JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 9, 2018

Pp. 997–1008
DOI: 10.1089/cmb.2017.0069

Prediction of Rare Palmitoylation Events in Proteins

BANDANA KUMARI, RAVINDRA KUMAR, and MANISH KUMAR

ABSTRACT

Palmitoylation directs many cellular processes such as protein trafficking, sorting, signaling, interactions with other biomolecules, to name a few. Palmitoylation commonly occurs on cysteine; however, occasional palmitoylation of few other amino acids has also been reported. To date, comprehensive analysis on occasional palmitoylation is unavailable. In the present study, we reported a computational method to predict palmitoylation of glycine and serine residues in a protein. The method is based on support vector machine (SVM). It was trained on position-specific scoring matrix of amino acids that surrounds palmitoylated glycine and serine. During training, SVM models were evaluated on leave-one-out cross validation, and the maximum prediction accuracies achieved during training were 100% glycine palmitoylation and 99.94% for serine palmitoylation. Similar prediction for performance was also shown on independent data sets. The two SVM models were used to develop a prediction method called *RARE*Palm. We provide web-server and standalone of *RARE*Palm, using the user that can predict the potential glycine and serine palmitoylation site(s) in a protein. Comparative analysis of glycine, serine, and cysteine palmitoylation was also done to analyze pathways and classes to which different forms of palmitoylation belong. We hope that our attempt will be useful in finding more glycine and serine that may undergo palmitoylation and expanding the information on these lesser known sites of palmitoylation.

Keywords: human palmitoylome, palmitoylation, *RARE*Palm, support vector machine.

1. INTRODUCTION

Palmitoylation is an important posttranslational modification (PTM) that generally occurs in secretory and membrane proteins (Smotrys and Linder, 2004; Bannan et al., 2008). Proteins depend on palmitoylation to participate in several biological processes such as regulation of protein sorting (Greaves et al., 2009), membrane/protein interaction (Blaskovic et al., 2013), protein/protein interaction (Noritake et al., 2009), trafficking (Greaves and Chamberlain, 2007), and vesicle fusion (Mitchell et al., 2006; Nadolski and Linder, 2007). Hence, palmitoylated proteins are an integral component of diverse cellular processes ranging from apoptosis to carcinogenesis (Milligan et al., 1995; Frohlich et al., 2014; Yeste-Velasco et al., 2015). At present, three forms of palmitoylation are known to occur in proteins: (1) S-palmitoylation, that is, attachment of a fatty acid (usually palmitic acid, C16:0) to cysteine side chains via thioester linkage; (2) N-palmitoylation, that is, amide linkage between palmitic acid and cysteine or glycine; and (iii) O-palmitoylation, that is, attachment of monounsaturated palmitate (C16:1) to a serine residue via oxyester bond (Kleuss and Krause, 2003; Greaves and Chamberlain, 2006; Zou et al., 2011).

Department of Biophysics, University of Delhi South Campus, New Delhi, India.

Generally, S-palmitoylation that involves cysteine residue is the most common form of palmitoylation. Sometimes, few other residues also undergo palmitoylation and they also have a very significant influence on cellular processes. A thorough search in SwissProt showed palmitoylation of four noncysteine amino acids, namely glycine, serine, lysine, and threonine. Palmitoylation of glycine residues in G-proteins makes cells more receptive for positive stimulus and less sensitive for inhibitory stimuli and thus facilitating regulation of G-protein-mediated signal transduction system (Kleuss and Krause, 2003). On the contrary, palmitoylation of serine has been reported in proteins of ''Wnt/Wg family'' (Willert et al., 2003; Takada et al., 2006), which contribute in the Wnt signaling pathways (Galli and Burrus, 2011), embryogenesis, and carcinogenesis (Reya and Clevers, 2005). The lysine palmitoylation site is important for pore formation in host membrane by *Escherichia coli* (Stanley et al., 1994; Basar et al., 1999) and also for adenylate cyclase toxin of *Bordetella pertussis* (Hackett et al., 1994). O-palmitoyl threonine is observed in bovine lipophilin (Stoffel et al., 1983) and in *Plectreurys tristes* toxin, where it helps the bacteria to act at an intracellular site for membrane penetration (Kabanov et al., 1989).

At present, while several efficient methods are available for prediction of cysteine palmitoylation sites, no predictor is available to detect the noncysteine palmitoylation sites. In this study, we introduce a prediction pipeline, *RARE*Palm, for prediction of noncysteine palmitoylated amino acids, namely glycine and serine. To the best of our knowledge, *RARE*Palm is the first method to identify potential glycine and serine residues that might undergo palmitoylation. *RARE*Palm is based on two models (one each for glycine and serine) built by using sequence conservation feature to train support vector machine (SVM). During training, the performance of glycine-based SVM model was 100% in terms of sensitivity, specificity, and accuracy, whereas in case of serine, 98.28% sensitivity, 100% specificity, and 99.94% accuracy were achieved. The performance and efficiency of models were also validated on independent data sets. For practical applications, we also provided *RARE*Palm in the form of web-server and stand-alone software at http://proteininformatics.org/mkumar/rarepalm/.

Potential glycine and serine palmitoylation sites in human proteome were also predicted with *RARE*Palm. We also cataloged cysteine palmitoylation sites of human proteome by using PalmPred, a predictor developed by our group for cysteine palmitoylation prediction. Finally, we compiled human palmitoylome in which different forms of palmitoylation (involving glycine, serine, and cysteine) were predicted and did a comparative analysis of their classes and pathways.

The overall schema followed during development of prediction module is outlined in Figure 1.

