International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Identification of Protein S-Farnesyl Cysteine Prenylation Sites Based on Substrate Specificities

Van-Nui Nguyen[1], Hai-Minh Nguyen[2]

[1,2]University of Information and Communication Technology (ICTU), Quyet Thang, Thai Nguyen, Vietnam

Abstract: ***Protein S-farnesyl cysteine prenylation (SFCP) is a specific kind of prenylation involved in the transfer of a farnesyl moiety to a cytoplasmic cysteine at or near the C-terminus of the target protein. It has been exhibited to play very important roles in promoting membrane interactions and biological activities of variety of cellular proteins. With the advancements in proteomic technology recently, the number of experimentally verified SFCP sites is increasing and becomes available. Due to the very important roles caused by S-farnesyl cysteine prenylation, the knowledge insight SFCP is one of the most hot issue nowaday. Howver, the number of proposed models for the identification of SFCP sites has still not met our current demands. Therefore, in this study we are motivated to propose a novel schema for the identification of S-farnesyl cysteine prenylation sites based on substrate specificities.***

Keywords: Protein prenylation, protein S-farnesyl cysteine prenylation, support vector machine, substrate motif, maximal dependence decomposition

1. Introduction

Protein prenylation (also known as isoprenylation or lipidation), which is first discovered in fungi in 1978 [1], is the addition of hydrophobic molecules to a protein or chemical compound. Protein prenylation assumes that prenyl groups (3-methyl-but-2-en-1-yl) facilitate attachment to cell membranes, similar to lipid anchors like GPI anchor, though direct evidence is missing. In eukaryote, protein prenylation is a PTM (Post-Translational modification) critical for promoting membrane interactions and biological activities of variety of cellular proteins. It is mediated by protein farnesyltransferase (PFT) by recognizing 'CAAX' motif on protein substrate [4]. The process of protein prenylation is facilitated by three eukaryotic enzymes with partially overlapping substrate specificities: farnesyl transferase, Caax protease and geranylgeranyl transferase [5]. Protein S-farnesyl cysteine prenylation involves the transfer of a farnesyl moiety to a cytoplasmic cysteine at or near the C-terminus of the target protein. Farnesyltransferase (FT) recognizes the so-called C-terminal CaaX box of substrate proteins to attach a farnesyl (15 carbons) anchor to the conserved cysteine via a thioether linkage [6].

Due to the very important role caused by protein S-farnesyl cysteine prenylation (SFCP), the amount of interests in the characterization of S-farnesyl cysteine prenylation has been increasing rapidly recently [5, 7-12]. Specifically, several predictors have been designed for the identification of S-farnesyl cysteine prenylation sites in recent years [6, 13, 14]. Also, these predictors have demonstrated their ability in the characterization of SFCP sites, however, at the moment, there is a lack of computation models or tools for identification of protein S-farnesyl cysteine prenylation sites. Furthermore, as more and more experimentally verified S-farnesyl cysteine prenylation sites become available, the lack of model for identification of S-farnesyl cysteine is serious.

Continuing with previous works [16-19], we are motivated to propose a novel scheme for the identification of protein S-farnesyl cysteine prenylation sites based on substrate specificities. Various featureas, that are extracted and encoded based on the substrate specificities, have been investigated in the work. The SVM-based model, that is constructed based on hybrid feature "AAC+AAPC+PSSM", appears to be the best with an accuracy of 94.14% and MCC of 0.850 when evaluated by five-fold cross-validation.

2. Materials and Methods

Figure 1 displays the system flow of this work, including of four main parts: data collection and pre-processing, feature extraction and encoding, model learning and Independent testing.

