Bioinformatics, 35(17), 2019, 2887–2890
doi: 10.1093/bioinformatics/btz003
Advance Access Publication Date: 12 January 2019
Discovery Note

Genome analysis

Identifying residues that determine palmitoylation using association rule mining

Bandana Kumari, Ravindra Kumar and Manish Kumar*

Department of Biophysics, University of Delhi South Campus, New Delhi 110021, India

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 30, 2018; revised on September 15, 2018; editorial decision on December 31, 2018; accepted on January 7, 2019

Abstract

Motivation: In eukaryotes, palmitoylation drives several essential cellular mechanisms like protein sorting, protein stability and protein–protein interaction. Several amino acids namely Cys, Gly, Ser, Thr and Lys undergo palmitoylation. But very little is known about the amino acid patterns that promote palmitoylation.

Results: We deduced presence of statistically significant amino acids around palmitoylation sites and their association with different palmitoylated residues i.e. Cys, Gly and Ser. The results suggest that palmitoylation, irrespective of its target residue, generally occurs at sites where Cys, Leu, Lys, Arg, Ser and Met are abundant. Furthermore, functional properties of the three types of palmitoylated proteins were compared. We observed similar functional behavior of Cys and Gly palmitoylated proteins but proteins with Ser palmitoylation showed distinctiveness from remaining two. Motif-wise functional conservation was also observed in Cys palmitoylated proteins. We also did functional annotation of predicted human palmitoylome.

Contact: manish@south.du.ac.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Palmitoylation is a common post-translational modification (PTM) in eukaryotes. It mostly occurs on Cys but sometimes other amino acids are also involved (detail about palmitoylation and their importance is in Supplementary File, Introduction). As physico-chemical attribute of flanking amino acid pattern(s) leads to PTM, hence identification of palmitoylation promoting amino acid patterns in vicinity of palmitoylation may help to find more such sites and collect finer details about the process. In last few years, many computational methods have been developed to identify various PTM sites in proteins (Jia *et al.*, 2014, 2016; Liu *et al.*, 2017; Qiu *et al.*, 2014, 2015, 2016a,b, 2017; Xu *et al.*, 2013a,b, 2014a,b, 2017; Zhang *et al.*, 2014). In past, attempts to discover motifs at palmitoylation sites were also made but they were targeted specifically for S-palmitoylation. Hence a few motifs are available to locate S-palmitoylation sites but they do not occur in all palmitoylated proteins (Gubitosi-Klug *et al.*, 2005). For non-Cys palmitoylation, very less information is available.

In this paper we report survey of statistically preferred amino acids in vicinity of palmitoylated amino acids and their association with the event of palmitoylation using association rule mining. This was done for both common (Cys) and rare palmitoylation (Gly and Ser). Obtained amino acid patterns were also verified by motif finding tool. We also investigated the probable functions of proteins involved in each category of palmitoylation. Furthermore, preferred amino acids near palmitoylation sites and functional involvement of palmitoylated proteins were also analyzed in human proteome.

2 Materials and methods

Dataset of proteins having minimum one palmitoylated amino acid was obtained from SwissProt. In this work, we determined (i) statistically preferred amino acids and their association with palmitoylation by MAPRes (Ahmad *et al.*, 2008), (ii) probable motifs for three types of palmitoylation using MEME (Bailey and Elkan, 1994), (iii) functional enrichment of palmitoylated proteins by DAVID Huang da *et al.*, 2009 and (iv) functions of human palmitoylome from PANTHER (Mi *et al.*, 2013).

Detailed methodology is described in Supplementary File, Methods.