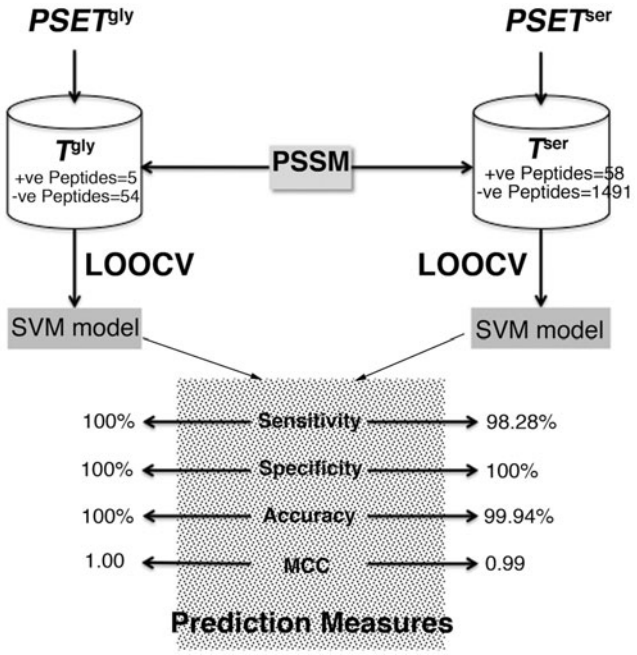


FIG. 1. Flow diagram showing the overall schema of *RARE*Palm.

2. METHODS

2.1. Data sets

2.1.1. Collection and preprocessing of data. A search in UniProt had shown that only glycine, serine, lysine, and threonine besides cysteine undergo palmitoylation. The keywords used during search were *Lipidation* (*FT*) under category ''PTM/Processing'' and *N-palmitoyl glycine*, *O-palmitoyl serine*, *palmitoyl lysine*, and *palmitoyl threonine* under category ''Term.'' Because of a very small number of proteins with lysine (8 proteins) and threonine palmitoylation (5 proteins), we restricted our work to glycine and serine palmitoylation only. We found 18 proteins with glycine palmitoylation and 167 proteins with serine palmitoylation. These two protein sets are, respectively, referred as $PSET^{gly}$ and $PSET^{ser}$ in the article. Each protein of $PSET^{gly}$ and $PSET^{ser}$ contained a single palmitoylated amino acid residue. Only 1 protein in $PSET^{gly}$ and 5 proteins in $PSET^{ser}$ had experimental evidence, the remaining proteins were annotated as probable/potential/by similarity. However, to keep maximum data for our analysis, we used all of them. We noted that in the recent UniProt release, serine palmitoylation in Wnt proteins (oxyester bond) is included under the term ''O-palmitoleyl serine.''

To study cysteine palmitoylation, we took the data set that was earlier used for developing a cysteine palmitoylation prediction method, PalmPred (Kumari et al., 2014). This data set has 151 proteins with 234 experimentally verified palmitoylated cysteines and 1303 nonpalmitoylated cysteines (referred as $PSET^{cys}$ in the article). Unlike $PSET^{gly}$ and $PSET^{ser}$, proteins of $PSET^{cys}$ had multiple occurrences of cysteine palmitoylation sites.

2.1.2. Human proteome data. With the aim to find more proteins having palmitoylation sites and critically analyze their role in cellular metabolism, we also annotated human proteome. The human proteome was retrieved from the Human Protein Reference Database (HPRD), Release 9 (Keshava Prasad et al., 2009) containing a total of 30,046 protein sequences.

2.2. Compilation of positive and negative data set

To develop the prediction method, we built positive and negative data sets by adopting a local window approach. A window of length ''w'' was extracted with a central glycine/serine symmetrically flanked by $(w - 1)/2$ residues. In peptides that did not have sufficient residues to complete the window length, symbol ''X'' was used to complete the length. If the central residue was annotated as palmitoylation site, the peptide was binned into a positive data set, while peptides with nonpalmitoylated central residue constituted the negative data set. In both positive and negative data sets, among multiple identical peptides, only one was kept.

2.3. Training and independent data sets

We randomly chose $\sim 3/4$ th fraction of peptides from both positive and negative data sets to construct the training data sets. The training data sets containing peptides from $PSET^{gly}$ and $PSET^{ser}$ are, respectively, referred as T^{gly} and T^{ser} (Supplementary Tables S1 and S2). Remaining $^1/_4$th data fractions of $PSET^{gly}$ and $PSET^{ser}$ were used as independent data sets (henceforth called I^{gly} and I^{ser}, respectively) (Supplementary Tables S3 and S4).

Since the size of data set used in the present study was small, it was not possible to evaluate the performance of prediction models on a very large set of examples. To manage this limitation we generated two additional data sets, one for glycine and another for serine palmitoylation. These data sets had 50 random sequences. Although these proteins were random sequences, the average amino acid composition of one database was similar to $PSET^{gly}$ and another to $PSET^{ser}$. The average protein composition was calculated by pepstats and random sequences were generated by makeprotseq from the EMBOSS package (Rice et al., 2000). The data set built using $PSET^{gly}$ was termed $RAND^{gly}$ and that built using $PSET^{ser}$ was called $RAND^{ser}$ (Supplementary Tables S5 and S6). $RAND^{gly}$ included 1048 glycines, whereas $RAND^{ser}$ included 1271 serines in total (Supplementary Tables S7 and S8).

2.4. Evolutionary information/position-specific scoring matrix profiles

Evolutionary information in the form of position-specific scoring matrix (PSSM) profiles has been widely used in prediction of PTMs and several other bioinformatic problems (Kumar et al., 2008; Chen

et al., 2012; Wang et al., 2012). The purpose behind using PSSM profile was to extract the occurrence probability of different amino acids at each position. In this work, we generated PSSM profile by using PSI-BLAST program (Altschul et al., 1997). For a protein sequence with N amino acid residues, PSI-BLAST generates a PSSM profile composed of ''*Nx20*'' dimensional vectors. Amino acid at each position (N) in protein has 20 values, representing occurrence probability of all amino acids at that position. A large PSSM profile value indicates an important role of a residue in proteins. The PSSM for each protein of $PSET^{gly}$ and $PSET^{ser}$ was generated against the NCBI NR protein database whose redundancy was reduced to 90%. The PSI-BLAST search was done for three iterations and using an *e*-value cutoff of 0.001 for inclusion of sequences in next iteration of searching. The PSSM values corresponding to each amino acid of the peptide pattern were extracted from their respective protein's PSSM. Any ''X'' in the peptide was given PSSM score 0 at all amino acid positions, that is, 20 times. Thus, the size of each peptide pattern was ''*wx20*,'' where *w* is the window size. These were later used as input pattern during training of SVM.

