JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

In Silico Identification of Protein S-Palmitoylation Sites and Their Involvement in Human Inherited Disease

Shuyan Li,[†] Jiazhong Li,[‡] Lulu Ning,[†] Shaopeng Wang,[†] Yuzhen Niu,[†] Nengzhi Jin,[§] Xiaojun Yao,[†] Huanxiang Liu,[‡] and Lili Xi*[,∥]

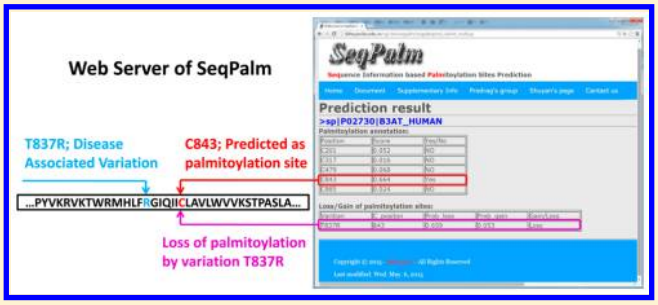
[†]Department of Chemistry and [‡]School of Pharmacy, Lanzhou University, Lanzhou, 730000, China

[§]Department of Technical Support, Gansu Computing Centre, Lanzhou, 730000, China

[∥]Department of Pharmacy, First Hospital of Lanzhou University, Lanzhou, 730000, China

Supporting Information

ABSTRACT: S-Palmitoylation is a key regulatory mechanism controlling protein targeting, localization, stability, and activity. Since increasing evidence shows that its disruption is implicated in many human diseases, the identification of palmitoylation sites is attracting more attention. However, the computational methods that are published so far for this purpose have suffered from a poor balance of sensitivity and specificity; hence, it is difficult to get a good generalized prediction ability on an external validation set, which holds back the further analysis of associations between disruption of palmitoylation and human inherited diseases. In this work, we present a reliable identification method for protein S-palmitoylation sites, called SeqPalm, based on a series of newly composed features from protein sequences and the synthetic minority oversampling technique. With only 16 extracted key features, this approach achieves the most favorable prediction performance up to now with sensitivity, specificity, and Matthew's correlation coefficient values of 95.4%, 96.3%, and 0.917, respectively. Then, all known disease-associated variations are studied by SeqPalm. It is found that 243 potential loss or gain of palmitoylation sites are highly associated with human inherited disease. The analysis presents several potential therapeutic targets for inherited diseases associated with loss or gain of palmitoylation function. There are even biological evidence that are coordinate with our prediction results. Therefore, this work presents a novel approach to discover the molecular basis of pathogenesis associated with abnormal palmitoylation. SeqPalm is now available online at http://lishuyan.lzu.edu.cn/seqpalm, which can not only annotate the palmitoylation sites of proteins but also distinguish loss or gain of palmitoylation sites by protein variations.



INTRODUCTION

S-Palmitoylation is a reversible covalent attachment of a saturated 16C palmitate to cysteines on the targeted proteins via the formation of a thioester linkage. It is a universal post-translational modification, usually occurring on membrane-associated proteins and playing critical roles in the regulation of protein trafficking, signaling, and behavior.[1] According to the studies in multiple organisms, palmitoylation is also found to be particularly important for a number of biological processes, including neuronal development,[2] cytoprotection,[3] parasite invasion,[4] ion-channel regulation,[5] osteoblast differentiation,[6] immune system regulation,[7] tumor suppression,[8] etc. Correspondingly, increasing evidence emerges to show the diverse roles of palmitoylation in pathophysiology. Many diseases are discovered to associate with the disruption of palmitoylation sites or palmitoyltransferase,[9] such as systematic amyloidosis,[10] osteoporosis,[6] neurodegenerative diseases (e.g., Alzheimer's disease,[11] Huntington disease,[12] neuronal ceroid lipofuscinosis,[13] and schizophrenia[14]), X-linked mental retardation,[15] diabetic vascular disease,[16] acute leukemia,[17] and a variety of other cancers.[18−21] Palmitoylated proteins are suggested to serve as biomarkers for cardiovascular disease,[22] and the palmitoylation/depalmitoylation related processes are also suggested to be potential therapeutic targets for several major diseases, e.g. hematologic cancer,[8] Huntington's disease,[12] schizophrenia,[23] and acute brain injury.[24]

Although there exists growing evidence for the importance of palmitoylation,[2,7,8,11,12] the identification of palmitoylation sites is not straightforward owing to the lack of distinct motifs on the substrates.[25] Therefore, multiple strategies, including mass spectrometric characterization,[26] metabolic labeling, click chemistry probe,[27,28] etc., were developed to recognize palmitoylation sites so as to further study and elucidate the molecular mechanisms and dynamics of palmitoylation. However, with the constant growth of protein data, these biological or chemical approaches are rather time- or money-consuming. Consequently, the computer aided methods for the

Received: May 13, 2015
Published: August 14, 2015

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

identification of palmitoylated proteins and corresponding sites become necessary.

Among the in silico methods for the identification of palmitoylation sites, the CSS-palm[29] and NBA-Palm[30] methods were presented as the pioneers in this field, but their prediction performances were merely adequate. Subsequently, Yang et al.[31] proposed a regularized biobasis artificial neural network (ANN) method, which made the performance slightly improved. The CSS-palm method was continuously updated to version 2.0, 3.0, and 4.0 to achieve further improved performance.[32,33] Meanwhile, Wang et al.[34] developed a CKSAAP-Palm method based on an encoding scheme as composition of *k*-spaced amino acid pairs and improved the performance of identification accuracy than former strategies. It got a remarkable Matthew's correlation coefficient (MCC) value as 0.754. After that, Hu et al.[35] proposed a IFS-Palm method based on the nearest neighbor (NN) algorithm using an assemble of protein structural and sequential features. Li et al.[36] designed a palmitoylation sites prediction method with a position weight matrices (PWMs) encoding scheme and support vector machines (SVMs). Shi et al.[37] also presented a predictor called WAP-Palm by SVMs method. But the identification performances for later three methods were lower than that of the CKSAAP-Palm method[34] considering the cross-validation results. After then, Fu et al.[38] proposed a new predictor on the random forest (RF) method and got a promising prediction performance with an MCC of 0.838 by the out-of-bag evaluation in the RF algorithm. However, this predictor cannot identify the palmitoylation sites within 10 residues along the C- and N-terminal of a protein sequence, where palmitoylation sites frequently occur. Therefore, the generalization ability of Fu's predictor is limited. Recently, Kumari et al. developed a PalmPred tool for this problem using sequence profiling information and SVMs as well.[39] Afterward, Pejaver et al.[40] developed a unified PTM sites predictor called ModPred including 23 types of PTMs where palmitoylation was also included. The performance of the last two methods on palmitoylation sites prediction were still lower than that of the CKSAAP-Palm method.[34]

These prediction methods have already been proven useful for the studying of molecular function in biological experiments.[38,39] For example, Zoltewicz et al.[41] used CSS-Palm2.0 to predict the palmitoylation sites on peripheral myelin protein 22 (PMP22). Their prediction results were proven later in the biochemical studies,[41] which indicated that the palmitoylation of PMP22 at C85 is critical for the role of this protein in modulating epithelial cell shape and motility. However, in order to further study the association between the disruption of palmitoylation sites and the pathology of diseases, a more reliable predictor with better generalization ability is still highly desired.

To achieve this goal, we propose a new method, SeqPalm, for the identification of palmitoylation sites. A large amount of possible types of features was explored in the first round of this study and it was finally settled with three categories: amino acid composition, autocorrelation of amino acid physicochemical properties and amino acid position weighted matrices. Synthetic minority oversampling technique (SMOTE)[42] was utilized to address the unbalance problem between positive and negative samples. With the final features selected by the RF algorithm, this method achieved a favorable prediction performance that was evaluated by both cross validation and external validation. SeqPalm was then applied to predict the gain and loss of palmitoylation sites upon single amino-acid variations and hence to link human inherited diseases with the corresponding molecular mechanisms.

