Prediction of Protein Modification Sites of Pyrrolidone Carboxylic Acid Using mRMR Feature Selection and Analysis

Lu-Lu Zheng[1,2], Shen Niu[3], Pei Hao[2,3], KaiYan Feng[2], Yu-Dong Cai[4]*, Yixue Li[1,2,3]*

1 Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong University of Science and Technology, Wuhan, China, **2** Shanghai Center for Bioinformation Technology, Shanghai, China, **3** Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, **4** Institute of Systems Biology, Shanghai University, Shanghai, China

Abstract

Pyrrolidone carboxylic acid (PCA) is formed during a common post-translational modification (PTM) of extracellular and multi-pass membrane proteins. In this study, we developed a new predictor to predict the modification sites of PCA based on maximum relevance minimum redundancy (mRMR) and incremental feature selection (IFS). We incorporated 727 features that belonged to 7 kinds of protein properties to predict the modification sites, including sequence conservation, residual disorder, amino acid factor, secondary structure and solvent accessibility, gain/loss of amino acid during evolution, propensity of amino acid to be conserved at protein-protein interface and protein surface, and deviation of side chain carbon atom number. Among these 727 features, 244 features were selected by mRMR and IFS as the optimized features for the prediction, with which the prediction model achieved a maximum of MCC of 0.7812. Feature analysis showed that all feature types contributed to the modification process. Further site-specific feature analysis showed that the features derived from PCA's surrounding sites contributed more to the determination of PCA sites than other sites. The detailed feature analysis in this paper might provide important clues for understanding the mechanism of the PCA formation and guide relevant experimental validations.

Citation: Zheng L-L, Niu S, Hao P, Feng K, Cai Y-D, et al. (2011) Prediction of Protein Modification Sites of Pyrrolidone Carboxylic Acid Using mRMR Feature Selection and Analysis. PLoS ONE 6(12): e28221. doi:10.1371/journal.pone.0028221

Editor: Vladimir N. Uversky, University of South Florida College of Medicine, United States of America

Received October 4, 2011; **Accepted** November 3, 2011; **Published** December 9, 2011

Funding: This work was supporeted by the National Basic Research Program of China (2011CB510102, 2011CB510101, and 2012CB517905), National High-Tech Research and Development Program of China (863) (grant 2009AA02Z308) and Innovation Program of Shanghai Municipal Education Commission (12ZZ087). PH gratefully acknowledges the support of the SA-SIBS Scholarship Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cai_yud@yahoo.com.cn (YDC); yxli@sibs.ac.cn (YL)

Introduction

Post-translational modifications (PTMs) are crucial for proteins to maintain their structural and functional diversities in both prokaryotes and eukaryotes. They influence a protein's state of activity, localization, turnover and ability to interact with other molecules [1], which are pivotal for many cellular processes, e.g. in signal transduction, kinase induced cascades are turned off and on by the removals or additions of phosphate groups in proteins [2]. Pyrrolidone carboxylic acid (PCA), also known as pyroglutamic acid or pGlu, is produced during one of the common PTMs, and may be formed naturally by an enzymatic synthesis from N-terminal glutamine under mildly acid conditions or as an artifact from N-terminal glutamic acid under very acid conditions in proteins or peptides [3,4]. As a glutamic acid derivative that lacks a H_2O molecule [3] in extracellular and multi-pass membrane proteins, many studies have demonstrated that pyroglutamic acid is formed either late in protein translation by cyclization of the glutamine at the N-terminus or as a post-translational event, just prior to the secretion of completed proteins from the cell [5]. Modified proteins of this type usually show an increase half-life, because PCA blocks proteins, minimizing their susceptibility to degradation by aminopeptidases [3,6]. Structures containing the cyclic product of glutamine at the N-terminus are common in nature [3]. In general, those proteins or peptides have an endocrine and/or regulative function in mammalian tissues and are of high interest, since their concentrations in the blood could influence synthesis pathways of important metabolites [7]. A particularly well-studied example is the tripeptide thyrotropin releasing factor (TRF), which has the sequence pGlu-His-Pro that stimulates the release of thyrotropin in vivo [8] and has been shown to be able to enhance prolactin synthesis and decrease growth hormone production in vitro [5,9]. Hinkle and Tashjian have proved that any structural substitution in the pGlu lactam ring of TRF will significantly decrease both hormone synthesis and receptor binding ability [8]. Aside from functions as an incorporated amino acid, functions of free pyrrolidone carboxylic acid are less clear, though its pharmacological properties have been described repeatedly [6]. It has been shown that pGlu stimulates GABA releasing from the cerebral cortex and therefore produces anti-anxiety effects in a simple approach-avoidance conflict situation in the rat [10]. In addition, Silva et al. have confirmed that L-pyroglutamic acid predominantly accumulates in the inherited metabolic diseases, glutathione synthetase deficiency (GSD) and γ-glutamylcysteine synthetase deficiency (GCSD), by reducing brain CO2 production, lipid biosynthesis and ATP levels [11]. High anion gap and metabolic