3 Results

3.1 Identification of statistically significant amino acids and association pattern mining

We calculated the statistically preferred amino acids in flanking regions of palmitoylated Gly, Ser and Cys and their association with the occurrence of palmitoylation. The length of peptides used for analysis was 31 amino acids. The positions marked −15 to +15 were flanking the palmitoylated residues at center (0th position) (Supplementary File, Methods). In case of Gly-palmitoylation, disorder/order neutral amino acid (Met) was preferred at position –1 (Supplementary File, Table S1). Positions +3, +5, +6, +8, +10, +11, +13, +14 and +15 had disorder-promoting amino acids, whereas at four positions (+1, +2, +4 and +12) order-promoting residues were found. Around palmitoylated Ser, disorder-promoting residues were preferred at 21 positions out of total 30 flanking positions (Supplementary File, Table S1). In palmitoylated Cys we noted enrichment of order-promoting residues at positions –12, –10, –7 and –5 to +3. But except +1, disorder-promoting residues were preferred at most downstream positions (+2 to +6, +8, +10 and +11) (Supplementary File, Table S1).

After finding statistically preferred amino acids, we derived the association rule at different minimum support levels (5, 10, 15,... 100%). For palmitoylated Gly, association pattern {<Met, –1><Cys, 1><Leu, 2><Gly, 3><Asn, 4><Ser, 5><Lys, 6><Thr, 7><Glu, 8><Asp, 9><Gln, 10><Arg, 11><Asn, 12><Glu, 13><Glu, 14><Lys, 15>=>Gly} was present from support levels 5% to 50% (Supplementary File, Table S2). It gradually tapered off from support level 55% and contained only two residues at support level 95 and 100%. In case of palmitoylated Ser, we could find association patterns in range of support levels 30–95%. There were 28 association patterns in total constituting of 21 unique patterns (Supplementary File, Table S3). For palmitoylated Cys association was found only up to 20% support level. At support level 5%, the mined association involved only two amino acid residues {<Leu,–3><Ser, 11>=>Cys} whereas from support levels 10–20%, multiple types of associations, each containing a single amino acid, were found (Supplementary File, Table S2). This suggests that Cys palmitoylation do not follow any particular amino acids pattern. We also observed that the number of amino acids in association patterns decrease with increase in support level. It was expected because support indicates the proportion of data that is covered by the association rule. We also confirmed the generality of the derived patterns by using N-fold cross-validation (Supplementary File, Results, Tables S4 and S5). An example of the amino acid association pattern at palmitoylation site and their location in a protein structure is shown in Figure 1.

3.2 Specificity and conservation analysis

The discovered association rules were evaluated on the basis of composition profiling of palmitoylated peptides/patterns using Two Sample Logo (Supplementary File, Methods). The inferences drawn from results of association mining and statistically preferred amino acids were also reconfirmed by their respective Two Sample Logos (Supplementary File, Results and Fig. S1). Our findings indicated that Cys, Leu, Lys, Arg, Ser and Met are the most frequently occurring residues near all three palmitoylation sites. This also indicates their important role in palmitoylation.

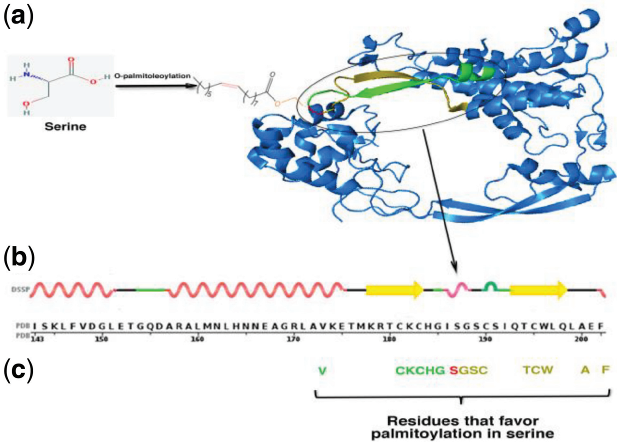
(a)
Serine
O-palmitoleoylation
(b)
DSSP
PDB I S K L F V D G L E T G Q D A R A L M N L H N N E A G R L A V K E T M K R T C K C H G I S G S C S I Q T C W L Q L A E F
PDB 143 150 160 170 180 190 200
(c)
V CKCHG SGSC TCW A F
Residues that favor palmitoylation in serine

Fig. 1. An example of placement of amino acids in association rule and its cognate palmitoylation in a protein 3D-structure. (**a**) O-palmitoleoylation of serine in PDB protein 4F0A. (**b**) Amino acid and secondary structure of protein in neighbouring regions of palmitoylated serine (position 187). (**c**) Amino acids which were found associated to the palmitoylated serine. Red indicates palmitoylated serine; green and brown indicates upstream and downstream amino acids respectively (Color version of this figure is available at *Bioinformatics* online.)

