected as key feature in HydPred for hydroxyproline.

Besides, residue H, as the only residue with both a five-member cyclic group and positive charge, is involved in 8 key component features for Hydroxyproline, but it does not show up for hydroxylysine at all. Interestingly, as same as a residue with positive charge, K shows up 7 times in key features for hydroxylysine but not once in hydroxyproline. It may implies

that the side chain with positive charge group is important for hydroxylation, where H contributes especially to the binding of prolyl-hydoxylases while K contributes especially to that of lysyl-hydorxoylases. Since H does not show up as the enrichment residues in hydroxyproline samples (Fig3A), it suggests H does not tent to locate at specialized positions of a hydrxoyproline sample. Meanwhile, K tends to locate at 4 and 10 in a hydroxylysine sample (Fig3B).

For category II, the volume involved features were selected as key features for both hydroxyproline (as GearynAuto_11_2) and hydroxylysine (as GearyAuto_6_2). Referring back to the component features, most the residues that are involved in the key component features are residues with small volumes, such as G, A, S, P, T, D, E. The residue with big volume, such as M, W and Y, does not appear in any of these key features. This phenomenon suggests that the hydroxylysine sites are likely to exist in the neighbourhood of residues with small volumes.

Loss or gain of hydroxylation sites in human inherited disease

Human disease associated variation data was gathered from the SwissVar[49] database (release 201504). There are 26087 disease related variations and 38058 neutral polymorphisms in 12569 proteins in all. From these, fragments of 13 residues centered on prolines and lysines where the variations are inside of these fragments were extracted as the disease-related or polymorphisms related samples. HydPred was then employed to predict the loss or gain of hydroxylation sites for both polymorphisms and disease related samples that were caused by the variations. One thing should be noted that the distribution of training data for HydPred is unlikely identical to the human variation data. Therefore, we addressed a 'high confidence' loss or gain of hydroxylation sites, inspired by Radivojac's work[50] to overcome the possibility of biased inference on false positive rate. Therefore, only the site with value of $P_l(s_i)$ or $P_g(s_i)$ that is larger than 0.500 for hydroxyproline and 0.400 for hydroxylysine (both higher than the cutoff value of HydPred) was assigned as a high confidence loss or gain of hydroxylation site. Correspondingly, only these sites were then analyzed in the further procedure.

The statistic results in Table 3 indicate that the loss of both hydroxyproline sites and hydroxylysine sites is significantly higher in disease associated variations than in polymorphisms according to the small p-values (as 1.29e-8 and 9.67e-4) that was calculated by chi-square test (L.p-value in Table 3), which suggests that loss of hydroxylation sites is likely to be involved with the pathogenesis of human inherited diseases. Meanwhile, since the p-values for gain of hydroxyproline and hydroxylysine sites are both higher than 0.05, which suggests that the gain of hydroxylation sites is not significant for disease-associated variations against polymorphisms and consequently implies that the gain of hydroxylation sites is not likely to be strongly associated with disease.

The predicted high confidence loss of hydroxylation sites in disease associated variations are listed in Table S3 for the further analyses. Among these disease-associated variations, 111 of them are predicted to cause 117 different high-confidence loss of hydroxylation sites by the HydPred method (Table S3). It indicated that about 0.43% (111/26087) disease-associated variations will cause high-confidence loss of hydroxylation sites. In other words, it is also implies that among these disease-associated variations, about 0.43% of the disease-causing mechanisms may be related with loss of hydroxylation sites.

Table 3. High confidence prediction results of loss or gain of hydroxylation sites.

	Dis.	Poly.	Loss			Gain		
			Dis.	Poly.	L.p-value	Dis.	Poly.	G.p-value
Hydroxyproline	19553	34063	81	53	1.29e-8	60	133	0.12
Hydroxylysine	16600	27361	36	26	9.67e-4	13	21	0.95

Note: Dis. indicated the number of disease-associated hydroxylation samples. Poly. indicated the number of polymorphism-related hydroxylation samples. L.p-value indicated the significance of loss of Dis. compared with loss of Poly. by chi-square test; G.p-value indicated the significance of gain of Dis. compared with gain of Poly. by chi-square test.

As indicated in Table S3, there are 44 proteins and 52 different types of human disease associated with loss of hydroxylation sites. Among these, 68 variations within 11 proteins are involved in 18 types of disease that are implicated with dysplasia of the skeleton, joint, muscle or skin, which are mostly due to variations in hydroxylation of collagen (noted by 'a' in Table S3). Indeed, most of the loss of hydroxylysine sites are involved in this category of diseases.