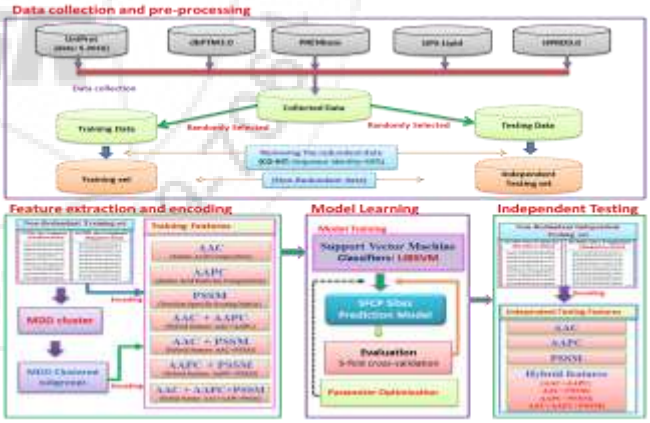


Figure 1: The system flowchart of the work

2.1. Data collection and pre-processing

Experimentally verified S-farnesyl cysteine prenylation (SFCP) sites are collected from open resources and published literatures, including 711 proteins from UniProt/Swiss-Prot [20] (date: May, 2016), 117 proteins from dbPTM3.0 [21], 113 proteins from PRENbase [6], 97 proteins from GPS-Lipid [13], and 27 proteins from HPRD9.0 [22]. Details of these datasets are displayed in Table ***1***. After some technical steps to remove duplicate or redundant proteins, we obtained the final non-redundant dataset containing 670 unique proteins with 718 SFCP sites (positive data). To prepare for independence testing, we

randomly select 70 proteins from the non-redundant dataset to serve as independent testing dataset. The remaining data is considered as training dataset. As a result, in this work, our final training dataset contains 600 unique proteins, and the final independent testing dataset contains 70 unique proteins.

Table 1: Data statistics of experimentally verified SFCP sites collected for the work

Resources	No. of S-Farnesylated proteins	No. of SFCP sites	No. of non-SFCP sites
UniProt_5.2016	711	735	-
dbPTM	117	169	-
PRENbase	113	113	-
GPS-Lipid	97	106	-
HPRD 9.0	27	39	-
Total	1065	1162	-
Combined non-redundant dataset	670	718	-
Training dataset	600	634	5808
Independent testing dataset	70	84	954

In this work, we focus on the sequence-based characterization of SFCP sites with substrate motifs. Therefore, window length of 2n + 1 is adopted to extract sequence fragments centering at the experimentally verified S-farnesyl cysteine (C) residue as well as containing n upstream and n downstream flanking amino acids. The obtained data is served as postitive data. To extract negative data, the sequence fragments containing window length of 2n + 1 amino acids and centering at lysine residue without the annotation of S-farnesyl cysteine prenylation residue were regarded as the negative training data (non-SFCP sites). According to a previous work [16, 17] and our preliminary evaluation by using various window lengths, the window size of 13 (n=6) has been shown to provide the optimal accuracy in the identification of SFCP sites. As a result, the training dataset contains 634 positive and 5808 negative data; the testing dataset consists of 84 positive and 954 negative data. Due to the fact that some data in the training dataset and testing dataset could be overlapped, so, the performance of the predictive model may be overestimated. Therefore, in order to avoid the overestimation of the model, the CD-HIT program [23] is applied to remove homologous data between datasets. As displays in
Table ***2***, in this work, with the use of 40% sequence identity, the final training dataset containing of 296 positive and 1051 negative data; the final independent testing dataset consisting of 28 positive and 332 negative data.

Table 2: Data statistics of removing homologous fragmets using CD-HIT with various values of sequence identity

Sequence identity	Training set (600 proteins)		Independent testing set (70 proteins)	
	Positive data	Negative data	Positive data	Negative data
100% (original)	634	5808	84	954
90%	500	4005	75	607
80%	450	3252	68	544
70%	421	2815	40	450
60%	380	2090	35	402
50%	341	1680	30	361
40%	296	1051	28	332

2.2. Features extraction ad encoding

In order to construct the predictive models for the identification of SFCP sites, support vector machine was adopted to distinguish SFCP sites from non-SFCP sites based on sequence-based features of substrate specificities, including: Amino Acid Composition (AAC), Amino Acid Pairwise Composition (AAPC), and Evolutionary information (PSSM, Position-Specific Scoring Matrix). These features are extracted and encoded for the final training and testing datasets that are achieved previously. The detail information of encoding for these features is as follows:

AAC feature: To encode for this feature, a 20-dimensional vector $(x_i, i = 1,2, \dots ,20)$ is utilized. This vector consisted of twenty elements, which represent the twenty types of amino acids, specifying the number of its occurrences normalized with the total number of residues in the fragment.

AAPC feature: A 20x20-dimensional matrix is used to encode feature, that has been extracted from a fragment. The 20x20-dimensional element $(x_{ij}, i, j = 1,2, \dots ,20)$ present the number of occurrences of amino acid pairwise normalized with the total number of amino acid pairs in the fragment.