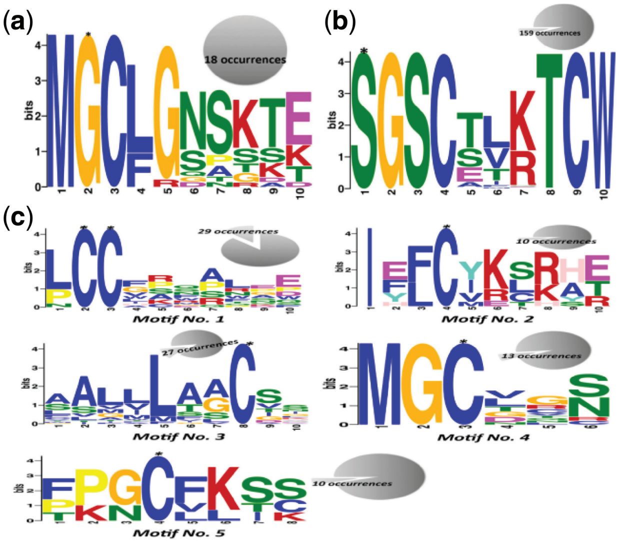

Fig. 2. Motifs found in (**a**) palmitoylated glycine, (**b**) palmitoylated serine and (**c**) palmitoylated cysteine. Palmitoylated residue in motif is marked by *. The number in pie chart of each motif shows number of times they appeared in the palmitoylated proteins

3.3 Glycine, serine and cysteine palmitoylation motif

A few Cys-palmitoylation promoting motifs have already been reported earlier. For example, in Src family of tyrosine kinases and α-subunits of trimeric G-proteins, if Gly at position 2 is myristoylated then Cys at position 3 will undergo palmitoylation (Aicart-Ramos *et al.*, 2011). In H-Ras and R-Ras proteins, Cys of –*CaaX* motif (*C* = cysteine, *a* = any aliphatic amino acid and *X* = amino acid specifying farnesylation or geranylation) undergoes palmitoylation (Hancock *et al.*, 1989). But these motifs represent a tiny fraction of Cys palmitoylation. Hence we searched for all possible palmitoylation motifs using MEME (version 4.11.1) (*E-value* < 0.05). In GLY+, we found a single pattern 'MGC[LF]G[NS]SKT[EK]' (Fig. 2a). A similar pattern $_N$-MGCLGNSKTE in Guanine nucleotide-binding protein G(s) subunit alpha is previously described as palmitoylation site (Kleuss and Krause, 2003; Mumby *et al.*, 1994) (Supplementary File, Table S6).

For peptides containing palmitoylated Ser, the identified motif occupied 159 sites out of 167 SER+ (Fig. 2b and Supplementary File, Table S6). PrositeScan associated this motif with Wnt protein family. MEME discovered 5 different motifs in 234 Cys+ peptides. These motifs were present at a minimum of 10 sites; we excluded motifs found at <10 sites due to very less number (Fig. 2c and Supplementary File, Table S6). All except one (*Motif No. 5*) have not been reported earlier, hence these might be considered as novel motifs for Cys palmitoylation.