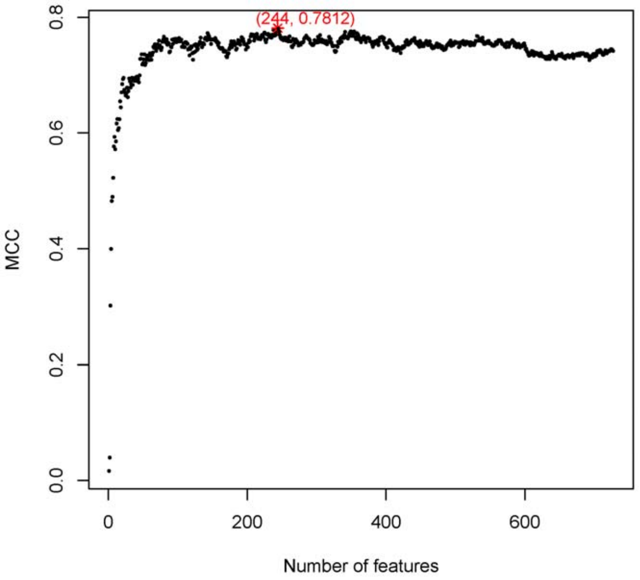

Figure 1. Distribution of MCC values against feature numbers. IFS predictive MCC values were plotted against feature numbers based on Table S1. The maximum MCC is 0.7812 using 244 features. These 244 features were considered as the optimal feature set of our classifier. doi:10.1371/journal.pone.0028221.g001

acidosis in adults is also proved to associate with accumulation of pGlu [12]. Thus, identification of protein pyroglutamic acid modification sites is of fundamental importance to understand the mechanism by which pGlu occurs in biological systems.

So far, the identification of pGlu modification sites has largely been focused on the mass spectrometry (MS) technique [13–16], which measures mass-to-charge ratio (m/z), yielding the molecular mass and the fragmentation pattern of peptides derived from proteins. Owing to its high-throughput and accuracy, MS technique represents a gold method for all modifications that change the molecular weight. Based on MS data, Wilkins et al. have developed a software tool, FindMod, to discover 22 post-translational modifications, including acetylation, phosphorylation and pyrrolidone carboxylic acid and so on [13]. They examined 5153 incidences of post-translational modifications annotated in the SWISS-PROT database, and made a total of 29 rules to predict the amino acids in the peptide that might carry the modification. For the PCA modification, they used the rules with monoisotopic and average delta masses being -17.0266 and -17.0306, the modified amino acid being Q and the position of amino acid in proteins being at the N-terminus. However, this MS method is time-consuming and costly. Developing novel methods *in silico* that merely depend on amino acid sequences is urgent, especially for large scale proteomic analysis.