PSSM feature: The PSSM (Position-Specific Scoring Matrix) is a type of evolutionary information that is commonly used for reprentation of motifs (patterns) in biological sequence. It is a matrix based on the amino acid frequencies (or nucleic acid frequencies) at every position of a multiple alignment.

Hybrid features: In addition to single features, the four hybrid features, that are formed by combining the single features, have been assessed, including: AAC+AAPC, AAC+PSSM, AAPC+PSSM, and AAC+AAPC+PSSM.

2.3. Model construction, learning and evaluation

Support vector machine (SVM) is adopted to construct the predictive models, and then learn the SVM classifiers based on extracted features. According to binary classification, the SVM using a kernel function maps the input samples into a higher dimensional space, and then finds out a a hyper-plane to discriminate between the two classes with maximal margin and minimal error. In this work, a public SVM library, LibSVM [25], is utilized to implement the predictive models for discriminating the SFCP sites from non-SFCP sites. Similar to previous works, the radial basis fuction (RBF) is selected as the kernel function for learning in the SVM classifiers, defined follows the formula:

$$K(S_i, S_j) = \exp(-\gamma ||S_i - S_j||^2) .$$

In the SVM learning, two supporting factors to enhance the performance of the models are cost and gamma. The RBF kernel function is determined by the gamma value, whereas the hyper-plane softness is controlled by cost value. To find the best final model, the predictive performance of models

using different parameters is evaluated by performing five-fold cross-validation. The training dataset is divided into five approximately equal sized subgroups. The five-fold cross-validation process is run five times with each subgroup is selected as testing dataset and the remaining subgroups are selected as training dataset. The five results are then combined to produce a single estimation for the five-fold cross-validation evaluation. The five-fold cross-validation has advantages in improving the reliability of evaluation because it considers all original data are regarded as both training and testing dataset, with each data is used for validation exactly once [16-19, 26]. In order to assess the predictive performance of trained models, the followings measures are often used: Sensitivity (SEN), Specificity (SPE), Accuracy (ACC), and Matthews Correlation Coefficient (MCC):

$$SEN = \frac{TP}{TP + FN} \quad ; \quad SPE = \frac{TN}{TN + FP} \quad ; \quad ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

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

Where: TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively.

2.4. Substrate motif discovery for the identification of S-farnesyl cysteine sites

Recent advancements of bio-technology and informatics on high-throughput of mass strometry-based proteomics, make a rapid increasing number of experimentally verified SFCP sites being available for researchers. However, the there still a lack of clues to help identify the SFCP sites. Therefore, we are motivated to discover the potential substrate motif of SFCP sites. In this work, Maximal Dependence Decomposition (MDD) [18, 28] is adopted to explore subtrate motif for the identification of SFCP sites. MDD has been shown to be effective in clustering splice sites for the purpose of splice site prediction, as well as identifying useful substrate motifs [16-18, 28].

MDD adopts the chi-square $\chi^2(A_i, A_j)$ test to assess the dependence of amino acid occurrence between two positions A_i and A_j that surround the S-farnesyl cysteine prenylation (Figure 2). To cluster SFC-data using MDD, the 20 different amino acids are first categorized into five subgroups. Subsequently, a contingency table representing the amino acid occurrences between two positions is constructed. The chi-square test is defined as:

$$\chi^2(A_i, A_j) = \sum_{m=1}^{5}\sum_{n=1}^{5}\frac{(X_{mn} - E_{mn})^2}{E_{mn}}.$$

When a strong dependence (defined as $\chi^2 \geq 34.3$, corresponding to a cut-off level of α=0.005 with 16 degrees of freedom) is detected, the decomposition will be preceeded [28].