MATERIALS AND METHODS

Palmitoylation Data Set. The palmitoylation data were assembled from UniProt (release 2014-03) and HPRD (release 9) databases. For UniProt, only experimentally verified information was extracted, including 162 palmitoylated proteins and 282 palmitoylation sites. For HPRD, 74 palmitoylation proteins and 149 palmitoylation sites were retrieved. The proteins with identical sequences were integrated as a unique protein and all the entries of palmitoylation sites inside of these proteins were also integrated as unique entries in the unique protein as well. Then, after the elimination of redundant data, 204 proteins with 2400 cysteine residues (361 are palmitoylation sites) were obtained for this study.

Since there is biological evidence of kinase-substrate binding within neighborhood residues of the modification site,[43] we believe that the palmitoyltransferase is also binding within the neighborhood residues of a cysteine site on the substrate. Therefore, the fragment with the cysteine of interest in its center was used as a palmitoylation sample to create features. In this work, the fragment length as 51 was utilized here (within ±25 residues). Then, if the cysteine in the center of a fragment had a palmitoylation entry, then this fragment was treated as a positive sample. Otherwise, it was assigned as a negative sample. All the samples with the same sequence of fragment were integrated as a unique sample. Then, the negative samples with the same sequence as positive samples were eliminated from negative sets. In the end, 361 positive samples and 2032 negative samples were gathered.

After then, the whole protein set was randomly divided into an internal cross-validation set and an external validation set by a ratio of 9 to 1 based on protein level. Thus, 184 proteins with 332 cysteine-palmitoylation sites and 1824 nonpalmitoylation cysteines were gathered into cross-validation set. The remaining 20 proteins with 29 cysteine-palmitoylation sites and 208 nonpalmitoylation cysteines were assigned to the external validation set. All the detailed information on data sets can be accessed in Supporting Table S1.

Human Disease Associated Variation Data Set. Human variation data were retrieved from the SwissVar database (UniProt release 2014_03). The data contained 69 086 variants from 12 529 human proteins. Of those, 24 608 variants were associated with human inherited diseases.

In accordance with our previous definition of samples, a variant occurring in the neighborhood of a cysteine site (within ±25 residues) was assigned as a palmitoylation-related variation. If the variation had disease-associated entries, the fragment where this variation was located in and with a cysteine in the center was treated as a disease-related sample. The neutral polymorphism samples were generated likewise. Therefore, 32 443 disease-related samples and 45 520 neutral polymorphism samples were included in this study. One should notice that there might be more than one cysteine in the neighborhood of a variant, hence one variant may produce multiple samples with different cysteines.

Feature Generation. As described above, each sample was represented by a 51-residue fragment where the cysteine of interest was in the center. If the upstream or downstream residue number was less than 25, the symbol X was used to represent the missing amino acid in the fragment. Following

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

this scenario, each fragment was transformed into three categories of features, as below:

I. Amino Acid Composition. Amino acid compositions were incorporated via two groups of features: single amino acid composition in a fragment and the composition of *k*-spaced amino acid pairs (CKSAAP).[44] A single amino acid composition is the fraction of each residue type in a sequence fragment. CKSAAP was first proposed by Chen et al.[44] for the prediction of protein flexible/rigid regions and was proven useful for the prediction of O-glycosylation sites[45] and palmitoylation sites.[34] It was calculated by considering the fraction of amino acid pairs that are separated by *k* amino acids within a protein (there are 441 possible pairs, e.g. AA, AC, AD, ..., XX). Note that these 441 pairs are derived from 21 amino acids (20 normal amino acid and one hypothetical amino acid "X"). We refer to such a feature vector as $(C_{AkA}C_{AkC}C_{AkD} \ldots C_{AkX})_{441}$. For instance, $C_{A3C} = N_{A3C}/(N-1)$, where N_{A3C} is the number of occurrences of the AC amino acid pair that is separated by three amino acids and N is the residue length of the peptide. In this work, N was set to 51 and k = 0, 1, ..., 9 were jointly considered. In total, 4431 compositional features were generated.

II. Autocorrelation of Amino Acid Physicochemical Properties. Thirteen amino acid properties were utilized for this feature set, including: (1) the hydrophobicity scale,[46] (2) the average flexibility index,[47] (3) the polarizability parameter,[48] (4) the free energy of solution in water, (5) the residue accessible surface area for a tripeptide,[49] (6) the average volumes of residues,[50] (7) the steric parameter[51] and (8) the relative mutability,[19] (9) the polarity factor, (10) the secondary structure factor, (11) the molecular volume factor, (12) the codon diversity factor and (13) electrostatic charge factor. The last five properties were derived from the literature[52−54] and assembled by Hu et al.[35] as transformed attributes. The properties for residue X were set to zeros.

Two autocorrelation calculating algorithms, as Geary autocorrelation and normalized Moreau-Broto autocorrelation were adopted here. The Geary autocorrelation features[55] 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 distance between two residues in protein sequence, P_i and P_{i+d} are the particular property value of the residues at position i and $i + d$, respectively. $\overline{P}$ is the average value of P_i and N is the peptide length.

The normalized Moreau-Broto autocorrelation features are defined as

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

where d, N, P_i and P_{i+d} are previously defined.

This block of features was inspired by both CKSAAP-Palm and PROFEAT methods.[56] There are 13 × 30 × 2 = 780 features in this block.

III. Amino Acid Position Weighted Matrices (PWMs). For a given data set of fixed-length sequence fragments, each amino acid at each position is associated with its frequency of occurrence. Using this frequency in both palmitoylated and nonpalmitoylated sets of fragments in training set, two position-weighted matrices (PWMs)[56,57] are calculated. Each row of the matrix corresponds to one kind of amino acids and every column corresponds to the position in the peptide. The element in the ith row and jth column of the matrix is defined as

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

where N_{ij} is the number of times the ith amino acid was observed in the jth position of the peptide and N_{pep} is the number of peptides in the whole data set, i = 1, 2, 3, ..., 21, j = 1, 2, 3, ..., 51. Since two PWMs were built, 51 × 2 = 102 features were extracted in this group.

After all, 5313 features were generated. In order to reduce redundant information, constant features and highly correlated features were excluded from the feature space. In particular, if any two features of all the samples shared a correlation coefficient greater than 0.85, one of them was removed randomly. The redundancy elimination was performed by software QSARINS.[58] Finally, 4959 features remained for the next step.

Generation of Balanced Samples. The unbalanced number of positive and negative samples is an inherent problem for the accurate prediction of palmitoylation sites. SMOTE method[42] was used in this work to fix this problem. With the generated features of positive sample in training set, SMOTE utilized a algorithm based on k-nearest neighbors method to create extra 500% positive samples from the current data. Thus, after applying SMOTE, a total of 1992 positive samples in training set from initial 332 samples were generated, resulting in a much less unbalanced data set. SMOTE was performed by the DMwR package of R.

Feature Selection and Modeling. Aiming to further reduce the dimension of the feature space and meanwhile find the restrict feature number that lead to the best classification model, feature ranking and modeling by RF method[59] were employed. The RF method was first introduced by Leo Breiman[59] and then proved to be a very powerful classification method in the fields of chemometrics and bioinformatics.[60−64] RF is a classifier consisting of collection of tree-structured classifiers with two major advantages being (1) using an out-of-bag (OOB) method[65] to monitor error, strength, and correlation and (2) measuring variable importance through permutation. With the whole set of features to build an RF classification model based on the cross-validation data set, the importance for each feature of its association with prediction target was presented. Then, the ranking importance of features was listed. With an increasing number of top ranking features, different RF models are built as well to select the best model which had the least feature space and equivalent prediction performance with whole feature space. The feature selection process also helped us to locate the key features of protein sequences that distinguished a palmitoylation site. RF was implemented using R package randomForest v 4.6-7.

In order to perform a correct supervised feature selection and to get a reliable prediction performance, the feature ranking and modeling processes in this work were proceeded only on the internal cross-validation set without any involvement of the external validation set as emphasized in Smialowski's work.[66]

Model Evaluation. The model evaluation here was performed on both the internal validation data set using 10-fold cross validation and on the external validation set. For 10-fold cross-validation, the internal validation data set in this work

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

was randomly split into 10 nonoverlapping partitions, each partition was set as test set once while the remaining data were set as training data. Thus, all the samples will be in the test data exactly once. This process was repeated 10 times with each time including different random partitions in order to obtain stable estimates of the classification performance. The averaged prediction result of validation sets over 10 runs was taken as the final 10-fold cross-validation result.

