CSS-Palm: Palmitoylation Site Prediction with a Clustering and Scoring Strategy (CSS)

Fengfeng Zhou[1,†], Yu Xue[2,†], Xuebiao Yao[2,3,*], and Ying Xu[1,*]

[1]Computational Systems Biology Laboratory, Department of Biochemical and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA;

[2]Laboratory of Cellular Dynamics, Hefei National Laboratory for Physical Sciences, and the University of Science and Technology of China, Hefei, China 230027

[3]Department of Physiology and Cancer Research Program, Morehouse School of Medicine, Atlanta, GA 30310; USA

***Running title*:** palmitoylation site prediction

Keywords: palmitoylation sites; clustering and scoring strategy; lipidation; human; CSS-Palm

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

*To whom correspondence should be addressed.

Ying Xu, Tel: 706-542-9779, Email: xyn@bmb.uga.edu

Xuebiao Yao, Tel: 86-551-3606304, Email: yaoxb@ustc.edu.cn

Supplementary materials – data retrieving

We searched PubMed with the key word “palmitoylation”, and collected 210 unambiguously experimental verified palmitoylation sites from ~300 scientific articles. Although the palmitoylation-related literature is increasing rapidly, we only adopted the palmitoylation sites published online before Apr. 7th, 2005. Then we retrieved the primary sequences from Swiss-Prot/TrEMBL database (http://cn.expasy.org).

Supplementary materials – algorithm designing and validation

Scoring strategy

We suppose that a short peptide surrounding the Cysteine (C) amino acid contains sufficient information for whether the residue could be palmitoylated. This hypothesis will be supported by the satisfying Jack Knife validation results. We define a peptide with m upstream and n downstream amino acids around a Cysteine amino acid as a *Palmitoylation Site Peptide* PSP(*m*, *n*).

For a short peptide PSP(m, n) of a known palmitoylation site and a given peptide PSP(m, n), if all the amino acids except one are the same according to their positions, we may assume with confidence that the given peptide could also be palmitoylated, especially when the pair of different amino acids have similar biochemical characteristics. For example, such pairs are Isoleucine (I) and Valine (V), or Tryptophan (W) and Tyrosine (Y). There are three examples in the known palmitoylation sites for the above pairs (see in Table 1). In the following work, we mainly focused on the peptide PSP (7, 7). For those located the N-terminus or C-terminus, we insert some space characters “-“ to lengthen them to 15 amino acids. Another interesting pair of palmitoylation site peptides is found (see in Table 2) that they have differences only in the left side with a few spaces.

Substrate	Position	PSP(m, n)	PMID
P21926	86	GAVQESQC	12575999
P60033	89	GAIQESQC	14966136
Q9H3Z4	122	CYCCCC	8034679
P49795	45	CWCCCC	8986788
Q9H3Z4	124	CYCCCC	8034679
P49795	47	CWCCCC	8986788

Table 1. Three pairs of palmitoylation site peptides with only one amino acid difference.

Substrate	Position	PSP(m, n)	PMID
Q6NT27	10	LESIMACCLSEEAKE	7536745
Q8MUI9	4	----MACCLSEEAKE	12892752

Table 2. An interesting pair of palmitoylation site peptides with only a few spaces in the left or right side.

We used the amino acid substitution matrix BLOSUM62 (Henikoff and Henikoff, 1992) to evaluate the similarity between two peptide sequences with length 15 amino acids. Although other matrices could be used, the BLOSUM62 matrix is chosen here. For two amino acids a and b, let the substitution score between them in BLOSUM62 be *S*(a, b). The score between amino acids a and b is defined as:

$$Score(a,b)=\begin{cases} S(a,b), & if\ ((a\neq\text{"–"})\ and\ (b\neq\text{"–"})) \\ 0, & otherwise \end{cases}$$

And the similarity between the two peptides A and B with length 15 AA is defined as:

$$S(A,B)=\begin{cases} \sum_{i=1}^{15} Score(A[i],B[i]), & if\ \sum_{i=1}^{15} Score(A[i],B[i])\geq 0 \\ 0, & otherwise \end{cases}$$

Due to the intrinsic characteristics of BLOSUM62, two peptides may have similar biochemical properties if the similarity score between them is high enough.