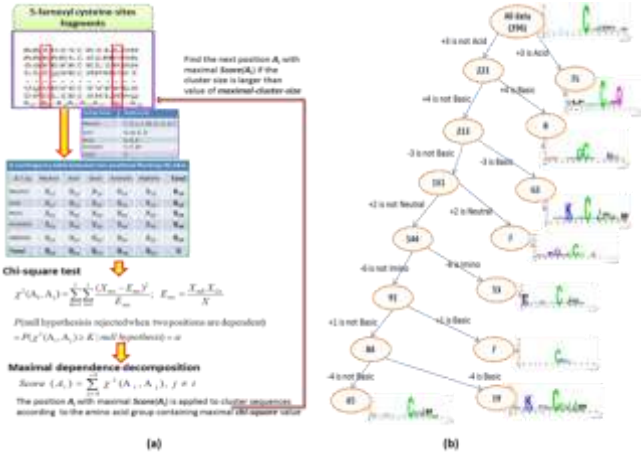


Figure 2: The analytical flowchart of motif discovery by using MDDLogo: (a). Detecting the maximal dependence of positions by using chi-square test; (b). The Tree-like visualization of MDDLogo-clustering result in this work.

3. Results and Discussion

3.1. Effection amino acid composition and single features in the identification of SFCP sites

To examine the position-specific amino acid compostion for S-farnesyl cysteine prenylation sites, WebLogo [29] is applied to generate the graphical sequence logo for the relative frequency of the corresponding amino acid at postions surrounding S-farnesyl cysteine sites.

The flanking sequences of substrate sites (at position 0) could be graphically visualized in the entropy plots of the sequence logo generated by WebLogo [29], such that the conservation of the amino acids around the SFCP sites could be easily observed. The identified motifs are subsequently evaluated on their ability to distinguish SFCP sites from non-SFCP sites by five-fold cross-validation. The AAC presents the fraction of each amino acid in a protein sequence, whereas the AAPC is used to encapsulate the global information about each protein sequence. Therefore, the investigation of the composition of flanking amino acids (AAC, AAPC) surrounding the SFCP could contribute to the identification of the potential SFCP sites.

Investigation of the differences between the AAC surrounding SFCP sites and those of non-SFCP sites shows that the overall trends are similar with slight variations. As shown in Figure 3 (c), prominent amino acid residues including Ala (A), Ser (S), Gly (G), and Lys (K), and Met (M); while Trp (W), Try (Y), and Phe (F), are three of the least significant amino acid residues. Sequence logo displays the most enriched residues surrounding the SFCP sites (Cysteine C). Also, as shown in Figure 3 (a), it shows that the most conserved amino acid residues including of Phe (F), Lys (K), Ser (S), Met (M), and Val (V). In addition, the difference between SFC-sites and non-SFC sites is visualized using TwoSampleLogo [30]. The enriched residues appear to be Phe (F), Pro (P), Ser (S), Gly (G) and Met (M); whereas the depleted amino acid residues include Val (V), Leu (L), Glu (E), Lys (K), and Gly (D) (Figure 3 (b)). An SVM model is trained to examine the effectiveness of AAC in SFCP sites.

This SVM model used a 20-dimensional vector comprising of the composition scores for twenty types of amino acids. In order to evaluate the AAC-based model, the five-fold cross-validation is applied. As shown in
Table ***3***, the model yields 91.91% accuracy, and an MCC value of 0.7998. Also, the AAPC-based model is trained to investigate the ability of AAPC and PSSM in identifying SFCP sites. The accuracy and MCC of the AAPC-based model reaches 88.27% and 73.78, respectively.

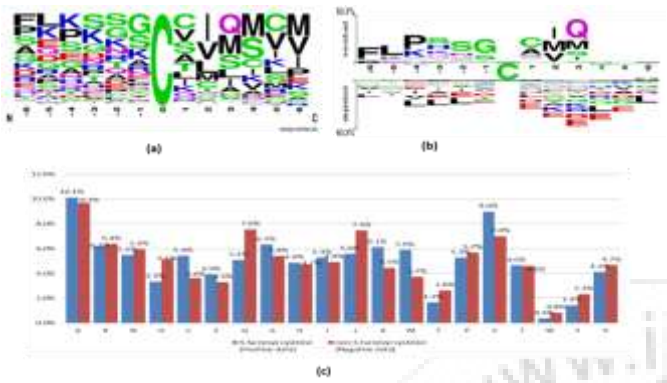


Figure 3: The graphical sequence logo showing the relative frequency of the corresponding amino acid at positions surrounding S-farnesyl cysteine sites

In addition to the composition of flanking amino acids, the evolutionary information (PSSM) is also investigated. Several amino acid residues of a protein can go through mutation without changing its structure, and two proteins may share similar structures with different amino acid composition. Evolutionary conservation usually reflects important biological function, and posttranslational modifications are prone to occur in conserved protein segments. The PSSM profiles are generated using the BLAST program through three iterations and default values of parameters. As presented in
Table ***3***, the five-fold cross-validation shows that the PSSM-based models yielded 92.68% accuracy, and the MCC value of 0.807.