Five frequently used indicators were utilized here to estimate the prediction performance of SeqPalm, including sensitivity (Sens), specificity (Spec), accuracy (ACC), MCC, and area under receiver operating characteristic curve (AUC). Receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier system as 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. AUC, as area under ROC curve, can give a comprehensive evaluation of a prediction method.

Application of SeqPalm on Human Inherited Disease. According to our previous study, the disruption of palmitoylation sites is enriched in disease-associated than in neutral variants,[67] which suggests a link between the disruption of palmitoylation and inherited disease. Besides, there are already biological evidence suggesting that the disruption of palmitoylation sites by variations may induce malfunction to essential proteins and hence to cause disease.[3,13,68,69] Therefore, SeqPalm was then applied on inherited disease-associated variants in order to further study this link and explore potential pathogenesis.

Roughly, SeqPalm estimates the probability that a residue s_i can be palmitoylated, as $P(s_i = s_i^p|S)$, given the protein sequence S.

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

$$P_l(s_i) = P(s_i = s_i^p|S) \times (1 - P(s_i = s_i^p|S_{xjy})) \quad (4)$$

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, V, W, X, Y). Here, symbol X is the assigned residue for short fragment as described in the section of feature generation, while others are 20 normal residues. Note that only the variation that exists in the neighborhood (±25 residues) of a cysteine is considered to have influence on gain or loss of palmitoylation sites here. In an extreme case, if the variation directly mutates the palmitoylation site (i.e., $i = j$), the mutant residue will change from C to other residues, which will definitely not be S-palmitoylated. In this case, $P(s_i = s_i^p|S_{xjy}) = 0$ and the probability of a loss of palmitoylation $P_l(s_i)$ equals to $P(s_i = s_i^p|S)$.With the same principle, the probability of a gain of palmitoylation at residue s_i is defined as

$$P_g(s_i) = (1 - P(s_i = s_i^p|S)) \times P(s_i = s_i^p|S_{xjy}) \quad (5)$$

The threshold that identifies a gain or loss of palmitoylation sites is set to the same value as the cutoff value of SeqPalm.

RESULTS

Feature Selection and Predicted Performance. Different models were built on increasing number of top ranking features and 10-fold cross-validation results for these models were shown in Figure 1. The MCC, ACC, and AUC scores all achieved approximately peak values as 0.917, 0.959, and 0.993, respectively, when top 16 features were selected. Then, the AUC value remained basically the same level while ACC and MCC values became apparently fluctuating when more features were included. Since the less of the selected number of features in a predictor, the less risk there will be for the overfitting problem to our best knowledge. Therefore, in order to avoid the overfitting problem and hence to get better generalization ability, the 16 top ranked features were selected to build our final SeqPalm predictor. Information of these selected features were listed in Table 1. The detailed prediction performance of SeqPalm is shown in Tables 2 and 3. In these two tables, the results of the external validation set are fairly stable compared with the cross validation results, which is an emblematical sign of favorable generalization ability for a classifier.[70]

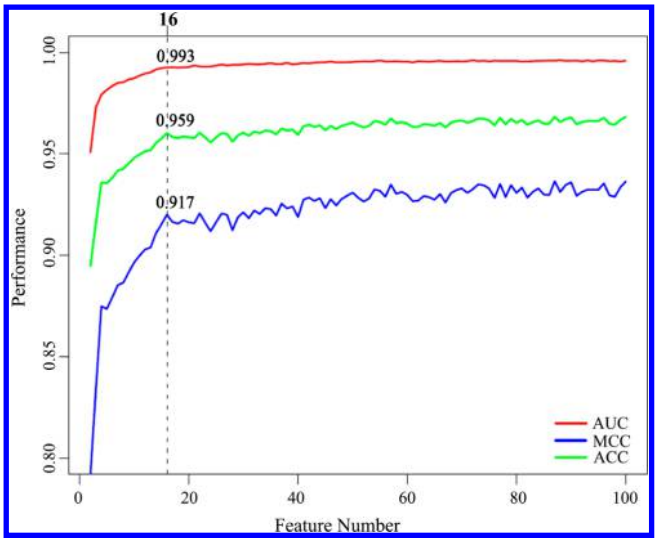


Figure 1. Cross-validation results of models which were built on top ranking features.

Performance Comparison. Since previous methods on the same topic used several different data sets and different evaluation strategies, it is difficult to get a rigorous performance comparison with all published methods. Therefore, first, the cross-validation results reported in previous works were compared with SeqPalm (Table 2) and then all the well-established methods of which the software was either available online or ready to be downloaded were tested on the same validation set against our SeqPalm method. By this criterion, CSSpalm4.0 (The latest version of CSSpalm2.0,[32] at http://csspalm.biocuckoo.org/), CKSAAP,[34] ModPred,[40] PalmPred,[39] and Fu's method[38] were utilized here to perform the comparative assessment on the same validation set. (Table 3)

Besides, the data set partitions inside of the cross validation process were usually split based on sites level as most of the previous predictors did.[29−36,39] Therefore, we also used the cross-validation on site level to get the final SeqPalm model. However, this partition could also be performed on protein level as it was in Pejaver's work[40] to avoid the intraprotein bias. Therefore, the cross-validation on protein level was also performed in this work so as to get a more fair and rigorous comparison.

Our SeqPalm method got a great improvement on MCC score in the cross-validation (Table 2) and also got the best performance on the external validation compared with the best score of previous predictors. For the first time, sensitivity and

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

Table 1. Top Ranking Features That Were Selected for the Final SeqPalm Model

Rank	Feature	Block	Note
1	GearyAuto_0_4	II	Geary autocorrelation of the hydrophobicity scale of amino acids with *d* value as 4
2	T	I	single amino acid composition of threonine
3	NorMBAuto_4_1	II	normalized Moreau−Broto autocorrelation of the free energy of solution in water of amino acids with *d* value as 1
4	C	I	single amino acid composition of cysteine
5	M	I	single amino acid composition of methionine
6	posPWM.27	III	positive position weighted matrices value at position 27
7	posPWM.12	III	positive position weighted matrices value at position 12
8	L4C	I	amino acid pair composition of leucine and cysteine pair that are separated by 4 amino acids
9	posPWM.22	III	positive position weighted matrices value at position 22
10	L6C	I	amino acid pair composition of leucine and cysteine pair that are separated by 6 amino acids
11	GearyAuto_0_7	II	Geary autocorrelation of the hydrophobicity scale of amino acids with *d* value as 7
12	GearyAuto_0_1	II	Geary autocorrelation of the hydrophobicity scale of amino acids with *d* value as 1
13	NorMBAuto_0_4	II	normalized Moreau−Broto autocorrelation of the free energy of solution in water of amino acids with *d* value as 4
14	GearyAuto_3_30	II	Geary autocorrelation of the hydrophobicity scale of amino acids with *d* value as 1
15	posPWM.25	III	positive position weighted matrices value at position 25
16	posPWM.13	III	positive position weighted matrices value at position 13

MCC score step over 0.900 in cross validation results in the same field. As for the external validation, only ModPred can get a comparable results with SeqPalm method, while the other predictors were significantly lower than SeqPalm, especially for sensitivity and MCC values. This result suggested that SeqPalm has the optimum performance and good generalization ability on the prediction of palmitoylation sites.

Loss or Gain of Palmitoylation Sites in Human Inherited Disease. Since the distribution of training data for SeqPalm is unlikely identical to the human variation data, we addressed a "high confidence" loss or gain of palmitoylation sites, inspired by Radivojac's work[71] 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 0.5 (higher than the cutoff value of SeqPalm) was assigned as a high confidence loss or gain of palmitoylation site. Correspondingly, only these sites were then analyzed in the further procedure. Following these steps, 138 sites were identified as high confidence gain and 105 as high confidence loss of palmitoylation sites by the influence of variations. The detailed information is listed in Supporting Table S2. In order to discover the relationship between loss or gain of palmitoylation sites with inherited disease, SeqPalm was also applied to the neutral polymorphism samples. In this way, there are 153 high confidence gain and 145 high confidence loss of palmitoylation sites in neutral polymorphism samples. Evaluated by the Chi-square test, both the probability that for gain (P-value = 1.18×10^{-37}) and loss (P-value = 0.429) of palmitoylation sites in disease associated variations are significant higher than neutral polymorphisms.