3.4 GO annotations of palmitoylation dataset

To analyze functions that might be modulated by palmitoylation, we performed gene ontology (GO) term enrichment analysis of proteins containing Gly, Ser and Cys palmitoylation site using DAVID (Supplementary File, Methods). The top ten statistically enriched GO terms in all three functional categories i.e. cellular components, molecular functions and biological processes were used for further analysis (Supplementary File, Tables S7–S9 respectively). Enrichment analysis (P-value < 0.05) had shown that most proteins with Gly and Cys palmitoylation had cellular description of plasma membrane (GO: 0005886) (Fig. 3 and Supplementary File, Table S7). In case of proteins containing Ser as palmitoylation site, cellular component of 94.7% of protein was found as extracellular matrix and extracellular region (Fig. 3 and Supplementary File, Table S7).

In molecular functions category also, similar enrichment of most of GO terms were found in palmitoylated Gly and Cys proteins. Surprisingly, only <10% of palmitoylated Ser containing proteins were enriched in any molecular functions. These functions were mainly structural molecule activity and extracellular matrix structural constituent (Fig. 3 and Supplementary File, Table S8).

The key biological processes were signaling for all the three types of palmitoylation (Fig. 3 and Supplementary File, Table S9).

3.5 Motif-wise function of Cys-palmitoylated proteins

The results of functional enrichment had shown involvement of majority of Cys palmitoylated proteins in a few specific functions. Hence we thought that proteins having palmitoylated Cys might have a specific Cys palmitoylation motif for a specific location to avoid non-specific signaling. So, we analyzed association of Cys-palmitoylated protein functions vis-à-vis motif with the objective to get insight in distribution of Cys-palmitoylated proteins across (i) cellular functions and (ii) organelles. To do this, Cys-palmitoylated proteins were clustered on the basis of motifs present in them. On analysis we observed cluster-wise distinct pattern of statistically over-represented amino acids. The most interesting pattern of conservation was observed for Cys. It was found over-represented at multiple positions in proteins, which does not have any consensus Cys palmitoylation motif. But in all other proteins (except that had *Motif No. 1*) no over-representation of Cys was observed (Supplementary File, Table S10). Another interesting observation was occurrence pattern of Leu, which was found in all motifs except *Motif No. 4* and *5*.

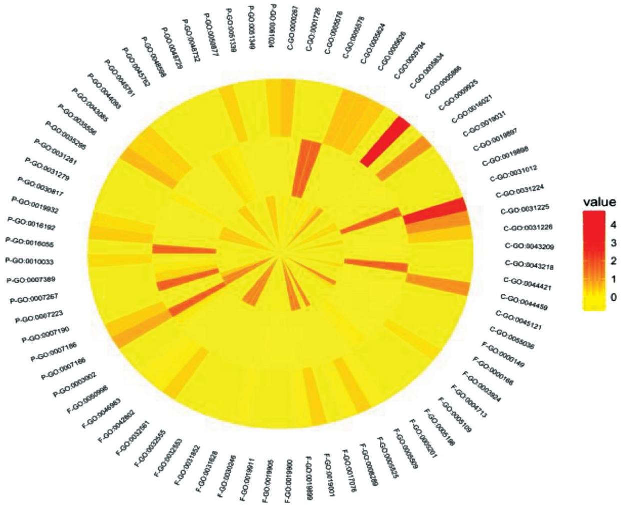

Fig. 3. GO-term enrichment of palmitoylated glycine, serine and cysteine proteins. The graph has three concentric circles; innermost circle is displaying enriched GO terms for glycine palmitoylation, central circle is displaying serine palmitoylation and outermost is showing cysteine palmitoylation. Annotations associated with GO terms are in Supplementary File (Supplementary Table S7–S9). Color scale indicates low (yellow) to high (red) percentage of proteins in each GO term with values ranging from 0 to 100% (Color version of this figure is available at *Bioinformatics* online.)