Table 3: Performance evaluation by Five-fold cross-validation

Feature	SEN	SPE	ACC	MCC
AAC	96.95%	90.49%	91.91%	0.800
AAPC	98.31%	85.44%	88.27%	0.738
PSSM	96.28%	91.76%	92.68%	0.807
AAC+AAPC	96.66%	92.96%	93.78%	0.839
AAC+PSSM	95.33%	93.62%	94.00%	0.842
AAPC+PSSM	95.33%	93.52%	93.93%	0.840
AAC+AAPC+PSSM	98.31%	92.96%	94.14%	0.850

3.2. Effection of hybrid features in identifying SFCP sites

It is straightforward and very beneficial to combine two or more different approaches in machine learning to exploit advantages from them. Various methods have been applied to predict protein sites [16-18, 31]. In our approach, hybrid features are built from the incorporation of two or more single features in order to form new features for the investigation. As a consequence, the hybrid features are found to be the most effective in predicting protein S-farnesyl cysteine prenylation sites.

The performance of the model when tested with the hybrid features using the training data and independent testing data is shown in
Table ***3*** and Table 4, respectively. The hybrid feature "AAC+AAPC+PSSM" has been demonstrated to generate the best model which achieves the highest performance, with 94.14% accuracy, and an MCC value of 0.8503. This indicates that the hybrid feature "AAC+AAPC+PSSM" would generate the most promising prediction results.

Table 4: Performance evaluation by Independent testing

Feature	SEN	SPE	ACC	MCC
AAC	85.71%	92.47%	91.94%	0.611
AAPC	89.29%	93.98%	93.61%	0.674
PSSM	89.29%	94.28%	93.89%	0.683
AAC+AAPC	92.86%	94.58%	94.44%	0.715
AAC+PSSM	89.29%	94.28%	93.89%	0.683
AAPC+PSSM	85.71%	94.28%	93.61%	0.661
AAC+AAPC+PSSM	**96.43%**	**94.88%**	**95.00%**	**0.747**

3.3. Independent testing performance

As mentioned previously, to assess the practicability of the trained models, an independent testing data set is constructed by randomly selected 70 uniqes proteins from the final-non-redudant data. After several technical steps and pre-processing, the independent testing data set comprises 28 positive and 332 negative data. The performance of the model when tested on the independent testing data set is shown in Table 4. The model constructed with the hybrid feature "AAC+AAPC+PSSM" delivers the best performance, with 95.00% accuracy, and an MCC value of 0.747. This evidences for the strength of our proposed method. Furthermore, this suggests that the hybrid approach of combining single features could be an effective and promissing approach.

In addition, our proposed method is also compared with a recent prediction tool on SFCP site, GPS-Lipid [13]. As shown in
Table ***5***, our proposed method achievied higher performance. This demonstrated the ability of our model in the prediction of SFCP sites.

Table 5: Performance compared with other prediction tools using Independent testing

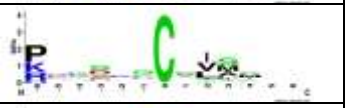
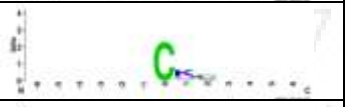
Tool	SEN	SPE	ACC	MCC
GPS-Lipid	9.06%	100.00%	21.94%	0.118
Our method	96.43%	94.88%	95.00%	0.747

3.4. Substrate motif discovery for the identification of S-farnesyl cysteine sites

MDD adopts a recursive chi-square test to evaluate the dependence of amio acid occurrence between two position surround the SFCP-sites. In this work, MDD is applied to sub-divide the positive training data (296 SFC-site fragments) to eight subgroups containing significant substrate motifs. The negative data for each MDD-clustered

subgroups are randomly selected from the negative training (1051 non-SFC-site fragments) with a ratio approximately equal to 1:3.551 (same as the ratio of positive training to negative training–296:1051). As a result, the eight useful substrate motifs are displayed in Table 6.