Seen in Table S2, loss or gain of palmitoylation sites is not only occurring consequently with the loss or gain of cysteine sites but also occurring because of the variation in the neighborhood of sequence around cysteine with a probability of occurrence about 34% (83/243 in Table S2). Each loss or

Table 3. Performance Comparison of SeqPalm with Other Methods on the Same External Validation Set

method	threshold	ACC	sens	spec	MCC
SeqPalm	ntree = 500, mtry = 4, cutoff = 0.410	**0.941**	**0.828**	**0.957**	**0.742**
Fu's method[38,a]	n.a.	0.937	0.517	0.995	0.669
ModPred[40]	PSSM, medium	0.920	0.897	0.923	0.703
	PSSM, low	0.890	0.897	0.889	0.636
	PSSM, high	0.895	0.241	0.986	0.370
PalmPred[39]	0.1	0.916	0.310	1.000	0.532
	−0.4	0.384	1.000	0.298	0.222
CSSpalm4.0[32]	high	0.890	0.759	0.909	0.578
	medium	0.865	0.793	0.875	0.541
	low	0.848	0.828	0.851	0.527
CKSAAP-Palm[34]	high	0.907	0.483	0.966	0.518
	low	0.869	0.655	0.899	0.485

[a]Note that a shortage was found in Fu's method in the external validation test that it cannot identify the palmitoylation sites within 10 residues at both C- and N-terminal of a protein sequence. Therefore, the unidentified positive samples in this situation were treated as false positive sample during the calculation. All the other predictors do not have this problem.

Table 2. Performance Comparison of SeqPalm with Other Methods by Cross-Validation on Both Site and Protein Levels

level	method	evaluation strategy	ACC	sens	spec	MCC
site	**SeqPalm**	10-fold CV	**0.959**	**0.954**	**0.963**	**0.917**
	CKSAAP[34]	10-fold CV	0.935	0.884	0.942	0.754
	PalmPred[39]	Jackknife	0.920	0.792	0.943	0.710
	Fu's method[38]	OOB method	0.919	0.889	0.947	0.838
	IFS-Palm[35]	Jackknife	0.906	0.686	0.947	0.638
	CSSpalm2.0[32]	Jackknife	0.896	0.772	0.924	0.671
	PWM_SVM[36]	10-fold CV	0.885	0.886	0.886	0.686
	NBA-Palm[30]	Jackknife	0.867	0.675	0.923	0.610
	WAP-Palm	10-fold CV	0.860	0.815	0.905	0.723
	Biobasis ANN[31]	Jackknife	n.a.	0.744	0.960	0.682
protein	**SeqPalm**	10-fold CV	**0.889**	**0.715**	**0.921**	**0.602**
	ModPred[40]	10-fold CV	n.a.	0.679	0.903	n.a.

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

gain of palmitoylation site by variations may imply a potential pathogenic for the corresponding inherited disease.

For example, among these variations, 17 of them (nos. 3, 4, 7, 8, 18, 21, 27, 57, 87, 88, 118, 136, 144, 145, 153, 196, and 238 in Table S2) within 7 proteins (NEU2_HUMAN, INS_HUMAN, HNF1A_HUMAN, HNF1B_HUMAN, V2R_HUMAN, AQP2_HUMAN, IRK11_HUMAN) were associated with diabetes and were also associated with predicted loss or gain of palmitoylation sites. Since there is already evidence proving that dysfunction of palmitoylation will cause diabetes,[16] these novel loss/gain of palmitoylation information may provide supplementary pathogenic mechanisms about diabetes.

Besides, six of these variations (nos. 133,138,183,193, 210, and 220 in Table S2) from collagen (CO1A1_HUMAN) predicted with gain of palmitoylation sites are related with osteogenesis imperfecta, which is a genetic disorder of increased bone fragility, low bone mass and other connective-tissue manifestations.[72] Previous study showed that palmitoylation plays an important role in osterix expression, osteoblast differentiation and associated with osteoporosis.[6,10] Thereby, gain of palmitoylation sites may also play critical role in osteogenesis imperfecta with the same mechanism.

In addition, recent study suggests the potential association of abnormal palmitoylation with dopaminergic diseases such as Parkinson disease, schizophrenia, attention deficit hyperactive disorder, and drug abuse.[73] Our study showed another proof for the association between abnormal palmitoylation and Parkinson disease (no. 126 in Table S2). In addition, Alzheimer disease is also related with abnormal of palmitoylation by our prediction results (no. 33 in Table S2), which is coordinated with previous evidence.[11]

Further Evaluation of the SeqPalm Method. In order to further prove our prediction results, we did a in-depth reference search and performed a further evaluation for SeqPalm method. Henry et al.[74] published a work recently and indicated that the immunity-related GTPase family M protein 1 (IRGM1) should be palmitoylated within the tight cluster of C371, 373, 374, 375 near the C-terminus of the protein by in vivo test. Protein IRGM1 was not included in the training database of SeqPalm, therefore the prediction results of this protein by SeqPalm should be of confidence without previous introduced bias. Among these sites, C371, C373, and C375 were predicted to be palmitoylation sites by SeqPalm method (red square, Figure 2). Besides, sites C257 and C258 were predicted without palmitoylation (blue square, Figure 2), which was also coordinate with the results of in vivo test in Henry's work, although these two sites were predicted as palmitoylation sites by CSS-Palm 2.0 method.[74] Henry et al.[74] also discovered that "The palmitoylation mutant, Irgm1(C371,373,374,375A), displayed a modest but statistically significant decrease in its ability to shift the mitochondrial equilibrium toward punctate forms and away from tubular forms". Variations of C371A, C373A, and C375A were also predicted as loss of palmitoylation sites by our SeqPalm method as well (purple square, Figure 2).

Besides, Greaves et al.[13] indicated that the variations of L115R and L116R human cysteine-string protein (CSP, UniProt ID: DNJC5_HUMAN) can lead to the loss of palmitoylation function and hence to cause neuronal ceroid lipofuscinosis. Meanwhile, variation L115A did not affect the function of CSP. By our SeqPalm method, the sites C122, 127, 128, 132, and 133 of protein CSP were all predicted to be

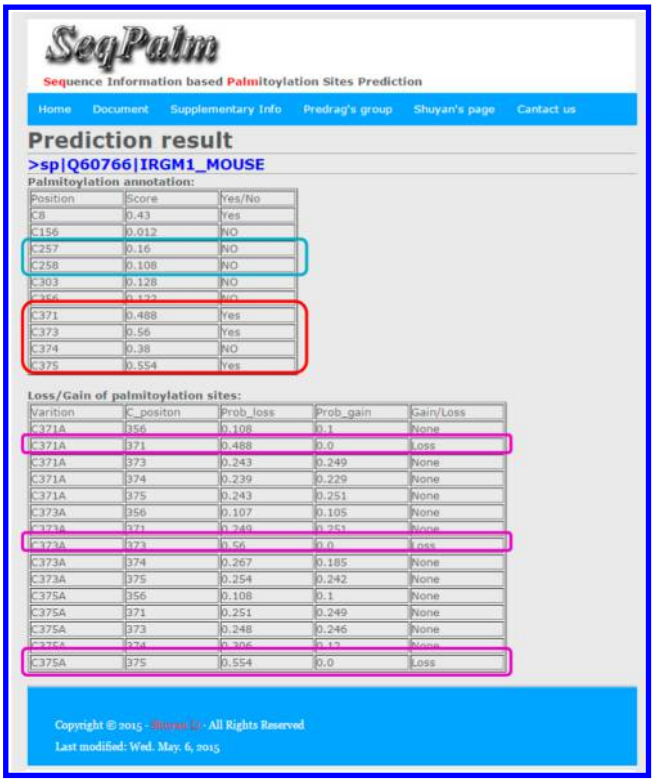


Figure 2. Prediction results of SeqPalm for protein IRGM1.

palmitoylation sites (red ellipse, Figure 3A). The single variation of L115A had very limited influence on the palmitoylation, where most of the palmitoylation sites remained in the mutant (Figure 3B). Although C128 lost palmitoylation, C119 gained palmitoylation instead, which should exhibit the same migration profile on SDS gels compared with wild-types since they had the similar molecular weight. This result is precisely coordinate with the experiment results in Greaves's work[13] and also presents more detailed location information on the corresponding palmitoylation sites. Furthermore, all of the palmitoylation sites were predicted to be lost in the mutant (purple ellipse, Figure 3C) with both variations as L115R and L116R. This result is also coordinate with Greaves's conclusions as that with the variations of L115R and L116R, there is a clear absence of palmitoylated monomeric form of protein CSP.[13] The prediction results from CSS-Palm method only suggested a minimal effect of the mutations of L115R and L116R on the palmitoylation of protein CSP.[75] Therefore, SeqPalm is proved to be a reliable method in the study of association between the disruption of palmitoylation sites and disease-associated variations.