2.5. Support vector machine

Prediction problem of glycine and serine palmitoylation can be addressed using the binary classification approach, which aims to identify whether a residue is palmitoylated or not. We have used SVM (Vapnik, 1995), a machine-learning algorithm for this. In recent years, SVM has been emerged as a popular tool to differentiate between positive and negative cases of several protein attributes, including palmitoylation (Kumari et al., 2014), glycosylation (Xie et al., 2013), phosphorylation (Lin et al., 2015), and subcellular location (Kumar et al., 2014). In the present work, SVM^{light} software (Jaochims, 1999) was used for SVM implementation.

In addition to SVM, we have also evaluated the performance of Naive Bayes, REPTree, and Random Forest using WEKA suit (Frank et al., 2016).

2.6. Optimization and evaluation of support vector machine model

The effectiveness of a predictor in practical application is often examined by three cross-validation methods: independent data set test, *n*-fold cross validation or subsampling test, and jackknife test or leave-one-out cross validation (LOOCV) (Chou and Zhang, 1995). Of the three, LOOCV is considered the most rigorous (Chou and Zhang, 1995). It evaluates a single data point (test data) using the model generated by remaining sample (training data). The process is repeated till each data point is used as test data. Since the LOOCV approach uses maximum data for training, unlike independent data set and subsampling tests, LOOCV does not suffer from the problem of randomness (Chou, 2011). By taking into account the above discussed features, we adopted the LOOCV approach to examine the prediction quality of SVM models developed from a training data set. We also evaluated the trained model on an independent data set. The prediction performance of each model was measured in terms of sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew's correlation coefficient (MCC).

These measures can be defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100$$

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \times 100$$

$$\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where TP, FN, TN, and FP denote the number of *true positive*, *false negative*, *true negative*, and *false positive*, respectively. In our analysis, all the correctly predicted palmitoylated sites and nonpalmitoylated sites fall under category *true positive* and *true negative*, respectively. If a palmitoylation site was predicted as nonpalmitoylated site, it was considered *false negative* prediction, whereas a nonpalmitoylated site predicted as palmitoylation site was considered *false positive* prediction.

To decide TN, TP, FN, and FP among the predictions, we have used the threshold approach. We have used values between −1 and +1 with step size 0.1 for threshold. If the SVM score for a pattern was

≥threshold, the central residue (Gly/Ser) was predicted as palmitoylated, otherwise nonpalmitoylated. On the basis of actual palmitoylation state of central residue of pattern, the final decision of TP, TN, FP, and FN was made. The thresholds at which we found minimum difference between sensitivity and specificity were considered the optimum thresholds. The evaluation of a trained SVM model using an independent data set was done at optimum SVM threshold.

3. RESULTS AND DISCUSSION

3.1. Performance of support vector machine models during leave-one-out cross validation

Using the PSSM profile of peptides of different lengths (7–27), we have trained SVM in LOOCV mode and evaluated the performance on the basis of evaluation parameters described above. As shown in Table 1, during LOOCV, the SVM model trained on T^{gly} attained the predictive sensitivity, specificity, and accuracy of 100% with both MCC and area under curve (AUC) values being 1.000. The predictive performance of T^{gly}-SVM model remained same throughout window size 7–27 (Supplementary Table S9).

During training of SVM with serine data set, T^{ser}, we observed improvement in performance from window size 7 to 9 (Supplementary Table S9). From window size 9 onward, the value of specificity was maintained at 100%, but the sensitivity gradually increased. This resulted in minor changes in accuracy and MCC. The SVM model for T^{ser} reached its maximum performance at window size 27 with 98.28% sensitivity, 100% specificity, 99.94% accuracy, 0.99 MCC, and 0.995 AUC (Table 1).

3.2. Performance on independent data set

As we can see from the LOOCV results (Supplementary Table S9), the performance of SVM models was more or less similar at all window lengths. Hence to select the best performing model, we evaluated the performance on independent data sets I^{gly} and I^{ser}. By taking into account the performance on independent data sets, we selected the SVM models trained on window size 13 and 27 for glycine and serine palmitoylation prediction, respectively. The selected models were able to predict all palmitoylation sites of independent data sets without any false positive (Table 2 and Supplementary Table S10).

When we evaluated performance efficiency of selected models on random data sets namely $RAND^{gly}$ and $RAND^{ser}$, corresponding SVM models did not predict any glycine and serine palmitoylation site in them (Supplementary Tables S7 and S8). This was expected since both data sets had random sequences. The performances also further reinforce robustness of our prediction model.

3.3. Performance of Naive Bayes, REPTree, and Random Forest

We have also used other classifiers, namely Naive Bayes, REPTree, and Random Forest of WEKA to assess the performance for prediction. To do one to one comparison, these three classifiers were also trained on the data set and features used for SVM. The results of LOOCV analysis of the three models are shown in Table 1. In case of glycine palmitoylation, we observed that Random Forest model had 100% of sensitivity,

Table 1. Performance of Classifiers on Training Data Sets

Classifiers	*Sensitivity (%)*	*Specificity (%)*	*Accuracy (%)*	*MCC*	*AUC*
T^{gly}					
Support vector machine	100.00	100.00	100.00	1.00	1.000
Random Forest	100.00	100.00	100.00	1.00	1.000
Naive Bayes	100.00	98.15	98.31	0.90	1.000
REPTree	80.00	98.15	96.61	0.78	0.931
T^{ser}					
Support vector machine	98.28	100.00	99.94	0.99	0.995
Random Forest	93.10	100.00	99.74	0.96	0.990
Naive Bayes	93.10	100.00	99.74	0.96	0.967
REPTree	77.59	99.26	98.45	0.78	0.974

AUC, area under curve; MCC, Matthew's correlation coefficient.