dbPTM is a database that contains information about protein post-translational modifications (PTMs), such as modified sites, solvent accessibility of amino acid residues, protein secondary and tertiary structures, protein domains and protein variations [17]. From the database, we found that 689 experimentally validated residues modified to be pyrrolidone carboxylic are amino acid Glns (Q), and only 2 are amino acid Glus (E). This seems to illustrate that pyrrolidone carboxylic acid derived from a glutamate residue is extremely rare and may be correlated with an extreme acidic context that the protein involved, and also explain why FindMod used the rule that the amino acid pGlu modified is Q exclusively. dbPTM also indicates that the Q residue carrying the modification with surrounding sites usually locates in a coiled secondary structure, but no specific sequence conservation patterns can be observed in these residues, while sites around E

Table 1. Top 20 features of the optimal feature set for pyrrolidone acid modification sites.

Order	Name	Site	Feature
1	AA10-Pssm_A	10	PSSM
2	AA1-Pssm_L	1	PSSM
3	AA13-Pssm_W	13	PSSM
4	AA2-Codon Diversity	2	AAFactor
5	AA4-Pssm_L	4	PSSM
6	AA10-Pssm_N	10	PSSM
7	AA3-Pssm_L	3	PSSM
8	AA8-Pssm_A	8	PSSM
9	AA5-Polarity	5	AAFactor
10	AA10-Codon Diversity	10	AAFactor
11	AA1-Secondary Structure Helix	1	Secondary structure
12	AA8-Pssm_C	8	PSSM
13	AA10-Pssm_I	10	PSSM
14	AA9-Pssm_I	9	PSSM
15	AA13-Secondary Structure Other	13	Secondary structure
16	AA1-Pssm_N	1	PSSM
17	AA5-Pssm_A	5	PSSM
18	AA6-Pssm_Q	6	PSSM
19	AA8-Side Chain Count of Atom_C Deviation from Mean	8	Deviation of side chain carbon atom number
20	AA2-Pssm_L	2	PSSM

doi:10.1371/journal.pone.0028221.t001

residue are restricted to LTGERL. Moreover, based on kinasePhos-like method, which is to computationally predict phosphorylation sites by applying profile hidden markov model (HMM) [18], dbPTM predicted 12,322 PCA sites using protein sequences taken from Swiss-Prot. However, we cannot get enough detailed information to compare this method with others. We also noticed that the number of predicted sites is much greater than that of experimentally validated sites, implying the predictor does not perform well or many PCA modification sites have not yet be identified by experimental methods.

In this paper, we present a new computational method to predict modification sites of PCA in protein sequences. Nearest neighbor algorithm (NNA), a kind of machine learning approach, incorporated by feature selection (IFS based on mRMR), was applied to make predictions. The features used in the method come from many different sources, and can be grouped into 7 categories: position-specific conversation scoring matrixes (PSSM), amino acid factors, disorder scores, secondary structure and solvent accessibility, gain/loss of amino acids during evolution, propensity of amino acid to be conserved at protein-protein interface and protein surface, and deviation of side chain carbon atom number. Our method achieved an overall MCC of 78.12% using the optimal feature set. Feature analysis shows that the conservation of amino acids at some certain residues in the upstream of PCA modification sites plays more important roles in the prediction; it also shows that the remaining features of amino acids in the flanking regions are important for the prediction.

Materials and Methods

Dataset

We downloaded 1366 protein sequences containing modification sites of PCA from uniprot (version 2011_02) [19,20]. These protein sequences totally contain 1528 modification sites of PCA, within which we selected 769 sites annotated as "Note = Pyrrolidone carboxylic acid" for further analysis. These 769 sites include 370 sites located in the internal region and 399 sites located at the N-terminal of the relevant proteins. We only considered the 370 sites located in the internal region of the 299 protein sequences.

We extracted one 21-residue peptide fragment for each PCA site with the modification site of PCA at the centre, plus 10 residues upstream and 10 residues downstream of the modification site. The peptides with length less than 21 residues were complemented by character "-". After removing identical peptide fragments, there remain 333 fragments.