Web Server of SeqPalm. The interface of SeqPalm is shown as Figure 4. There are two kinds of jobs that can be done in this server as follows:

I. Annotation of Palmitoylation Sites in Proteins. Users can input single or multiple protein sequences in a FASTA format into the first textbox as showed in Figure 4 and, then, leave the second textbox empty and press the submit button. After that, all of the cysteines in the proteins will be predicted whether to be a palmitoylation site or not. After the calculation, the outputs will be listed in the result page as shown in Figure 2.

II. Distinguishing Loss or Gain of Palmitoylation Sites by Protein Variations. Users can input only one wild-type protein sequence in the first textbox and input single or multiple

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

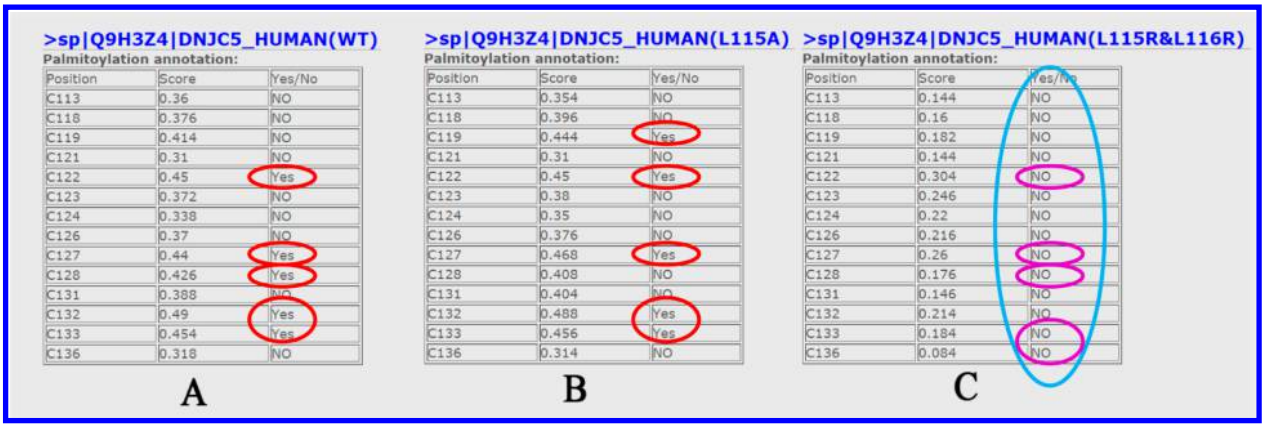


Figure 3. Annotation results of SeqPalm for protein CSP. A. Wild type. B. Mutant with variation of L115A. C. Mutant with variations of L115R and L116R.

Figure 4. Interface of SeqPalm web server with protein IRGM1 as input.

protein variations in the second textbox. The input format of variation is "T130C", where the first letter stands for the wild-type of the residue, the last letter stands for the mutant-type of the residue, and the number in between is the position in the protein sequence where this variation occurs. Multiple input of variations are separated by comma. After pressing the submit button, the results of loss or gain of palmitoylation sites by the variations will showed up below the annotation results as shown in Figure 2.

One should note that each site in a input of multiple variations will be treated as single variation mutant each time. Then, SeqPalm will evaluate the influence of the cysteines in the neighborhood of this single variation rather than the multiple variations together in the mutant. For example, if the input of variation is "C371A, C373A, C375A" with input sequence of protein IRMG1 as shown in Figure 4, SeqPalm will distinguish the loss or gain of palmitoylation sites in three mutants, each with a single variation inside. If someone wants to evaluate the gain or loss of palmitoylation sites by a mutant with multiple variations together, they should input the wild-type protein sequence and the mutant sequence in the first textbox separately, leave the second textbox empty, and then check the difference in the annotation results for both wild type and mutant, as shown in Figure 3.

DISCUSSION

Sample encoding scheme and model construction algorithms are two critical keys for prediction method. Many types of machine learning algorithms have already been employed in the prediction studies of palmitoylation sites such as naïve bayes algorithm,[30] ANN,[31] SVMs,[34,36] NN algorithm,[35] and RF method.[38] However, considering the predictors that can recognize all the palmitoylation sites along whole protein sequence, the identification performance comes to a standstill right after the CKSAAP method, which was published six years ago in 2009. This phenomenon implied an idea that the bottleneck for the computational identification of palmitoylation sites may not be the machine learning methods but may be the encoding schemes of features. Actually, many other types of feature-encoding schemes other than the three categories that were used in SeqPalm were included at first in this work, e.g. amino acid binary encoding scheme,[76] pseudo amino acid composition,[76] etc. However, none of the features from other encoding schemes was ranked at the top of the importance list as key features in this study. Therefore, the final model was only built on the three categories of composed features. We also believed that this prediction performance will be further improved when other suitable encoding schemes were introduced.

For the machine learning method, SVMs was also utilized at first in this work. However, it seemed that the predictions from SVMs in this study were easily failed with overfitting performance somehow, which presented a satisfactory cross-validation results but very poor external validation results. Therefore, the SVMs method was eliminated from our option. Meanwhile, the RF method presented a good prediction performance with both satisfactory cross-validation and external validation results. Therefore, RF algorithm was chose in this work. It is also suggested that researchers of predictors should be aware of the overfitting problems in cross-validation results as Alexander et al. emphasized.[77] It is also the reason why we were very cautious with the compare of cross-validation and

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

external validations results so as to make sure of that our SeqPalm method did not have the overfitting problem.

Besides, there is high possibility that palmitoyltransferases identify the palmitoylation sites mainly based on sequence information according to the success of SeqPalm. In other words, the determinant that whether a site will be palmitoylated or not may be encoding in a sequence window with cysteine in the center. Actually, different window sizes from 5 to 61 were also tested in this work with an increasing step of 2 in the first round of this study. Then, the sequence length as 51 got the best AUC score as 0.987 using the whole feature space by the default parameters of RF method. According to the previous studies, only up to 25 of the peptide length were tried before.[38] Therefore, although palmitoylation has no distinct motif in sequence of substrate unlike other types of lipidation, a fragment with 51 residues centered by cysteine may possess a latent motif for the identification of palmitoylation site. Correspondingly, the 16 features (Table 1) that were extracted from sequence information in SeqPalm method is probably the most promising option as the latent motif of palmitoylation for now. Again, we believe that the more intuitive grasp of palmitoylation motif should be discovered in the future with more suitable encoding scheme were introduced in this field.