Several other types of serious diseases are also indicated in Table S3, such as malfunction of the neurological system (noted by 'b' in Table S3), abnormalities affecting cardiovascular and cerebrovascular function, arteries or blood (noted by 'c'), severe immunodeficiency (noted by 'd') and kidney disease (noted by 'e').

Previous studies have reported the association of disruption of hydroxylation sites with diseases caused by defect of collagen[5, 43], malfunction of the neurological system[7] and kidney disease[3]. All these diseases are included in the above analyses results. However, the other associations with abnormalities of cardiovascular and cerebrovascular function or blood and severe immunodeficiency are rarely reported. Although variations in primary protein sequence and function may underlie the pathologies of these diseases, in some examples, such as Brugada syndrome-1[51], this is unclear. Our analyses indicate that there is also a strong likelihood that loss of hydroxylation sites may play a key role in the pathogenesis of these diseases. Indeed, our predicted hypo-hydroxylation of protein variants associated with these diseases may well be provided as a basis for the development of novel therapies in the future.

Web Server of HydPred

The web server for HydPred can be found online at Http://lishuyan.lzu.edu.cn/hydpred. It is a user-friendly web server for researchers which can be easily utilized without any mathematical or computational background.

The interface of HydPred is shown as Fig4A. Users can input single or multiple protein sequences in a FASTA format into the textbox and press the 'submit' button. Then, all of the prolines and lysines in a protein will be predicted to be hydroxylation sites or not. After the calculation, the outputs will be listed in the results page, as shown in Fig4B. It takes approximately 2 seconds to calculate and output the prediction results for both hydroxyproline and hydroxylysine for a protein of approximately 200 residues. Correspondingly, the time cost is multiplied when more protein sequences are inputted.

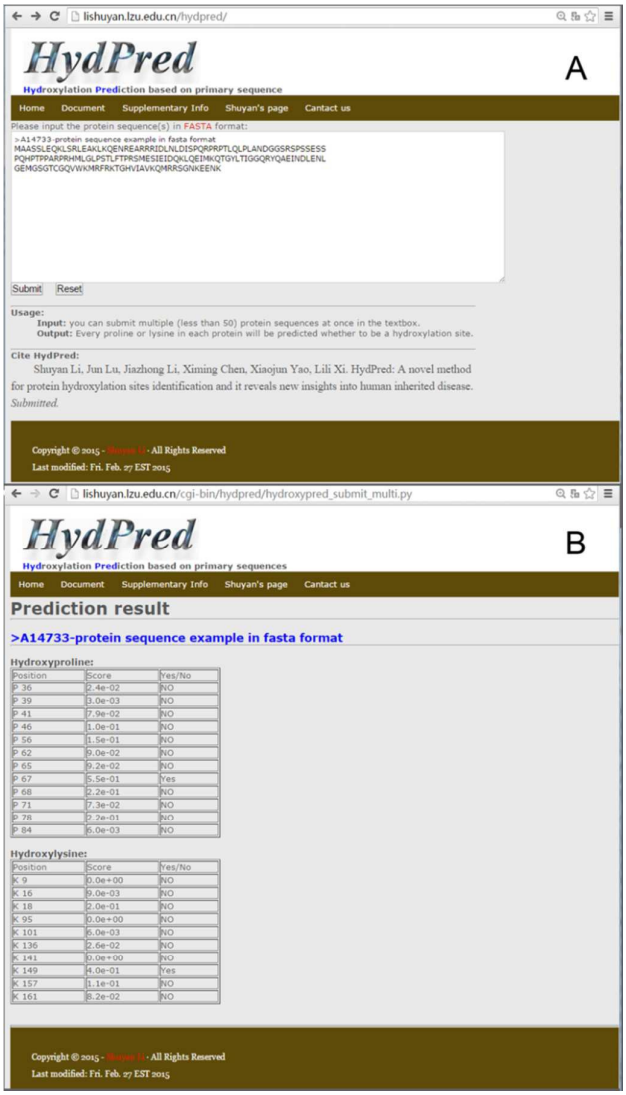


Fig4. HydPred web server. A. Interface of HydPred with input; B. Results page of HydPred.