Clustering strategy

The features for palmitoylation are quite elusive and almost all the existing studies proposed that there is no common canonical consensus sequence/motif for palmitoylation (Bijlmakers and Marsh, 2003; el-Husseini Ael and Bredt, 2002; Linder and Deschenes, 2003; Smotrys and Linder, 2004; ten Brinke, et al., 2002). So we suppose that the *bona fide* pattern might consist of multiple different features. Based on this hypothesis, we divide the set of known palmitoylation sites into several clusters according to the similarities between them.

The clustering algorithm is described as follows:

Algorithm. ***B****LOSUM62-****b****ased* ***C****lustering method (BBC)*

Input: A set of known palmitoylation site peptides $P=\{P_1, P_2, ..., P_n\}$, and a cut-off score S.

Output: The set of clusters C.

Operations:

1. Construct a graph $G=(V, E)$, where the node set $V=\{v_1, v_2, ..., v_n\}$, and $E=\{\}$.
2. for any pair of peptides P_i and P_j
   if $S(P_i, P_j)>=S$, then $E=E\cup\{(v_i, v_j)\}$.
3. *Return* all the peptides in one connected component in graph G as one cluster.

After this processing, the sets of known palmitoylation site peptides are divided into several clusters according to a given cut-off score. In our case, we chose 12 as the cut-off score, and got three clusters.

Clustering and Scoring Strategy

Based on the scoring and clustering strategies described above, we present the following algorithm to score the possibility of a potential palmitoylation site peptide *P*.

Algorithm. ***C****lustering and* ***S****coring* ***S****trategy for* ***Palm****itoylation Sites (CSS-Palm)*

Input: The set of clusters *C* got in procedure ***BBC***, and a potential palmitoylation site peptide *Seq*.

Output: The score for the possibility that *Seq* is a palmitoylation site peptide.

Operations:

For each cluster C_i in C

{

  For each peptide P_j in C_i

  {

    Calculate the score $S(Seq, P_j)$

  }

  $$S_i = \left\{ \sum_j S(Seq, P_j) \right\} / |C_i|$$

}

$$S(Seq) = \max_i \{S_i\}$$

Return $S(Seq)$.

The higher the score of a peptide sequence by *CSS-Palm* is, the higher confidently we may assert that this peptide may be palmitoylated. The basic idea, the similarity score between two peptides, is so simple that biologists could calculate it by hand. And even the similarity score itself delineates the similarity for palmitoylation between two peptides, e.g. the pairs of peptides in Table 1 and Table 2.

Supplementary materials – performance evaluation of CSS-Palm

Performance criteria

As far as we know, CSS-Palm is the first general-purpose *in silico* Palmitoylation site prediction system in the scientific literature. So in the following of this article, we will mainly focus on the performance of CSS-Palm itself by Jack-Knife validation.

The sensitivity (*Sn*), specificity (*Sp*), and accuracy (*Ac*) are adopted to evaluate the performance of CSS-Palm. Sensitivity and specificity illustrate the correct prediction ratios of positive and negative data sets respectively, while accuracy represents the correct ratio among both data sets. But when the number of positive data and negative data differ too much from each other, the *correlation coefficient* (*CC*) should be calculated to assess the prediction performance. The value of *CC* ranges from -1 to 1, and bigger *CC* stands for better prediction performance. As in the Jack-Knife validation, we remove one palmitoylation site per time, and then evaluate the prediction performance of CSS-Palm. The averaged performance measures are denoted as the prediction performance of CSS-Palm system.

The curated palmitoylation site peptides are the positive data, while all other Cysteine (C) amino acids in the palmitoylation substrates are regarded as the negative data. Among the data with positive predictions by CSS-Palm, the real positives are defined as *true positives* (*TP*), while the others are defined as *false positives* (*FP*). Among the data with negative predictions by CSS-Palm, the real positives are defined as *false negatives* (*FN*), while the others are defined as *true negatives* (*TN*).