For the key features that were selected for the final SeqPalm model (Table 1), hydrophobicity scales should be the most important amino acid property for the identification of palmitoylation sites, since one-fourth of the 16 features (nos. 1, 11, 12, and 13) are about hydrophobic properties. This result is highly coordinate with the main function of palmitoylation, which is to enhance the hydrophobicity of proteins and contribute to their membrane association.[78] Besides, the composition of threonine, cysteine and leucine, whether in single amino acid composition (nos. 2 and 4) or in pairs composition (nos. 8 and 10), is critical as well. In order to intuitively visualize the association between amino acid composition and the identification of palmitoylation sites, Two Sample Logo (TSL)[79] method was also utilized here. Seen from the TSL results (Figure 5), cysteine and leucine are

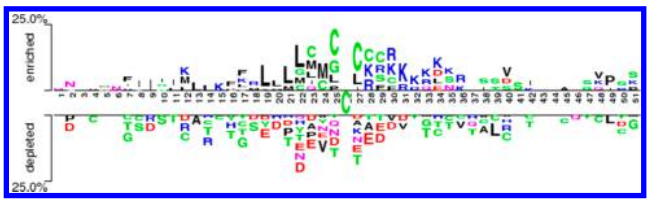


Figure 5. Amino acids enrichment and depletion at different positions of palmitoylation samples.

significantly enriched around the central cysteine, while threonine is depleted in more than one-third of positions in the fragment, which is coordinate with our results. As seen from Table 1, the positions of 12, 13, 22, 25, 27 (nos. 6, 7, 9, 15, and 16 in Table 1) play important roles for the identification. Comparing this result to Figure 5, positions 22, 25, and 27 are just around the central cysteine; all of which are enriched with leucine or cysteine and significantly depleted with threonine. But position 12 and 13 had no direct association between the amino acid composition and palmitoylation identification according to Figure 5. This implies that these two positions may also have potential but important associations with palmitoylation, although the mechanism is unknown yet.

For the disease association analysis, our previous study presented that about 5% of disease-associated variations may affect known PTM sites and only 1% of them directly mutated the modification site.[67] Then in this study, by the application of SeqPalm method, 243 high confidence disruption of palmitoylation sites were discovered with association with human inherited disease. Currently, although there are only limited experimental results presenting direct variations on palmitoylation sites and studying their associations with disease, it is fortunate that two works[13,74] are found to be the further proof the reliability of SeqPalm as described in the Results section. Our study for the first time comprehensively reveal the association of disruption of palmitoylation site with disease-related variations in silico and hence to decipher of certain pathogenesis, especially for monogenic disease. Therefore, if the disruption of genes of palmitoyltransferase is related with certain disease, such as osteogenesis disorder,[9,10,80] and the same disease is also directly associated with the disruption of palmitoylation site (nos. 1, 4, 6, 10, 14, 16, 19, 21, and 24 in Table S2), the linkage between palmitoylation and the pathogenesis of the very disease will be enhanced and even clinched. In this case, palmitoylation related process would be a potential therapeutic target for this kind of disease.

CONCLUSIONS

This work presented the most reliable annotation tool for protein S-palmitoylation sites identification, called SeqPalm, which can further distinguish the loss or gain of palmitoylation sites by protein variations. The online web server of SeqPalm was provided at http://lishuyan.lzu.edu.cn/seqpalm. With the application of SeqPalm, 243 disease-associated variations were discovered for the first time that they were associated with loss or gain of palmitoylation sites. Among the associated diseases, diabetes and osteogenesis imperfecta were found highly related with disruption of palmitoylation, which may provide some insights into the pathological mechanism on molecular lever of these inherited diseases. In the further, SeqPalm is expected to be further developed as a comprehensive in silico platform of palmitoylation study, including functions like calculating the difference of protein−ligand binding affinity between wild type and mutant with disruption of palmitoylation sites, providing suggestions of therapeutic targets of related disease, etc.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00276.

Supplementary tables, including (S1) all the palmitoylation data sets and (S2) high confidence gain/loss of palmitoylation sites associated with human inherited disease (PDF)

AUTHOR INFORMATION

Corresponding Author

*Phone: +86-0931-8356510. E-mail: xill@lzu.edu.cn. Mailing address: Room 103, Building of Pharmacy, First Hospital of Lanzhou University, Lanzhou, 730000, China.

Author Contributions

S.L. and L.X. conceived and designed this work. S.L. and J.L. constructed the SeqPalm tool and built the web server. S.W. produced the balanced data with SMOTE method. Y.N. and N.J. validated the built model. S.L., L.N., H.L., and X.Y.

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

analyzed the data and wrote the paper. All authors have given approval to the final version of the manuscript

Funding

This work was supported by National Natural Science Foundation of China (no. 21405068 to S.L., No. 21205055 to J.L., and No. 21305057 to L.X.), and the Fundamental Research Funds for the Central Universities (no. lzujbky-2015-31 to S.L. and no. lzujbky-2013-153 to L.X.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Predrag Radivojac for his great help and advices for the quality improvement of this work, Dr. Paola Gramatica and Dr. Nicola Chirico for providing software QSARINS, and Dr. Fiona Luke and Dr. Aju Wang for language checking. In addition, we also like to express gratitude to all the anonymous reviewers who helped to make this work better.

■ REFERENCES

(1) Martin, B. R.; Wang, C.; Adibekian, A.; Tully, S. E.; Cravatt, B. F. Global profiling of dynamic protein palmitoylation. *Nat. Methods* **2011**, *9*, 84−89.

(2) Fukata, Y.; Fukata, M. Protein palmitoylation in neuronal development and synaptic plasticity. *Nat. Rev. Neurosci.* **2010**, *11*, 161−175.

(3) Sardjono, C. T.; Harbour, S. N.; Yip, J. C.; Paddock, C.; Tridandapani, S.; Newman, P. J.; Jackson, D. E. Palmitoylation at Cys595 is essential for PECAM-1 localisation into membrane microdomains and for efficient PECAM-1-mediated cytoprotection. *Thromb. Haemostasis* **2006**, *96*, 756−766.

(4) Jones, M. L.; Collins, M. O.; Goulding, D.; Choudhary, J. S.; Rayner, J. C. Analysis of protein palmitoylation reveals a pervasive role in plasmodium development and pathogenesis. *Cell Host Microbe* **2012**, *12*, 246−258.

(5) Shipston, M. J. Ion channel regulation by protein palmitoylation. *J. Biol. Chem.* **2011**, *286*, 8709−8716.

(6) Leong, W. F.; Zhou, T.; Lim, G. L.; Li, B. Protein palmitoylation regulates osteoblast differentiation through BMP-Induced osterix expression. *PLoS One* **2009**, *4*, e4135.

(7) Ivaldi, C.; Martin, B. R.; Kieffer-Jaquinod, S.; Chapel, A.; Levade, T.; Garin, J.; Journet, A. Proteomic analysis of S-Acylated proteins in human B cells reveals palmitoylation of the immune regulators CD20 and CD23. *PLoS One* **2012**, *7*, e37187.

(8) Xu, J.; Hedberg, C.; Dekker, F. J.; Li, Q.; Haigis, K. M.; Hwang, E.; Waldmann, H.; Shannon, K. Inhibiting the palmitoylation/depalmitoylation cycle selectively reduces the growth of hematopoietic cells expressing oncogenic Nras. *Blood* **2012**, *119*, 1032−1035.

(9) Korycka, J.; Łach, A.; Heger, E.; Bogusławska, D. M.; Wolny, M.; Toporkiewicz, M.; Augoff, K.; Korzeniewski, J.; Sikorski, A. F. Human DHHC proteins: A spotlight on the hidden player of palmitoylation. *Eur. J. Cell Biol.* **2012**, *91*, 107−117.

(10) Saleem, A. N.; Chen, Y.-H.; Baek, H. J.; Hsiao, Y.-W.; Huang, H.-W.; Kao, H.-J.; Liu, K.-M.; Shen, L.-F.; Song, I. w.; Tu, C.-P. D.; Wu, J.-Y.; Kikuchi, T.; Justice, M. J.; Yen, J. J. Y.; Chen, Y.-T. Mice with alopecia, osteoporosis, and systemic amyloidosis due to mutation in zdhhc13, a gene coding for palmitoyl acyltransferase. *PLoS Genet.* **2010**, *6*, e1000985.

(11) Meckler, X.; Roseman, J.; Das, P.; Cheng, H.; Pei, S.; Keat, M.; Kassarjian, B.; Golde, T. E.; Parent, A. T.; Thinakaran, G. Reduced Alzheimer's disease β-Amyloid deposition in transgenic mice expressing S-palmitoylation-deficient APH1aL and nicastrin. *J. Neurosci.* **2010**, *30*, 16160−16169.