Table 2. Performance of Classifiers on Independent Data Sets

Classifiers	*Sensitivity (%)*	*Specificity (%)*	*Accuracy (%)*	*MCC*
I^{gly}				
Support vector machine	100.00	100.00	100.00	1.00
Random Forest	100.00	100.00	100.00	1.00
Naive Bayes	100.00	89.47	90.48	0.67
REPTree	100.00	57.89	61.90	0.34
I^{ser}				
Support vector machine	100.00	100.00	100.00	1.00
Random Forest	100.00	100.00	100.00	1.00
Naive Bayes	100.00	93.60	93.86	0.61
REPTree	100.00	99.40	99.42	0.93

specificity, and accuracy. When tested with Naive Bayes, the attained sensitivity, specificity, accuracy, MCC, and AUC were 100.00%, 98.15%, 98.31%, 0.90, and 1.000, respectively. On the contrary, the REPTree model attained an accuracy of 96.61% and an AUC score of 0.931 with 80% sensitivity, 98.15% specificity, and an MCC of 0.78. When we compared the models developed for predicting serine palmitoylation, we found that models of Naive Bayes and Random Forest had similar sensitivity (93.10%), specificity (100%), accuracy (99.74%), and MCC (0.96). REPTree showed comparatively poor performance with 77.59% sensitivity, 99.26% specificity, 98.45% accuracy, 0.78 MCC, and AUC value 0.974.

When we assessed the performance of Random Forest, Naive Bayes, and REPTree models on the independent data set, we found that performance of Naive Bayes was, however, good on training data set, but on independent data set it was comparatively lesser for both the glycine (100% sensitivity, 89.47% specificity, and 90.48% accuracy) and serine palmitoylation (100% sensitivity, 93.60% specificity, 93.86% accuracy) (Table 2). Random Forest models for both glycine and serine palmitoylation were found accurate in terms of all four measures, that is, they had 100% sensitivity, specificity, and accuracy with MCC of 1. REPTree models developed for glycine palmitoylation had a large difference between sensitivity (100%) and specificity (57.89%). Also, the attained accuracy was 61.90%. In I^{ser}, REPTree model attained 100% sensitivity and 99.40% specificity.

This shows the performance of SVM is more consistent than other classifiers, on both training and independent data sets. Hence, for further work, SVM models were chosen as the predictor.

3.4. Web-server

Experimental annotation of any PTM site is both a labor-extensive and resource-consuming exercise. Hence, development of computational tools can help to reduce the search time as well as to save the

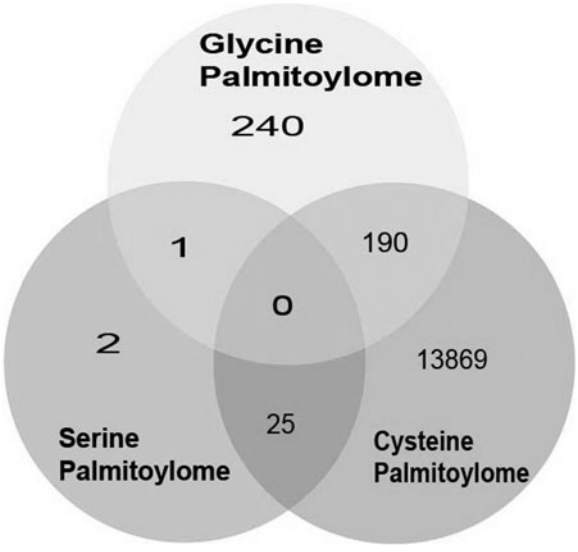


FIG. 2. Number of proteins predicted as part of glycine, serine, and cysteine palmitoylome of human.

resources also. Therefore, we designed a tool called *RARE*Palm by using SVM models trained on T^{gly} and T^{ser}. *RARE*Palm is available both as web-interface and stand-alone software at http://proteininformatics.org/mkumar/rarepalm/. When a user submits the protein sequence(s), the prediction system displays the position of serine and glycine residues and whether it will be palmitoylated or not. *RARE*Palm web-server allows users to submit a maximum of 5 protein sequences at a time. We suggest users to locally install the predictor (tested on LINUX and Mac systems) if the prediction is to be made for >5 sequences. The detailed description of *RARE*Palm tool is available at its web-server.