For negative samples, we extracted totally 3635 Q sites located in the internal region of the 299 protein sequences. Then we removed Q sites identical to positive samples or within the 3635 Q sites themselves, resulting in totally 2997 peptide fragments, from which we randomly selected 1665 (333*5 = 1665) peptide

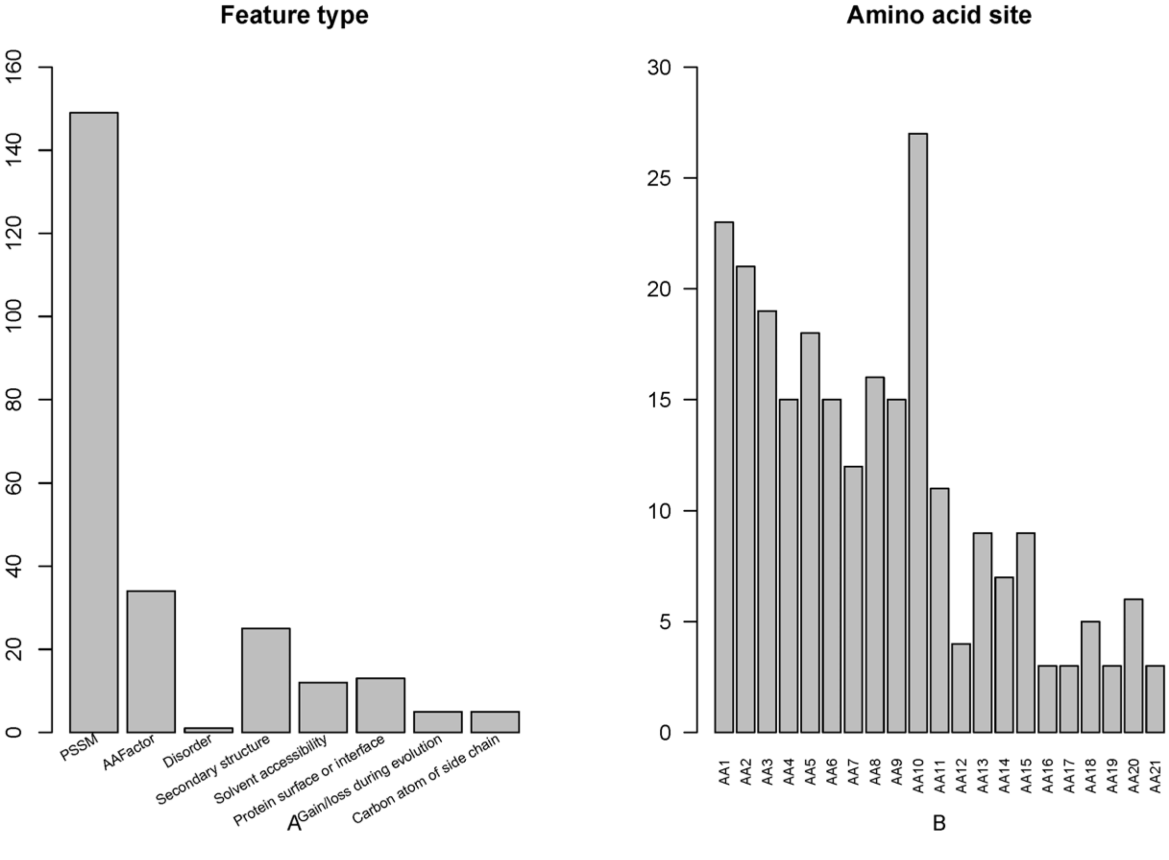

Figure 2. Feature and site specific distribution of the optimal feature set. (A) Feature distribution of the optimal feature set. Among the optimized 244 features, there were 149 features of PSSM conservation score, 34 features of amino acid factor, 1 feature of disorder, 25 features of secondary structure, 12 features of solvent accessibility, 13 features of propensity on protein surface or protein-protein-interaction interface ("Protein surface or interface"), 5 features of gain/loss during evolution and 5 features of deviation of side chain carbon atom number ("Carbon atom of side chain"). (B) Site specific distribution of the optimal feature set. The site-specific distribution of the optimal feature set revealed that site 10 played the most important role in the prediction of PCA modification. Site 1, 2, 3 and 5 also played relatively more role than the remaining sites.
doi:10.1371/journal.pone.0028221.g002

fragments as the negative samples. Thus, there were totally 1998 samples, including 333 positive samples and 1665 negative samples. Both positive and negative samples and their surrounding amino acids were given in Dataset S1.