When we did cluster-wise association mining, patterns were obtained at confidence level 100% (Supplementary File, Table S11). In *Motif No. 1*, association was found up to support level 45% and all association patterns had Cys, Leu, Ser or Glu. Unlike *Motif No. 1*, in which maximum number of amino acids in an association pattern was 4, *Motif No. 2* had 8 amino acids and over-representation of Phe, Ile, Lys and Arg, which were not observed in other motifs. In *Motif No. 3*, the association patterns consisted of Ala, Gly, Leu and Ser were found at support level 15 and 25%. From support level 30–50%, association was restricted to Ala, Leu and Ser, at 55% it was Ala and Leu and afterwards only Leu was associated to palmitoylated Cys in *Motif No. 3*. For *Motif No. 4*, we found association of only Met and Gly {<Met,-2><Gly,-1>=>Cys} from minimum support level 5% to 90%. Interestingly in *Motif No. 5*, we did not find any association pattern. Association pattern for palmitoylated proteins, which did not have any motifs, were mainly consisted of Cys, Leu, Lys, Phe and Gly. This was in line with the association rule mined for complete dataset of palmitoylated Cys (Supplementary File, Table S11).

Motif-wise functional enrichment of GO terms revealed a clear functional difference among them. Cellular components enrichment analysis showed plasma membrane (GO: 0005886) was enriched across proteins of all motifs (Supplementary File, Table S12). Except this, each motif had enrichment of a distinct set of GO-terms. Unique enriched terms for proteins having *Motif No. 1* were Golgi apparatus and photoreceptor outer segment membrane, for *Motif No. 2*, AMPA glutamate receptor complex and postsynaptic density, for *Motif No. 3* cell outer membrane, for *Motif No. 4*, cytosol and heterotrimeric G-protein complex and for *Motif No. 5* focal adhesion (Supplementary File, Table S12). Proteins with *Motif No. 4* and *5* were also enriched in membrane raft (GO: 0045121).

Among molecular functions, protein binding (GO: 0005515) was common enriched term across proteins of Cys palmitoylation motifs. Other terms in *Motif No. 1* were peptide hormone binding, endothelin receptor activity, vasopressin receptor activity and protein kinase A binding (Supplementary File, Table S13). Receptor activity and glutamate receptor activity related terms were present in *Motif No. 2* (Supplementary File, Table S13). Proteins of *Motif No. 3* were shown to have identical protein binding, peptidoglycan binding and lipid binding. In *Motif No. 4* metal ion binding, G-protein beta/gamma-subunit complex binding, signal transducer activity and GTPase activity were enriched. We observed similar molecular functions in *Motif No. 4* and *5*, which were related to GTPase activity and GTP binding.

Biological process terms showed that proteins having *Motif No. 1* was involved in cell proliferation, enteric smooth muscle cell differentiation, posterior midgut development and endothelin receptor signaling pathway (Supplementary File, Table S14). It also has a role in negative regulation of neuron maturation. Proteins with *Motif No. 2* were involved in transport, receptor recycling and synaptic transmission. No enriched biological terms were observed for proteins with *Motif No. 3*. In *Motif No. 4*, response to drug was identified as the most enriched biological process. It was probably because of its other processes that were also related to G-protein and adenylate cyclase (Supplementary File, Table S14). We found proteins grouped under *Motif No. 5* were involved in functions related to hormone, forebrain differentiation, Rac protein and senescence.

The difference of association patterns and functional enrichment among motifs clearly indicates that a distinct amino acid patterns are responsible for the distinct protein function in proteins having Cys palmitoylation motif.

3.6 Annotation of human palmitoylome

We also cross-verified the identified pattern of this study in the human proteome. For this, we used human palmitoylome from our previous work (Kumari *et al.*, 2018) (Supplementary File, Results). The human palmitoylome was predicted by our previously developed palmitoylated Cys prediction tools PalmPred (Kumari *et al.*, 2014) and Gly and Ser prediction tool *RARE*Palm (Kumari *et al.*, 2018). Our analysis revealed presence of similar amino acids and association pattern in predicted palmitoylation also (Supplementary File, Fig. S2). GO terms associated with human palmitoylome (Supplementary File, Table S15–S17) were also similar to the functional enrichment result of proteins in palmitoylated dataset (Supplementary File, Table S7–S9).