Table 6: Substrate motif detected by MDD

Sub-group	No. of positive data	No. of negative data	Substrate site motif
1	62	220	
2	53	188	
3	19	67	
4	65	231	
5	7	25	
6	7	25	
7	8	28	
8	75	267	

In addition, as shown in Table 6 and Table 7, MDD-clusters containing Lysine (K), Proline (P) and Glutamine (Q) residues in conserved motifs appear to generate better performances. For example, the evaluation by independent testing shows that the MDD cluster 2, consisting of Proline (P) and Lysine (K) residues at position -6 of conserved motifs, yields an accuracy of 98.61%. Similarly, MDD cluster 3, which is comprised of Lysine (K) residues at position -4 in conserved motifs, obtains 98.19% accuracy. In general, almost all clusters containing conserved Lysine (K), Proline (P) and Glutamine (Q) at specific positions, could yield good performance. In contrast, other clusters without clearly conserved motifs containing Lysine (K), Proline (P) and Glutamine (Q) residues generally show lower sensitivity. This suggests that, for the identification of SFCP sites, the substrate site specificities may depend on the conserved position of Lysine (K), Proline (P) and Glutamine (Q) residues.

Table 7: Independent testing performance for MDDLogo-clustered models.

Models	SEN	SPE	ACC	MCC
Single Model (all data) (Without MDD)	96.43%	94.88%	95.00%	0.747
MDD-Model 1	96.43%	98.49%	98.33%	0.893
MDD-Model 2	100.00%	98.49%	98.61%	0.914
MDD-Model 3	100.00%	98.19%	98.33%	0.899
MDD-Model 4	75.00%	93.07%	91.67%	0.557
MDD-Model 5	78.57%	93.07%	91.94%	0.580
MDD-Model 6	92.86%	96.08%	95.83%	0.766
MDD-Model 7	96.43%	96.08%	96.11%	0.788
MDD-Model 8	100.00%	97.89%	98.06%	0.885
Combined-MDD Models	**92.41**%	**96.42**%	**96.11**%	**0.785**

4. Conclusions

Protein S-farnesyl cysteine prenylation was a kind of post-translational modification that plays critical roles for many cellular processes such as DNA replication, signaling and trafficking, found in all eukaryotic cells. It comprises an attachment of S-farnesyl isoprenoid, which are typically involved in mediating not only protein-membrane but also protein-protein interactions. Inhibition of S-farnesyl cysteine prenylation has been extensively investigated to suppress the activity of oncogenic Ras protein to achieve antitumor activity. The current status of prenyltransferase inhibitors have been accounted to be as potentially therapeutics against several diseases, including: cancers, progeria, aging, parasitic diseases, bacterial and viral infections. In this study, we present a new schema for the identification of S-farnesyl cysteine prenylation sites based on substrate specificities. The SVM models based on various features are constructed and investigated. The hybrid feature “AAC+AAPC+PSSM” has been found to generate the best model that yields the highest performance. Evaluation of the proposed model using an independent testing reveals the strength of our proposed method in comparison with existing prediction tools. In addition, the eight useful substrate motifs, which are discovered by MDD, provide promissing clues for biologist to recognize and identify the protein SFCP sites.

References

[1] Kamiya, Y., et al., Structure of rhodotorucine A, a novel lipopeptide, inducing mating tube formation in Rhodosporidium toruloides. Biochem Biophys Res Commun, 1978. 83(3): p. 1077-83.
[2] Farnsworth, C.C., et al., Human lamin B contains a farnesylated cysteine residue. J Biol Chem, 1989. 264(34): p. 20422-9.
[3] Wolda, S.L. and J.A. Glomset, Evidence for modification of lamin B by a product of mevalonic acid. J Biol Chem, 1988. 263(13): p. 5997-6000.
[4] Soni, R., et al., Structure-based binding between protein farnesyl transferase and PRL-PTP of malaria parasite: an interaction study of prenylation process in Plasmodium. J Biomol Struct Dyn, 2016: p. 1-12.
[5] Novelli, G. and M.R. D'Apice, Protein farnesylation and disease. J Inherit Metab Dis, 2012. 35(5): p. 917-26.
[6] Maurer-Stroh, S., et al., Towards complete sets of