Feature Construction

The features of PSSM conservation scores. Evolutionary conservation plays important roles in biological analysis. A more conserved residue within a protein sequence usually indicate that it is more important for protein functioning and thus under stronger selective pressure.

We used Position Specific Iterative BLAST (PSI BLAST) [21] to measure the conservation status for a specific residue. It used a 20-dimensional vector to denote probabilities of conservation against mutations to 20 different amino acids for a specific residue. For a given peptide, all such 20-dimentional vectors for all residues composed a matrix called position specific scoring matrix (PSSM). More conserved residues through cycles of PSI BLAST were suggested to be more important for biological functioning.

In this study, we used PSSM conservation score to quantify the conservation status of each amino acid in a protein sequence.

The features of amino acid factors. Since each of the 20 amino acids has various and specific properties, the composition of different residues within a protein can influence the specificity and diversity of the protein structure and function. AAIndex [22] is a database containing various physicochemical and biochemical properties of amino acids. Atchley et al. [23] performed multivariate statistical analyses on AAIndex and transformed AAIndex to five multidimensional and highly interpretable numeric patterns of attribute covariation reflecting polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. These five numerical pattern scores (denoted as "amino acid factors") have been applied successfully in many researches [24–26]. We also used the amino acid factors to represent the respective properties of each amino acid in a given protein.

The features of disorder score. Protein segments lacking fixed three-dimensional structures under physiological conditions play important roles in biological functions [27,28]. The disordered regions of proteins allow for more modification sites and interaction partners and always contain PTM sites, sorting signals, and protein ligands. Thus it is quite importance for protein structure and function [27,29,30]. In this study, VSL2 [31], which can accurately predict both long and short disordered regions in proteins, was used to calculate disorder score that denotes the disorder status of each amino acid in a given protein sequence.

The features of secondary structure and solvent accessibility. Protein structures play important roles in protein functioning and the post-translational modification of specific residues may be influenced by the solvent accessibility of the relevant residues. So we also considered protein structures including secondary structure and solvent accessibility to encode each peptide. These features were predicted by SSpro 4 [32], which classify