(12) Young, F. B.; Butland, S. L.; Sanders, S. S.; Sutton, L. M.; Hayden, M. R. Putting proteins in their place: Palmitoylation in Huntington disease and other neuropsychiatric diseases. *Prog. Neurobiol.* **2012**, *97*, 220−238.

(13) Greaves, J.; Lemonidis, K.; Gorleku, O. A.; Cruchaga, C.; Grefen, C.; Chamberlain, L. H. Palmitoylation-induced Aggregation of Cysteine-string Protein Mutants That Cause Neuronal Ceroid Lipofuscinosis. *J. Biol. Chem.* **2012**, *287*, 37330−37339.

(14) Mukai, J.; Dhilla, A.; Drew, L. J.; Stark, K. L.; Cao, L.; MacDermott, A. B.; Karayiorgou, M.; Gogos, J. A. Palmitoylation-dependent neurodevelopmental deficits in a mouse model of 22q11 microdeletion. *Nat. Neurosci.* **2008**, *11*, 1302−1310.

(15) Raymond, F. L.; Tarpey, P. S.; Edkins, S.; Tofts, C.; O'Meara, S.; Teague, J.; Butler, A.; Stevens, C.; Barthorpe, S.; Buck, G.; Cole, J.; Dicks, E.; Gray, K.; Halliday, K.; Hills, K.; Hinton, J.; Jones, D.; Menzies, A.; Perry, J.; Raine, K.; Shepherd, R.; Small, A.; Varian, J.; Widaa, S.; Mallya, U.; Moon, J.; Luo, Y.; Shaw, M.; Boyle, J.; Kerr, B.; Turner, G.; Quarrell, O.; Cole, T.; Easton, D. F.; Wooster, R.; Bobrow, M.; Schwartz, C. E.; Gecz, J.; Stratton, M. R.; Futreal, P. A. Mutations in ZDHHC9, Which Encodes a Palmitoyltransferase of NRAS and HRAS, Cause X-Linked Mental Retardation Associated with a Marfanoid Habitus. *Am. J. Hum. Genet.* **2007**, *80*, 982−987.

(16) Wei, X.; Schneider, J. G.; Shenouda, S. M.; Lee, A.; Towler, D. A.; Chakravarthy, M. V.; Vita, J. A.; Semenkovich, C. F. De novo lipogenesis maintains vascular homeostasis through endothelial nitric-oxide synthase (eNOS) palmitoylation. *J. Biol. Chem.* **2011**, *286*, 2933−2945.

(17) Yu, L.; Reader, J. C.; Chen, C.; Zhao, X. F.; Ha, J. S.; Lee, C.; York, T.; Gojo, I.; Baer, M. R.; Ning, Y. Activation of a novel palmitoyltransferase ZDHHC14 in acute biphenotypic leukemia and subsets of acute myeloid leukemia. *Leukemia* **2011**, *25*, 367−371.

(18) Anami, K.; Oue, N.; Noguchi, T.; Sakamoto, N.; Sentani, K.; Hayashi, T.; Hinoi, T.; Okajima, M.; Graff, J. M.; Yasui, W. Search for transmembrane protein in gastric cancer by the Escherichia coli ampicillin secretion trap: expression of DSC2 in gastric cancer with intestinal phenotype. *J. Pathol.* **2010**, *221*, 275−284.

(19) Mansilla, F.; Birkenkamp-Demtroder, K.; Kruhoffer, M.; Sorensen, F. B.; Andersen, C. L.; Laiho, P.; Aaltonen, L. A.; Verspaget, H. W.; Orntoft, T. F. Differential expression of DHHC9 in microsatellite stable and instable human colorectal cancer subgroups. *Br. J. Cancer* **2007**, *96*, 1896−1903.

(20) Yamamoto, Y.; Chochi, Y.; Matsuyama, H.; Eguchi, S.; Kawauchi, S.; Furuya, T.; Oga, A.; Kang, J. J.; Naito, K.; Sasaki, K. Gain of 5p15.33 is associated with progression of bladder cancer. *Oncology* **2007**, *72*, 132−138.

(21) Kang, J. U.; Koo, S. H.; Kwon, K. C.; Park, J. W.; Kim, J. M. Gain at chromosomal region 5p15.33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer. *Cancer Genet. Cytogenet.* **2008**, *182*, 1−11.

(22) Ferri, N.; Paoletti, R.; Corsini, A. Lipid-modified proteins as biomarkers for cardiovascular disease: a review. *Biomarkers* **2005**, *10*, 219−237.

(23) Charych, E. I.; Jiang, L.-X.; Lo, F.; Sullivan, K.; Brandon, N. J. Interplay of palmitoylation and phosphorylation in the trafficking and localization of phosphodiesterase 10A: Implications for the treatment of schizophrenia. *J. Neurosci.* **2010**, *30*, 9027−9037.

(24) Yang, G.; Cynader, M. S. Palmitoyl acyltransferase zD17 mediates neuronal responses in acute ischemic brain injury by regulating JNK activation in a signaling module. *J. Neurosci.* **2011**, *31*, 11980−11991.

(25) Munday, A. D.; López, J. A. Posttranslational Protein Palmitoylation: Promoting Platelet Purpose. *Arterioscler., Thromb., Vasc. Biol.* **2007**, *27*, 1496−1499.

(26) Hoffman, M. D.; Kast, J. Mass spectrometric characterization of lipid-modified peptides for the analysis of acylated proteins. *J. Mass Spectrom.* **2006**, *41*, 229−241.

(27) Martin, B. R.; Cravatt, B. F. Large-scale profiling of protein palmitoylation in mammalian cells. *Nat. Methods* **2009**, *6*, 135−138.

(28) Tsai, F. D.; Wynne, J. P.; Ahearn, I. M.; Philips, M. R. Metabolic labeling of Ras with tritiated palmitate to monitor palmitoylation and depalmitoylation. *Methods Mol. Biol.* **2014**, *1120*, 33−41.