farnesylated and geranylgeranylated proteins. PLoS Comput Biol, 2007. 3(4): p. e66.
[7] Palsuledesai, C.C. and M.D. Distefano, Protein prenylation: enzymes, therapeutics, and biotechnology applications. ACS Chem Biol, 2015. 10(1): p. 51-62.
[8] Hechinger, A.K., et al., Inhibition of protein geranylgeranylation and farnesylation protects against graft-versus-host disease via effects on CD4 effector T cells. Haematologica, 2013. 98(1): p. 31-40.
[9] Charron, G., et al., Prenylome profiling reveals S-farnesylation is crucial for membrane targeting and antiviral activity of ZAP long-isoform. Proc Natl Acad Sci U S A, 2013. 110(27): p. 11085-90.
[10] Geryk-Hall, M., Y. Yang, and D.P. Hughes, Driven to death: Inhibition of farnesylation increases Ras activity and promotes growth arrest and cell death [corrected]. Mol Cancer Ther, 2010. 9(5): p. 1111-9.
[11] Goodsell, D.S., The molecular perspective: protein farnesyltransferase. Oncologist, 2003. 8(6): p. 597-8.
[12] Einav, S. and J.S. Glenn, Prenylation inhibitors: a novel class of antiviral agents. J Antimicrob Chemother, 2003. 52(6): p. 883-6.
[13] Yubin Xie, Y.Z., Hongyu Li, Xiaotong Luo, Zhihao He, Shuo Cao, Yi Shi, Qi Zhao, Yu Xue, Zhixiang Zuo and Jian Ren*, GPS-Lipid: a robust tool for the prediction of multiple lipid modification sites. Scientific Reports. 2016. (Accepted), 2016.
[14] Maurer-Stroh, S. and F. Eisenhaber, Refinement and prediction of protein prenylation motifs. Genome Biol, 2005. 6(6): p. R55.
[15] Chen, W.N., et al., Particle Swarm Optimization With an Aging Leader and Challengers. IEEE Transactions on Evolutionary Computation, 2013. 17(2): p. 241-258.
[16] Nguyen, V.N., et al., A new scheme to characterize and identify protein ubiquitination sites. IEEE/ACM Trans Comput Biol Bioinform, 2016.
[17] Nguyen, V.N., et al., Characterization and identification of ubiquitin conjugation sites with E3 ligase recognition specificities. BMC Bioinformatics, 2015. 16 Suppl 1: p. S1.
[18] Lee, T.Y., et al., Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinformatics, 2011. 27(13): p. 1780-7.
[19] Lee, T.Y., et al., SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. PLoS One, 2011. 6(7): p. e21849.
[20] Boeckmann, B., et al., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003. 31(1): p. 365-70.
[21] Lu, C.T., et al., DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res, 2013. 41(Database issue): p. D295-305.
[22] Keshava Prasad, T.S., et al., Human Protein Reference Database--2009 update. Nucleic Acids Res, 2009. 37(Database issue): p. D767-72.
[23] Huang, Y., et al., CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics, 2010. 26(5): p. 680-2.
[24] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997. 25(17): p. 3389-402.
[25] Lin, C.-C.C.a.C.-J., LIBSVM: a library for support vector machines. Acm Transactions on Intelligent Systems and Technology, 2011.
[26] Nguyen, V.N., et al., UbiNet: an online resource for exploring the functional associations and regulatory networks of protein ubiquitylation. Database (Oxford), 2016. 2016.
[27] Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta, 1975. 405(2): p. 442-51.
[28] Burge, C. and S. Karlin, Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997. 268(1): p. 78-94.
[29] Crooks, G.E., et al., WebLogo: a sequence logo generator. Genome Res, 2004. 14(6): p. 1188-90.
[30] Vacic, V., L.M. Iakoucheva, and P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics, 2006. 22(12): p. 1536-7.
[31] Tung, C.W. and S.Y. Ho, Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics, 2008. 9: p. 310.

Author Profile

Van-Nui Nguyen was born in Vietnam. He obtained his PhD degree in Department of Computer Science & Engineering from Yuan Ze University, Taiwan. His research interests include computer science, fuzzy constraint network, bioinformatics, computational biology and data mining.

Hai-Minh Nguyen was born in Vietnam. He obtained his PhD degree in School of Information Technology, Kyungpook National University, Korea. His research interests include computer science, biomedical, bioinformatics and data mining.