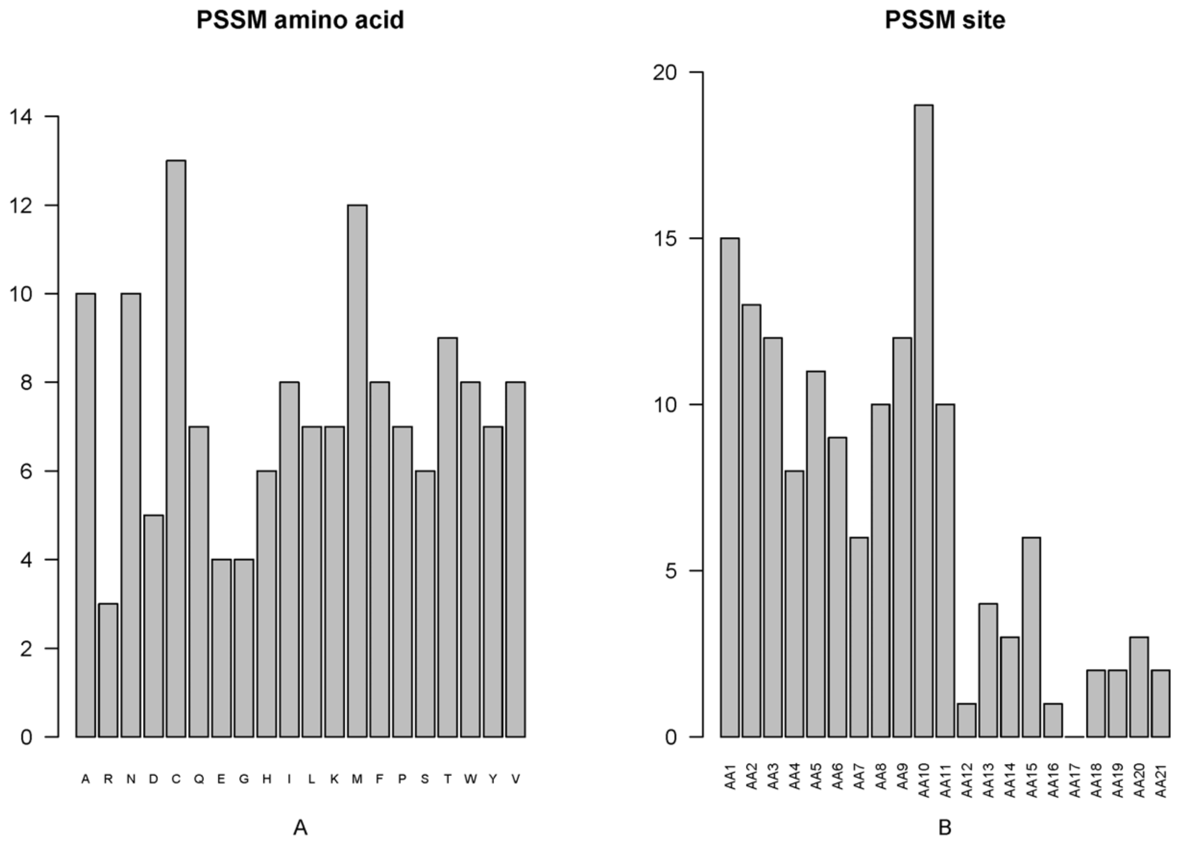

Figure 3. Feature and site specific distribution of the PSSM features in the optimal feature set. (A) Feature distribution of the PSSM features in the optimal feature set. The conservation against mutations to amino acid C, M A, N and T influent more on PCA modification determination than the mutations to other amino acids. (B) Site-specific distribution of the PSSM features in the optimal feature set. The conservation status of "AA10", "AA1" and "AA2" sites were most important for the PCA modification site prediction.
doi:10.1371/journal.pone.0028221.g003

secondary structure of each amino acid as 'helix', 'strand', or 'other', and solvent accessibility as 'buried' or 'exposed'. The usage of the secondary structure feature predicted by SSpro4 in our study could indicate whether a specific type of these three secondary structure types is associated with protein PCA modification sites.

Gain/loss of amino acids during evolution,propensity of amino acid to be conserved at protein-protein interface and protein surface. It has been suggested by Goldsmidt et al that protein folding has evolved to remove regions of high propensity and remain proper conformation for fibrillation from protein surfaces [33]. We included features of gain/loss of amino acids during evolution [34] in our analysis. Since the location of a residue on protein surface or interaction interface may also influence the determination of PCA modification site, we also included the features of conservation of an amino acid on protein-protein interaction interface and protein exposed surface [35].

Deviation of side chain carbon atom number. Different atoms may have their various intrinsic properties [36]. Different composition of atoms within a residue or peptide could also influence protein properties and thus influence protein structures and functions. So, we calculated the deviation of side chain carbon atom number for each residue within the 21-residue segment. This feature was calculated by subtracting the mean carbon atom number of side chains within a 21-residue segment by the side chain carbon number of each residue.

The feature space. Since the residue at site 11 of the 21-peptide is Q, so for this site we incorporated 27 features, including 20 features of PSSM conservation score, 1 feature of disorder score, 3 features of secondary structure, 2 features of solvent accessibility and 1 feature of deviation of side chain carbon atom number. For other residues, we incorporated totally 35 features by adding other 8 features, including 5 features of AAFactor, 2 features of propensity of amino acid to be conserved at protein-protein interface and protein surface and 1 feature of gain/loss of amino acids during evolution. Overall for the 21-residue protein segment, there are totally 35*20+27 = 727 features.