Conclusions

In this work, we describe a reliable identification method for protein hydroxylation sites called HydPred and comprehensively studied the potential loss or gain of hydroxylation sites in human inherited disease for the first time. The introduction of the SMOTE algorithm in HydPred fixed the unbalanced problem of positive and negative samples; hence, it showed a more favorable performance, especially on the external validation set compared to other published predictors in the same field. The web server of HydPred is a fast and reliable annotation tool for hydroxylation sites in newly discovered protein sequences. Moreover, the application of HydPred suggests that the loss of hydroxylation sites is more likely to be involved in human inherited disease than gain of hydroxylation sites. 52 different types of disease were found to be associated with potential loss of hydroxylation sites in the corresponding protein variants, which presents a new angle for understanding the basis of disease pathogenesis and developing targeted therapies at the molecular level.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 21405068 to S. L., No. 21205055 to J. L. and No. 31400437 to X. C.) and the Fundamental Research Funds for the Central Universities of China (No. lzujbky-2015-31 to S. L. and No. lzujbky-2013-153 to L. X.). We would like to thank Paul Dyson in Swansea University for his great help on the language improvement of this manuscript. We also like to thank Gansu Computing Center for providing the computing resources.

Conflict of interest

None.

References

1. D. M. Hudson and D. R. Eyre, *Connect Tissue Res*, 2013, **54**, 245-251.
2. D. W. Sepkovic and H. L. Bradlow, *Ann NY Acad Sci*, 2009, **1155**, 57-67.
3. R. H. Wenger and D. Hoogewijs, *Am J Physiol Renal Physiol*, 2010, **298**, F1287-F1296.
4. A. Siddiq, L. R. Aminova and R. R. Ratan, *Front Biosci*, 2008, **13**, 2875-2887.
5. J. C. Marini, W. A. Cabral, A. M. Barnes and W. Chang, *Cell Cycle*, 2007, **6**, 1675-1681.
6. N. Napoli and R. Armamento-Villareal, *Adv Clin Chem*, 2007, **43**, 211-227.
7. S. Schlisio, *J Cell Mol Med*, 2009, **13**, 4104-4112.
8. K. Salnikow and K. S. Kasprzak, *Environ Health Perspect*, 2005, **113**, 577-584.
9. J. Palka, A. Surazynski, E. Karna, K. Orlowski, Z. Puchalski, K. Pruszynski, J. Laszkiewicz and H. Dzienis, *Hepato-Gastroenterol*, 2002, **49**, 1699-1703.
10. I. Gecit, M. Aslan, M. Gunes, N. Pirincci, R. Esen, H. Demir and K. Ceylan, *J Cancer Res Clin*, 2012, **138**, 739-743.
11. M. E. Cockman, J. D. Webb, H. B. Kramer, B. M. Kessler and P. J. Ratcliffe, *Mol Cell Proteomics*, 2009, **8**, 535-546.
12. M. Getie, C. E. H. Schmelzer and R. H. H. Neubert, *Proteins*, 2005, **61**, 649-657.
13. Z. R. Yang, *J Comput Biol*, 2009, **16**, 691-702.
14. L.-L. Hu, S. Niu, T. Huang, K. Wang, X.-H. Shi and Y.-D. Cai, *PLoS ONE*, 2010, **5**, e15917.