The performance measures sensitivity (Sn), specificity (Sp), accuracy (Ac), and correlation coefficient (CC) are all defined as follows:

$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP},$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN},$$

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

Cut-off	*Sn* (%)	*Sp* (%)	*Ac* (%)	*CC*
4	50.65	97.62	87.00	0.5946
2.6	82.16	83.17	82.94	0.5877
1.5	97.12	44.86	56.67	0.3672

Table 3. Performance evaluation of CSS-Palm (Jack-knife validation).

Cut-off	*Sn* (%)	*Sp* (%)	*Ac* (%)	*CC*
4	40.84	100	43.34	0.1667
2.6	67.78	88.89	68.75	0.2398
2	81.20	77.78	81.03	0.2857
1.5	90.66	44.44	88.70	0.2247

Table 4. Performance evaluation of CSS-Palm (Three-fold cross validation).

Performance Discussions

As in Table 3, with the high cut-off score 4, the sensitivity and specificity of CSS-Palm reach 50.65% and 97.62% respectively. And the accuracy is 87.00%. With such high specificity and accuracy, the prediction results of CSS-Palm may be a power hint for further experimental considerations. The correlation coefficient, 0.5946, also shows the good performance of CSS-Palm.

The cut-off score of 2.6 stands for a balanced pair of sensitivity (82.16%) and specificity (83.17%), while the accuracy (82.94%) and correlation coefficient (0.5877) are certainly acceptable.

When high sensitivity is required, we may choose the cut-off score 1.5, of which sensitivity and specificity are 97.12% and 44.86% respectively. The accuracy (56.67%) and correlation coefficient (0.3672) suggest that the prediction results may need further validations by other strategies, including *in vivo* or *in vitro* experiments.

CSS-Palm provides similar prediction performance for the three-fold cross validation (Table 4).

Supplementary materials – implemented as a web server

For experimentalists' convenience, we implemented our CSS-Palm strategy as an easy-to-use web server, which can be accessed from http://bioinformatics.lcd-ustc.org/css_palm/ . The description of the server and the result discussions could be found in the article.

Supplementary figures – legends(WebLogo, Crooks, et al., 2004)

Figure S1: Sequence logo of subset 1.

Figure S2: Sequence logo of subset 2.

Figure S3: Sequence logo of subset 3.

Figure S4: Sequence logo of the total set.

ACKNOWLEDGEMENTS

X. Yao and Y Xue's work is supported by grants from Chinese Natural Science Foundation (39925018, 30270654 and 30270293), Chinese Academy of Science (KSCX2-2-01), Chinese 973 project (2002CB713700), Chinese Minister of Education (20020358051), American Cancer Society (RPG-99-173-01) and National Institutes of Health (DK56292; CA92080). X. Yao is a Georgia Cancer Coalition Eminent Research Scholar. F. Zhou and Y Xu's work is supported by the Georgia Cancer Coalition, National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204), and Department of Energy's Genomes to Life Program (http://doegenomestolife.org/) under project "Carbon Sequestration in Synechochoccus sp.: From Molecular Machines to Hierarchical Modeling.

REFERENCES

Bijlmakers, M.J. and Marsh, M. (2003) The on-off story of protein palmitoylation, *Trends Cell Biol*, **13**, 32-42.

Crooks, G.E., Hon G., Chandonia J.M., Brenner S.E. (2004) WebLogo: A sequence logo generator, Genome Research, 14:1188-1190

el-Husseini Ael, D. and Bredt, D.S. (2002) Protein palmitoylation: a regulator of neuronal development and function, *Nat Rev Neurosci*, **3**, 791-802.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, **89**, 10915-10919.

Linder, M.E. and Deschenes, R.J. (2003) New insights into the mechanisms of protein palmitoylation, *Biochemistry*, **42**, 4311-4320.

Smotrys, J.E. and Linder, M.E. (2004) Palmitoylation of intracellular signaling proteins: regulation and function, *Annu Rev Biochem*, **73**, 559-587.

ten Brinke, A., van Golde, L.M. and Batenburg, J.J. (2002) Palmitoylation and processing of the lipopeptide surfactant protein C, *Biochim Biophys Acta*, **1583**, 253-265.