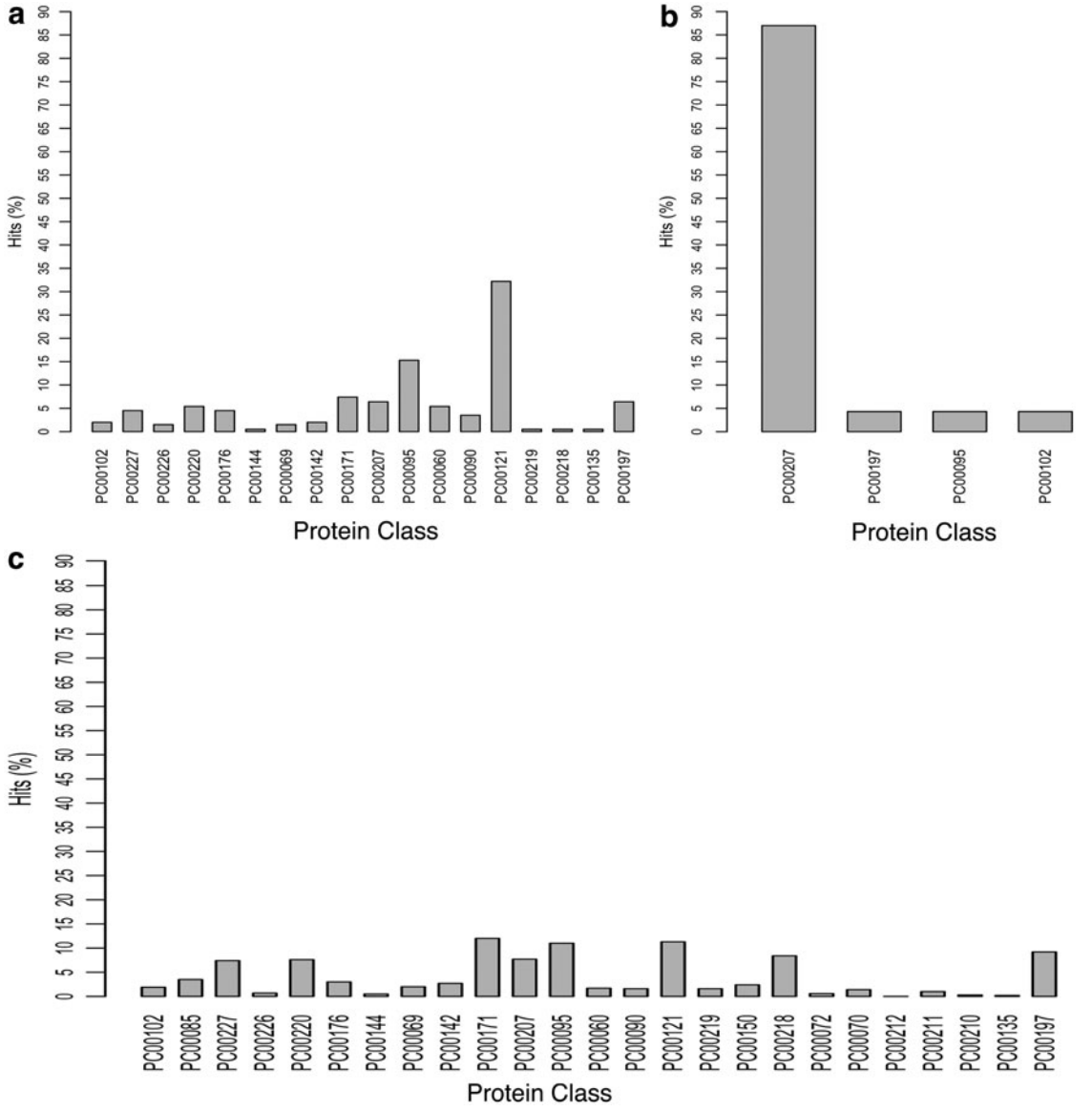


FIG. 3. Functional classes of human proteins predicted to have **(a)** glycine, **(b)** serine, and **(c)** cysteine palmitoylation. Protein class terms shown in figure indicates following categories: PC00060, calcium-binding protein; PC00069, cell adhesion molecule; PC00070, cell junction protein; PC00072, chaperone; PC00085, cytoskeletal protein; PC00090, defense/immunity protein; PC00095, enzyme modulator; PC00102, extracellular matrix protein; PC00121, hydrolase; PC00135, isomerase; PC00142, ligase; PC00144, lyase; PC00150, membrane traffic protein; PC00171, nucleic acid binding; PC00176, oxidoreductase; PC00197, receptor; PC00207, signaling molecule; PC00210, storage protein; PC00211, structural protein; PC00212, surfactant; PC00218, transcription factor; PC00219, transfer/carrier protein; PC00220, transferase; PC00226, transmembrane receptor regulatory/adapter protein; and PC00227, transporter.

3.5. Proteome-wide search of palmitoylation sites in human

Proteins usually undergo PTM to diversify and extend their functions (Wang et al., 2014). Hence, PTMs are considered key regulators of several complex intracellular processes. Therefore, analysis of palmitoylation events at proteome level might provide comprehensive information about their role in the cellular metabolism. To characterize human palmitoylome, we predicted all glycine and serine palmitoylation sites in the human proteome using *RARE*Palm. Human proteome has 1,085,582 glycine and 1,371,606 serine residues in total. *RARE*Palm predicted palmitoylation in 464 glycines present in 431 proteins and 28 serines present in 28 proteins (Fig. 2 and Supplementary Table S11). This trivial number of predicted palmitoylation sites from a large set of glycine and serine residues shows that the predictor did not give random predictions but it is highly specific in nature. To do a comparative analysis of glycine and serine palmitoylation vis-à-vis cysteine palmitoylation, PalmPred (Kumari et al., 2014) was used to predict palmitoylated cysteine in human proteome. Of 370,162 cysteines in human proteome, 30,595 cysteine residues were identified as potential palmitoylation sites (in 14,084 proteins). Thus, we found 431 proteins containing glycine, 28 proteins with serine, and 14,084 proteins with cysteine in human proteome undergoing palmitoylation, and henceforth, we referred them as glycine palmitoylome, serine palmitoylome, and cysteine palmitoylome, respectively.