15. Y. Xu, X. Wen, X.-J. Shao, N.-Y. Deng and K.-C. Chou, *Int J Mol Sci*, 2014, **15**, 7594-7610.
16. L. Breiman, *Machine Learning*, 2001, **45**, 5-32.
17. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, *J. Artif. Intell. Res.*, 2002, **16**, 321-357.
18. S. Shi, X. Chen, H. Xu and J. Qiu, *Mol Biosyst*, 2014, DOI: 10.1039/c4mb00646a.
19. Y.-Z. Chen, Y.-R. Tang, Z.-Y. Sheng and Z. Zhang, *BMC Bioinformatics*, 2008, **9**, 1-12.
20. X.-B. Wang, L.-Y. Wu, Y.-C. Wang and N.-Y. Deng, *Protein Eng. Des. Sel.*, 2009, **22**, 707-712.
21. H. Cid, M. Bunster, M. Canales and F. Gazitua, *Protein Eng.*, 1992, **5**, 373-375.
22. R. Bhaskaran and P. K. Ponnuswamy, *J Pept Protein Res*, 1988, **32**, 241-255.
23. M. Charton and B. I. Charton, *J Theor Biol*, 1982, **99**, 629-644.
24. C. Chothia, *J Mol. Biol.*, 1976, **105**, 1-12.
25. J. Pontius, J. Richelle and S. J. Wodak, *J Mol. Biol.*, 1996, **264**, 121-136.
26. J. L. Fauchère, M. Charton, L. B. Kier, A. Verloop and V. Pliska, *Int. J. Pept. Protein Res.*, 1988, **32**, 269-278.
27. F. Mansilla, K. Birkenkamp-Demtroder, M. Kruhoffer, F. B. Sorensen, C. L. Andersen, P. Laiho, L. A. Aaltonen, H. W. Verspaget and T. F. Orntoft, *Brit. J. Cancer*, 2007, **96**, 1896-1903.
28. W. R. Atchley, J. Zhao, A. D. Fernandes and T. Drüke, *P. Natl. Acad. Sci. USA*, 2005, **102**, 6395-6400.
29. N. D. Rubinstein, I. Mayrose and T. Pupko, *Mol. Immunol.*, 2009, **46**, 840-847.
30. T. Huang, P. Wang, Z.-Q. Ye, H. Xu, Z. He, K.-Y. Feng, L. Hu, W. Cui, K. Wang, X. Dong, L. Xie, X. Kong, Y.-D. Cai and Y. Li, *PLoS One*, 2010, **5**, e11900.
31. L.-L. Hu, S.-B. Wan, S. Niu, X.-H. Shi, H.-P. Li, Y.-D. Cai and K.-C. Chou, *Biochimie*, 2011, **93**, 489-496.
32. R. R. Sokal and B. A. Thomson, *Am. J. Phys. Anthropol.*, 2006, **129**, 121-131.
33. Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen and Y. Z. Chen, *Nucleic Acids Res.*, 2006, **34**, W32-W37.
34. W.-C. Chang, T.-Y. Lee, D.-M. Shien, J. B.-K. Hsu, J.-T. Horng, P.-C. Hsu, T.-Y. Wang, H.-D. Huang and R.-L. Pan, *J. Comput. Chem.*, 2009, **30**, 2526-2537.
35. Y. X. Li, Y. H. Shao and N. Y. Deng, *Protein and peptide letters*, 2011, **18**, 186-193.
36. *R Development Core Team, R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2009.
37. F. Petralia, P. Wang, J. Yang and Z. Tu, *Bioinformatics*, 2015, **31**, i197-i205.
38. C. C. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen and P. Macrossan, *IEEE/ACM Trans Comput Biol Bioinform*, 2011, **8**, 1580-1591.
39. W. Gu, A. R. Vieira, R. M. Hoekstra, P. M. Griffin and D. Cole, *Epidemiol Infect*, 2015, DOI: 10.1017/s095026881500014x, 1-9.
40. Z. Lin, C. M. Vicente Goncalves, L. Dai, H. M. Lu, J. H. Huang, H. Ji, D. S. Wang, L. Z. Yi and Y. Z. Liang, *Anal Chim Acta*, 2014, **827**, 22-27.
41. T. Bylander, *Machine Learning*, 2002, **48**, 287-297.
42. P. Smialowski, D. Frishman and S. Kramer, *Bioinformatics*, 2010, **26**, 440-443.
43. M. Shiiba, S. B. Arnaud, H. Tanzawa, E. Kitamura and M. Yamauchi, *J Bone Miner Res*, 2002, **17**, 1639-1645.
44. S. Li, L. M. Iakoucheva, S. D. Mooney and P. Radivojac, *Pac. Symp. Biocomput.*, 2010, **15**, 337-347.
45. P. Gramatica, *QSAR Combin. Sci.*, 2007, **26**, 694-701.
46. A. Golbraikh and A. Tropsha, *J Mol Graph Model*, 2002, **20**, 269-276.
47. V. Vacic, L. M. Iakoucheva and P. Radivojac, *Bioinformatics*, 2006, **22**, 1536-1537.
48. G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *Journal of Molecular Biology*, 1963, **7**, 95-99.
49. A. Mottaz, F. P. David, A. L. Veuthey and Y. L. Yip, *Bioinformatics*, 2010, **26**, 851-852.
50. P. Radivojac, P. H. Baenziger, M. G. Kann, M. E. Mort, M. W. Hahn and S. D. Mooney, *Bioinformatics*, 2008, **24**, i241-i247.
51. S. G. Priori, C. Napolitano, M. Gasparini, C. Pappone, P. Della Bella, U. Giordano, R. Bloise, C. Giustetto, R. De Nardis, M. Grillo, E. Ronchetti, G. Faggiano and J. Nastoli, *Circulation*, 2002, **105**, 1342-1347.