4 Conclusions

Palmitoylation is an important PTM of proteins. Most palmitoylation occurs at Cys but sometimes other amino acids also get palmitoylated. In the present study, we discovered specific amino acid sequence patterns to locate Cys, Gly and Ser palmitoylation. With use of Association Rule Mining approach, we found amino acid patterns that favor palmitoylation. We also established degree of correlation between palmitoylation and each mined pattern in terms of confidence and minimum support level. Overall our results suggested that conserved pattern of a few specific amino acids are associated with Ser and Gly palmitoylation, whereas no conserved amino acid pattern was associated with Cys palmitoylation. We also found probable consensus motif for identifying Gly and Ser palmitoylation and five motifs for Cys palmitoylation. In different motifs identified near Cys palmitoylation sites, we found motif-wise specificity for association pattern and function. We also observed that all the protein functions involved in Gly palmitoylation were part of Cys palmitoylation but *vice versa* was not true. This functional behavior was mimicked by the human palmitoylome also. This indicates that the present analysis can be extrapolated to proteome level.

Acknowledgements

The author's wish to acknowledge Prof. Amit Bardhan, Faculty of Management Sciences, University of Delhi, for suggestions offered during the work.

Funding

B.K. was supported under Senior Research Fellowship of Indian Council of Medical Research, India [BIC/11(33)/2014]. R. K. was supported under CSIR/UGC NET Fellowship [20-12/2009(ii)EU-IV].

Conflict of Interest: none declared.

References

Ahmad,I. *et al.* (2008) MAPRes: mining association patterns among preferred amino acid residues in the vicinity of amino acids targeted for post-translational modifications. *Proteomics*, **8**, 1954–1958.

Aicart-Ramos,C. *et al.* (2011) Protein palmitoylation and subcellular trafficking. *Biochim. Biophys. Acta*, **1808**, 2981–2994.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Gubitosi-Klug,R.A. *et al.* (2005) The human Kv1.1 channel is palmitoylated, modulating voltage sensing: identification of a palmitoylation consensus sequence. *Proc. Natl. Acad. Sci. USA,* **102**, 5964–5968.

Hancock,J.F. *et al.* (1989) All ras proteins are polyisoprenylated but only some are palmitoylated. *Cell*, **57**, 1167–1177.

Huang da,W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Jia,C. *et al.* (2014) Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. *Int. J. Mol. Sci.*, **15**, 10410–10423.

Jia,J. *et al.* (2016) pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics*, **32**, 3133–3141.

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

Kumari,B. *et al.* (2014) PalmPred: an SVM based palmitoylation prediction method using sequence profile information. *PLoS One*, **9**, e89246.

Kumari,B. *et al.* (2018) Prediction of rare palmitoylation events in proteins. *J. Comput. Biol.*, **25**, 997–1008.

Liu,L.M. *et al.* (2017) iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. *Med. Chem.*, **13**, 552–559.

Mi,H. *et al.* (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.

Mumby,S.M. *et al.* (1994) Receptor regulation of G-protein palmitoylation. *Proc. Natl. Acad. Sci. USA,* **91**, 2800–2804.

Qiu,W.R. *et al.* (2014) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed. Res. Int.*, **2014**, 947416

Qiu,W.R. *et al.* (2015) iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J. Biomol. Struct. Dyn.*, **33**, 1731–1742.

Qiu,W.R. *et al.* (2016a) iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget*, **7**, 44310–44321.

Qiu,W.R. *et al.* (2016b) iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, **32**, 3116–3123.

Qiu,W.R. *et al.* (2017) iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general pseaac via grey system theory. *Mol. Inform.*, **36**, 1600010.

Xu,Y. *et al.* (2013a) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One*, **8**, e55844.

Xu,Y. *et al.* (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ*, **1**, e171.

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

Xu,Y. *et al.* (2014b) iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One*, **9**, e105018.

Xu,Y. *et al.* (2017) iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med. Chem.*, **13**, 544–551.

Zhang,J. *et al.* (2014) PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. *Int. J. Mol. Sci.*, **15**, 11204–11219.