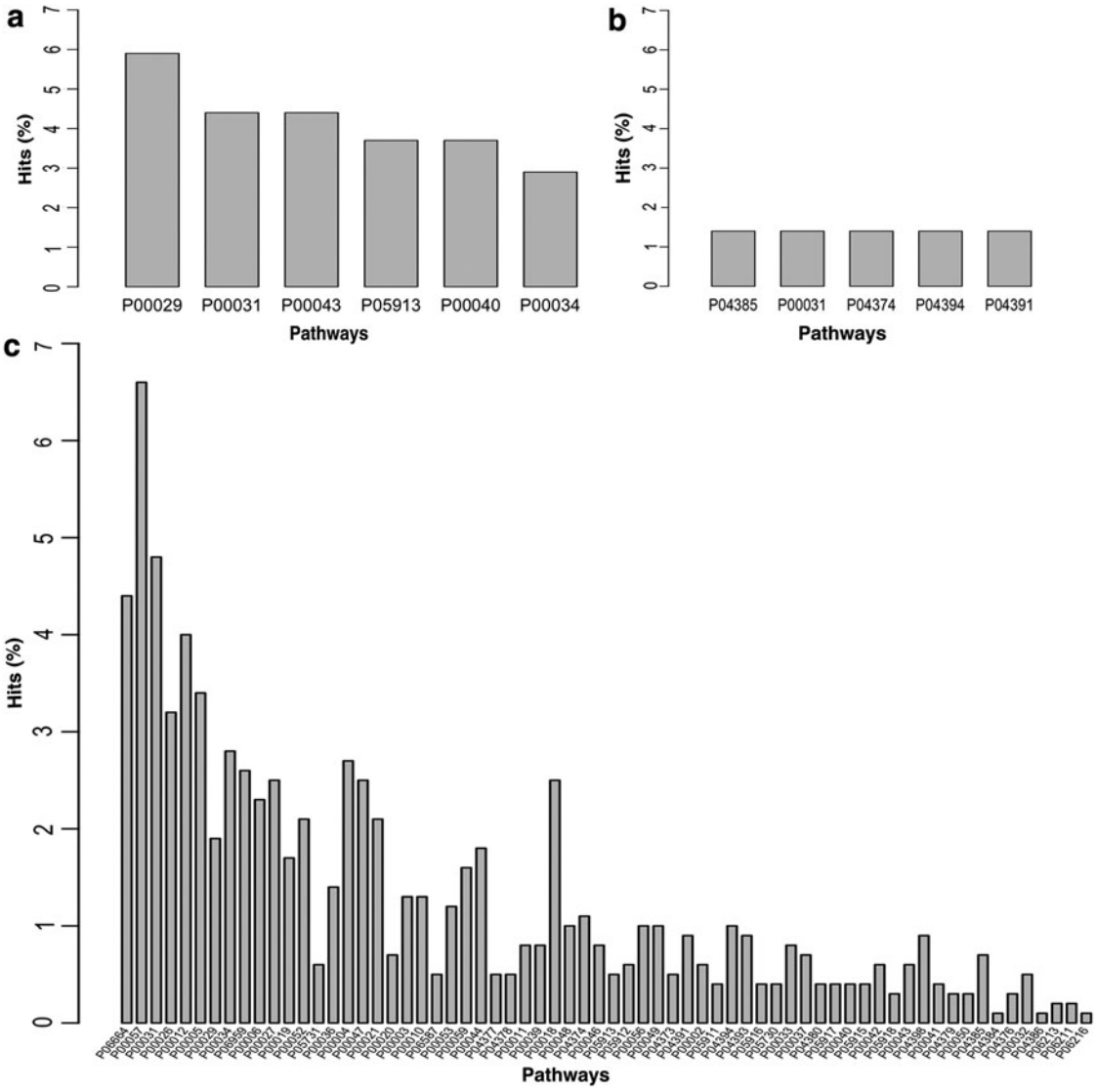
Our analysis also showed that in only one protein (NP_002929.1), both glycine and serine palmitoylation sites were predicted and that too at nearby positions (palmitoylation of glycine at position 155 and palmitoylation of serine at position 151). We also observed the presence of glycine (190 proteins) or serine (25 proteins) palmitoylation sites along with the cysteine (Fig. 2; Supplementary Table S11). Our analysis also revealed that in 30 proteins, palmitoylation sites of glycine and cysteine residue are present at adjacent positions; in 29 proteins, palmitoylation occurred at the N-terminal second and third amino acids (*Nterm*-**AA1–Gly2–Cys3–**), while in 1 protein (NP_002341.1) it occurred at the 23rd and 24th residues (*Nterm*-**AA1-22–Gly23–Cys24–**) (Supplementary Table S11). Among the proteins in which the palmitoylation site was predicted for both cysteine and serine amino acids, in 3 proteins (NP_078613.1, NP_004176.2, and NP_490645.1), palmitoylated cysteine/serine positions were present at a distance of 5 amino acids (237th/243th, 218th/224th, 180th/186th).

FIG. 4. PANTHER pathway analysis in human proteome predicted to contain palmitoylated **(a)** glycine, **(b)** serine, and **(c)** cysteine. On *x*-axis, shown pathway terms are those significant at *p*-value <0.1 after Bonferroni correction: P00002, alpha adrenergic receptor signaling pathway; P00003, Alzheimer's disease-amyloid secretase pathway; P00004, Alzheimer's disease-presenilin pathway; P00005, angiogenesis; P00006, apoptosis signaling pathway; P00010, B cell activation; P00011, blood coagulation; P00012, cadherin signaling pathway; P00018, EGF receptor signaling pathway; P00019, endothelin signaling pathway; P00020, FAS signaling pathway; P00021, FGF signaling pathway; P00026, heterotrimeric G-protein signaling pathway-Gi alpha- and Gs alpha-mediated pathway; P00027, heterotrimeric G-protein signaling pathway-Gq alpha- and Go alpha-mediated pathway; P00029, Huntington disease; P00031, inflammation mediated by chemokine and cytokine signaling pathway; P00032, insulin/IGF pathway-mitogen-activated protein kinase kinase/MAP kinase cascade; P00033, insulin/IGF pathway-protein kinase B signaling cascade; P00034, integrin signaling pathway; P00036, interleukin signaling pathway; P00037, ionotropic glutamate receptor pathway; P00039, metabotropic glutamate receptor group III pathway; P00040, metabotropic glutamate receptor group II pathway; P00041, metabotropic glutamate receptor group I pathway; P00042, muscarinic acetylcholine receptor 1 and 3 signaling pathway; P00043, muscarinic acetylcholine receptor 2 and 4 signaling pathway; P00044, nicotinic acetylcholine receptor signaling pathway; P00046, oxidative stress response; P00047, PDGF signaling pathway; P00048, PI3 kinase pathway; P00049, Parkinson's disease; P00050, plasminogen activating cascade; P00052, TGF-beta signaling pathway; P00053, T cell activation; P00056, VEGF signaling pathway; P00057, Wnt signaling pathway; P00059, p53 pathway; P04373, 5HT1-type receptor-mediated signaling pathway; P04374, 5HT2-type receptor-mediated signaling pathway; P04376, 5HT4-type receptor-mediated signaling pathway; P04377, beta1 adrenergic receptor signaling pathway; P04378, beta2 adrenergic receptor signaling pathway; P04379, beta3 adrenergic receptor signaling pathway; P04380, corticotropin releasing factor receptor signaling pathway; P04384, gamma-aminobutyric acid synthesis; P04385, histamine H1 receptor-mediated signaling pathway; P04386, histamine H2 receptor-mediated signaling pathway; P04391, oxytocin receptor-mediated signaling pathway; P04393, Ras pathway; P04394, thyrotropin-releasing hormone receptor signaling pathway; P04398, p53 pathway feedback loops 2; P05730, endogenous cannabinoid signaling; P05731, GABA-B receptor II signaling; P05911, angiotensin II-stimulated signaling through G proteins and beta-arrestin; P05912, dopamine receptor-mediated signaling pathway; P05913, enkephalin release; P05915, opioid proenkephalin pathway; P05916, opioid prodynorphin pathway; P05917, opioid proopiomelanocortin pathway; P05918, p38 MAPK pathway; P06211, BMP/activin signaling pathway-drosophila; P06213, DPP signaling pathway; P06587, nicotine pharmacodynamic pathway; P06664, gonadotropin-releasing hormone receptor pathway; P06959, CCKR signaling map.