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

(29) Zhou, F.; Xue, Y.; Yao, X.; Xu, Y. CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). *Bioinformatics* **2006**, *22*, 894−896.
(30) Xue, Y.; Chen, H.; Jin, C.; Sun, Z.; Yao, X. NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm. *BMC Bioinf.* **2006**, *7*, 458.
(31) Yang, Z. Predicting palmitoylation sites using a regularised biobasis function neural network. In *Bioinformatics Research and Applications*; Măndoiu, I.; Zelikovsky, A., Eds.; Springer: Berlin Heidelberg, 2007; Vol. *4463*, Chapter 37, pp 406−417.
(32) Ren, J.; Wen, L.; Gao, X.; Jin, C.; Xue, Y.; Yao, X. CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng., Des. Sel.* **2008**, *21*, 639−644.
(33) Xue, Y.; Liu, Z.; Cao, J.; Ren, J. Computational prediction of post-translational modification sites in proteins. In *Systems and computational biology - molecular and cellular experimental systems*, Yang, N.-S., Ed.; InTech: Rijeka, Croatia, 2011; Chapter 6, pp 105−124.
(34) Wang, X.-B.; Wu, L.-Y.; Wang, Y.-C.; Deng, N.-Y. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Eng., Des. Sel.* **2009**, *22*, 707−712.
(35) Hu, L.-L.; Wan, S.-B.; Niu, S.; Shi, X.-H.; Li, H.-P.; Cai, Y.-D.; Chou, K.-C. Prediction and analysis of protein palmitoylation sites. *Biochimie* **2011**, *93*, 489−496.
(36) Li, Y. X.; Shao, Y. H.; Deng, N. Y. Improved prediction of palmitoylation sites using PWMs and SVM. *Protein Pept. Lett.* **2011**, *18*, 186−193.
(37) Shi, S. P.; Sun, X. Y.; Qiu, J. D.; Suo, S. B.; Chen, X.; Huang, S. Y.; Liang, R. P. The prediction of palmitoylation site locations using a multiple feature extraction method. *J. Mol. Graphics Modell.* **2013**, *40*, 125−130.
(38) Fu, L.; Xie, H.-L.; Xu, X.-R.; Yang, H.-J.; Nie, X.-D. Combining random forest with multi-amino acid features to identify protein palmitoylation sites. *Chemom. Intell. Lab. Syst.* **2014**, *135*, 208−212.
(39) Kumari, B.; Kumar, R.; Kumar, M. PalmPred: An SVM Based Palmitoylation Prediction Method Using Sequence Profile Information. *PLoS One* **2014**, *9*, e89246.
(40) Pejaver, V.; Hsu, W.-L.; Xin, F.; Dunker, A. K.; Uversky, V. N.; Radivojac, P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.* **2014**, *23*, 1077−1093.
(41) Zoltewicz, S. J.; Lee, S.; Chittoor, V. G.; Freeland, S. M.; Rangaraju, S.; Zacharias, D. A.; Notterpek, L. The palmitoylation state of PMP22 modulates epithelial cell morphology and migration. *ASN Neuro* **2012**, *4*, 409.
(42) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.
(43) Songyang, Z.; Blechner, S.; Hoagland, N.; Hoekstra, M. F.; Piwnica-Worms, H.; Cantley, L. C. Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.* **1994**, *4*, 973−982.
(44) Chen, K.; Kurgan, L. A.; Ruan, J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct. Biol.* **2007**, *7*, 25.
(45) Chen, Y.-Z.; Tang, Y.-R.; Sheng, Z.-Y.; Zhang, Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinf.* **2008**, *9*, 101.
(46) Cid, H.; Bunster, M.; Canales, M.; Gazitua, F. Hydrophobicity and structural classes in proteins. *Protein Eng., Des. Sel.* **1992**, *5*, 373−375.
(47) Bhaskaran, R.; Ponnuswamy, P. K. Positional flexibilities of amino acid residues in globular proteins. *Int. J. Pept. Protein Res.* **1988**, *32*, 241−255.
(48) Charton, M.; Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.* **1982**, *99*, 629−644.
(49) Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1976**, *105*, 1−12.
(50) Pontius, J.; Richelle, J.; Wodak, S. J. Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *J. Mol. Biol.* **1996**, *264*, 121−136.
(51) Fauchère, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278.
(52) Atchley, W. R.; Zhao, J.; Fernandes, A. D.; Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6395−6400.
(53) Rubinstein, N. D.; Mayrose, I.; Pupko, T. A machine-learning approach for predicting B-cell epitopes. *Mol. Immunol.* **2009**, *46*, 840−847.
(54) Huang, T.; Wang, P.; Ye, Z.-Q.; Xu, H.; He, Z.; Feng, K.-Y.; Hu, L.; Cui, W.; Wang, K.; Dong, X.; Xie, L.; Kong, X.; Cai, Y.-D.; Li, Y. Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLoS One* **2010**, *5*, e11900.
(55) Sokal, R. R.; Thomson, B. A. Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population. *Am. J. Phys. Anthropol.* **2006**, *129*, 121−131.
(56) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2006**, *34*, W32−W37.
(57) Chang, W.-C.; Lee, T.-Y.; Shien, D.-M.; Hsu, J. B.-K.; Horng, J.-T.; Hsu, P.-C.; Wang, T.-Y.; Huang, H.-D.; Pan, R.-L. Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J. Comput. Chem.* **2009**, *30*, 2526−2537.
(58) Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J. Comput. Chem.* **2013**, *34*, 2121−2132.
(59) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(60) Petralia, F.; Wang, P.; Yang, J.; Tu, Z. Integrative random forest for gene regulatory network inference. *Bioinformatics* **2015**, *31*, i197−i205.
(61) Chen, C. C.; Schwender, H.; Keith, J.; Nunkesser, R.; Mengersen, K.; Macrossan, P. Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8*, 1580−1591.
(62) Singh, H.; Singh, S.; Singla, D.; Agarwal, S. M.; Raghava, G. P. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct* **2015**, DOI: 10.1186/s13062-015-0046-9.
(63) Gu, W.; Vieira, A. R.; Hoekstra, R. M.; Griffin, P. M.; Cole, D. Use of random forest to estimate population attributable fractions from a case-control study of Salmonella enterica serotype Enteritidis infections. *Epidemiol. Infect.* **2015**, 1−9.
(64) Lin, Z.; Vicente Goncalves, C. M.; Dai, L.; Lu, H. M.; Huang, J. H.; Ji, H.; Wang, D. S.; Yi, L. Z.; Liang, Y. Z. Exploring metabolic syndrome serum profiling based on gas chromatography mass spectrometry and random forest models. *Anal. Chim. Acta* **2014**, *827*, 22−27.
(65) Bylander, T.; Hanzlik, D. Estimating generalization error using out-of-bag estimates. In *National Conference on Artificial Intelligence*, Orlando, FL, July 18−22, 1999; pp 321−327.
(66) Smialowski, P.; Frishman, D.; Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* **2010**, *26*, 440−443.
(67) Li, S.; Iakoucheva, L. M.; Mooney, S. D.; Radivojac, P. Loss of post-translational modification sites in disease. *Biocomputing 2010* **2010**, *15*, 337−347.
(68) O'Dowd, B. F.; Hnatowich, M.; Caron, M. G.; Lefkowitz, R. J.; Bouvier, M. Palmitoylation of the human beta 2-adrenergic receptor. Mutation of Cys341 in the carboxyl tail leads to an uncoupled nonpalmitoylated form of the receptor. *J. Biol. Chem.* **1989**, *264*, 7564−7569.
(69) Adlanmerini, M.; Solinhac, R.; Abot, A.; Fabre, A.; Raymond-Letron, I.; Guihot, A.-L.; Boudou, F.; Sautier, L.; Vessières, E.; Kim, S. H.; Lière, P.; Fontaine, C.; Krust, A.; Chambon, P.; Katzenellenbogen,

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025

J. A.; Gourdy, P.; Shaul, P. W.; Henrion, D.; Arnal, J.-F.; Lenfant, F. Mutation of the palmitoylation site of estrogen receptor α in vivo reveals tissue-specific roles for membrane versus nuclear actions. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E283−E290.
(70) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.
(71) Radivojac, P.; Baenziger, P. H.; Kann, M. G.; Mort, M. E.; Hahn, M. W.; Mooney, S. D. Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* **2008**, *24*, i241−i247.
(72) Rauch, F.; Glorieux, F. H. Osteogenesis imperfecta. *Lancet* **2004**, *363*, 1377−1385.
(73) Foster, J. D.; Vaughan, R. A. Palmitoylation controls dopamine transporter kinetics, degradation, and protein kinase C-dependent regulation. *J. Biol. Chem.* **2011**, *286*, 5175−5186.
(74) Henry, S. C.; Schmidt, E. A.; Fessler, M. B.; Taylor, G. A. Palmitoylation of the Immunity Related GTPase, Irgm1: Impact on Membrane Localization and Ability to Promote Mitochondrial Fission. *PLoS One* **2014**, *9*, e95021.
(75) Benitez, B. A.; Alvarado, D.; Cai, Y.; Mayo, K.; Chakraverty, S.; Norton, J.; Morris, J. C.; Sands, M. S.; Goate, A.; Cruchaga, C. Exome-Sequencing Confirms DNAJC5Mutations as Cause of Adult Neuronal Ceroid-Lipofuscinosis. *PLoS One* **2011**, *6*, e26741.
(76) *Regulation of protein trafficking and function by palmitoylation*, Oxford, U.K., Aug 23−25, 2012.
(77) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316.
(78) Conibear, E.; Davis, N. G. Palmitoylation and depalmitoylation dynamics at a glance. *J. Cell Sci.* **2010**, *123*, 4007−4010.
(79) Vacic, V.; Iakoucheva, L. M.; Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **2006**, *22*, 1536−1537.
(80) Undale, A.; Srinivasan, B.; Drake, M.; McCready, L.; Atkinson, E.; Peterson, J.; Riggs, B. L.; Amin, S.; Moedder, U. I.; Khosla, S. Circulating osteogenic cells: Characterization and relationship to rates of bone loss in postmenopausal women. *Bone* **2010**, *47*, 83−92.

DOI: 10.1021/acs.jcim.5b00276
J. Chem. Inf. Model. 2015, 55, 2015−2025