3.6. Annotation of protein functional class and pathways of human palmitoylome

In the present work, PANTHER (version 10.0) was used to assign pathway and class to proteins of three palmitoylome. PANTHER (**P**rotein **An**alysis **Th**rough **E**volutionary **R**elationships) is a very popular tool for large-scale functional analysis of protein set. In PANTHER, using one gene representative per gene, sequences are organized into families of homologous genes. Then, phylogeny is used to identify subfamilies with each family. Experimentally derived gene ontology (GO) functional terms are then assigned across the related gene.

PANTHER identified 18 protein classes in glycine palmitoylome and 4 protein classes in serine palmitoylome. Glycine palmitoylome included mainly enzymes, calcium and nucleic acid binding proteins, receptor, signaling molecule, and transporter (Fig. 3a), whereas most of the human proteins with palmitoylation at serine were identified as signaling molecules (Fig. 3b). Cysteine palmitoylome includes all the classes present in glycine and serine palmitoylation along with few additional classes such as cell junction protein, chaperone, cytoskeletal protein, membrane traffic protein, storage protein, structural protein, and surfactant and transcription factor (Fig. 3c). We found an interesting observation that glycine and cysteine palmitoylome had very similar patterns of protein classes; in fact, for both of them top three of the most highly abundant classes were hydrolase, enzyme modulator, and nucleic acid binding. Other moderately occupying classes in them were calcium-binding protein, transferase, receptor, transporter, and signaling molecule. In both palmitoylome, isomerase, lyase, and transfer/carrier protein were among the least found



classes. In contrast, serine palmitoylome had an entirely different composition of functional classes. It had signaling molecule as the most abundant protein class (87%). Signaling molecule was also observed in genes of glycine and cysteine palmitoylome, but at the lower range (6%–7%).

To find a clear and concise picture of pathways in which the abovementioned protein classes are involved, we also did enrichment analysis of pathways (Bonferroni-$P < 0.1$) in different palmitoylome. We found that genes in glycine palmitoylome were enriched in Huntington disease (P00029), enkephalin release (P05913), metabotropic glutamate receptor group II, and various signaling pathways (P00031, P00034, P00043) (Fig. 4a; Supplementary Table S12). Genes associated with serine palmitoylome were enriched for various signaling pathways sharing an equal proportion (Fig. 4b; Supplementary Table S12). The enriched pathways in cysteine palmitoylome genes were related to different neurodegenerative disorders, signaling, immune systems, and tumors (Fig. 4c; Supplementary Table S12). Some of the other pathways were angiogenesis, blood coagulation, oxidative stress response, and metabotropic glutamate receptor group pathway. We noted that the key pathway of palmitoylated proteins, irrespective of their target residue, is signaling. Similar to protein class, all pathway terms of glycine and serine palmitoylome also appeared in cysteine palmitoylome.

4. CONCLUSIONS

In the present study, a highly accurate predictor for glycine and serine palmitoylation in protein sequences was developed. Using the predictor, we predicted palmitoylation sites of glycine and serine residues in human proteome. Cysteine palmitoylation was also predicted using PalmPred for comparative analysis. Functional annotation of all predicted palmitoylation sites, in terms of functional class and pathways, was also carried out by PANTHER. Results showed that among the proteins in which palmitoylation at both glycine and cysteine was observed, they were present in consecutive amino acids. Proteins of all the three palmitoylation were primarily involved in signaling. Analysis of each palmitoylation class showed that palmitoylated glycine and cysteine residues bearing proteins participate in similar pathways and functional classes. We hope that the developed prediction system will be useful for discovery of new palmitoylation events. It is also anticipated to reveal many unknown facts about this phenomenon.

ACKNOWLEDGMENTS

We are thankful to Dr. Neelja Singhal for her valuable comments in article preparation. B.K. is a recipient of ICMR-SRF (Grant No. BIC/11(33)/2014) and R.K. was supported by UGC-SRF (20-12/2009(ii)EU-IV).

AUTHORS' CONTRIBUTIONS

M.K. conceived, designed, and supervised the research. B.K. and R.K. performed the experiments. M.K. and B.K. performed the analysis and wrote the article. All authors read, discussed, and approved the article.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Bannan, B.A., Van Etten, J., Kohler, J.A., et al. 2008. The Drosophila protein palmitoylome: Characterizing palmitoylthioesterases and DHHC palmitoyl-transferases. *Fly (Austin)* 2, 198–214.

Basar, T., Havlicek, V., Bezouskova, S., et al. 1999. The conserved lysine 860 in the additional fatty-acylation site of *Bordetella pertussis* adenylate cyclase is crucial for toxin function independently of its acylation status. *J. Biol. Chem.* 274, 10777–10783.

Blaskovic, S., Blanc, M., and Van Der Goot, F.G. 2013. What does S-palmitoylation do to membrane proteins? *FEBS J.* 280, 2766–2774.

Chen, Y.Z., Chen, Z., Gong, Y.A., and Ying, G. 2012. SUMOhydro: A novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One* 7, e39195.

Chou, K.C. 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247.

Chou, K.C., and Zhang, C.T. 1995. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 30, 275–349.

Frank, E., Hall, M.A., and Witten, I.H. 2016. *The WEKA Workbench. Online Appendix for ''Data Mining: Practical Machine Learning Tools and Techniques,''* Fourth Edition. Morgan Kaufmann, San Francisco, CA.

Frohlich, M., Dejanovic, B., Kashkar, H., et al. 2014. S-palmitoylation represents a novel mechanism regulating the mitochondrial targeting of BAX and initiation of apoptosis. *Cell Death Dis.* 5, e1057.

Galli, L.M., and Burrus, L.W. 2011. Differential palmit(e)oylation of Wnt1 on C93 and S224 residues has overlapping and distinct consequences. *PLoS One* 6, e26636.

Greaves, J., and Chamberlain, L.H. 2006. Dual role of the cysteine-string domain in membrane binding and palmitoylation-dependent sorting of the molecular chaperone cysteine-string protein. *Mol. Biol. Cell.* 17, 4748–4759.

Greaves, J., and Chamberlain, L.H. 2007. Palmitoylation-dependent protein sorting. *J. Cell Biol.* 176, 249–254.

Greaves, J., Prescott, G.R., Gorleku, O.A., and Chamberlain, L.H. 2009. The fat controller: Roles of palmitoylation in intracellular protein trafficking and targeting to membrane microdomains (review). *Mol. Membr. Biol.* 26, 67–79.

Hackett, M., Guo, L., Shabanowitz, J., et al. 1994. Internal lysine palmitoylation in adenylate cyclase toxin from *Bordetella pertussis*. *Science* 266, 433–435.

Jaochims, T. 1999. Making large-scale SVM learning practical. Scholkopf, B., Burges, C., and Smola, A., eds. *Advances in Kernel Methods–Support Vector Learning.* MIT Press, Cambridge, MA.

Kabanov, A.V., Levashov, A.V., and Alakhov, V. 1989. Lipid modification of proteins and their membrane transport. *Protein Eng.* 3, 39–42.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. 2009. Human Protein Reference Database—2009 update. *Nucleic Acids Res.* 37, D767–D772.

Kleuss, C., and Krause, E. 2003. Galpha(s) is palmitoylated at the N-terminal glycine. *EMBO J.* 22, 826–832.

Kumar, M., Gromiha, M.M., and Raghava, G.P. 2008. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 71, 189–194.

Kumar, R., Jain, S., Kumari, B., and Kumar, M. 2014. Protein sub-nuclear localization prediction using SVM and Pfam domain information. *PLoS One* 9, e98345.

Kumari, B., Kumar, R., and Kumar, M. 2014. PalmPred: An SVM based palmitoylation prediction method using sequence profile information. *PLoS One* 9, e89246.

Lin, S., Song, Q., Tao, H., et al. 2015. Rice_Phospho 1.0: A new rice-specific SVM predictor for protein phosphorylation sites. *Sci. Rep.* 5, 11940.

Milligan, G., Parenti, M., and Magee, A.I. 1995. The dynamic role of palmitoylation in signal transduction. *Trends Biochem. Sci.* 20, 181–187.

Mitchell, D.A., Vasudevan, A., Linder, M.E., and Deschenes, R.J. 2006. Protein palmitoylation by a family of DHHC protein S-acyltransferases. *J. Lipid Res.* 47, 1118–1127.

Nadolski, M.J., and Linder, M.E. 2007. Protein lipidation. *FEBS J.* 274, 5202–5210.

Noritake, J., Fukata, Y., Iwanaga, T., et al. 2009. Mobile DHHC palmitoylating enzyme mediates activity-sensitive synaptic targeting of PSD-95. *J. Cell Biol.* 186, 147–160.

Reya, T., and Clevers, H. 2005. Wnt signalling in stem cells and cancer. *Nature* 434, 843–850.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–277.

Smotrys, J.E., and Linder, M.E. 2004. Palmitoylation of intracellular signaling proteins: Regulation and function. *Annu. Rev. Biochem.* 73, 559–587.

Stanley, P., Packman, L.C., Koronakis, V., and Hughes, C. 1994. Fatty acylation of two internal lysine residues required for the toxic activity of *Escherichia coli* hemolysin. *Science* 266, 1992–1996.

Stoffel, W., Hillen, H., Schroder, W., and Deutzmann, R. 1983. The primary structure of bovine brain myelin lipophilin (proteolipid apoprotein). *Hoppe Seylers Z Physiol. Chem.* 364, 1455–1466.

Takada, R., Satomi, Y., Kurata, T., et al. 2006. Monounsaturated fatty acid modification of Wnt protein: Its role in Wnt secretion. *Dev. Cell.* 11, 791–801.

Vapnik, V. 1995. *The Nature of Statical Learning Theory.* Springer Verlag, New York.

Wang, X., Mi, G., Wang, C., et al. 2012. Prediction of flavin mono-nucleotide binding sites using modified PSSM profile and ensemble support vector machine. *Comput. Biol. Med.* 42, 1053–1059.

Wang, Y.C., Peterson, S.E., and Loring, J.F. 2014. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.* 24, 143–160.

Willert, K., Brown, J.D., Danenberg, E., et al. 2003. Wnt proteins are lipid-modified and can act as stem cell growth factors. *Nature* 423, 448–452.

Xie, H.L., Fu, L., and Nie, X.D. 2013. Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. *Protein Eng. Des. Sel.* 26, 735–742.

Yeste-Velasco, M., Linder, M.E., and Lu, Y.J. 2015. Protein S-palmitoylation and cancer. *Biochim. Biophys. Acta.* 1856, 107–120.

Zou, C., Ellis, B.M., Smith, R.M., et al. 2011. Acyl-CoA:Lysophosphatidylcholine acyltransferase I (Lpcat1) catalyzes histone protein O-palmitoylation to regulate mRNA synthesis. *J. Biol. Chem.* 286, 28019–28025.

Address correspondence to:
Dr. Manish Kumar
Department of Biophysics
University of Delhi South Campus
Benito Juarez Road
New Delhi 110021
India

E-mail: manish@south.du.ac.in