For residues denoted by "-", the carbon atom number of side chains was set to be the mean of the side chain carbon atom numbers of the 20 amino acids, and other features were set to be 0.

mRMR method

We used Maximum Relevance Minimum Redundancy (mRMR) method to rank the importance of the 727 features [37]. mRMR method could rank features based on both their relevance to the target and the redundancy of features. A smaller index of a feature denotes that it has a better trade-off between maximum relevance to target and minimum redundancy.

Both relevance and redundancy was quantified by mutual information (MI), which estimates how much one vector is related to another. The MI equation was defined as below:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \qquad (1)$$

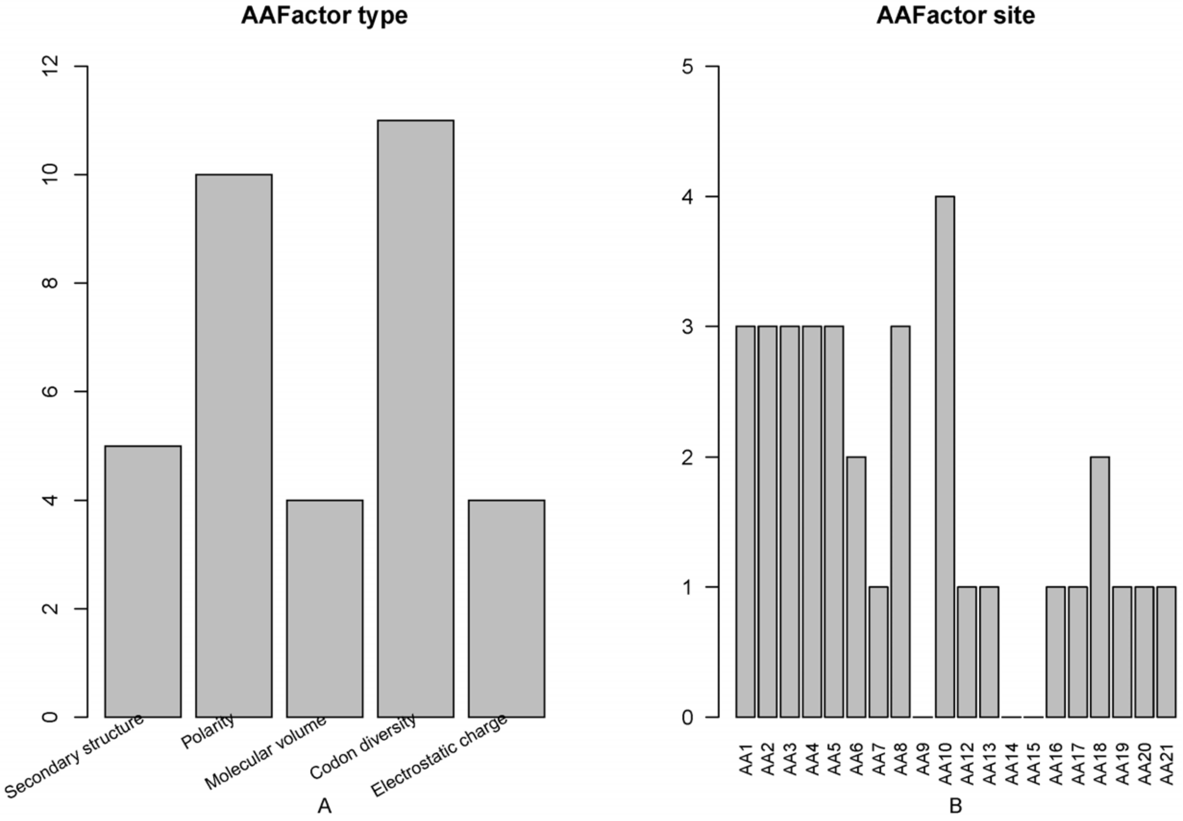

Figure 4. Feature and site specific distribution of the amino acid factor features in the optimal feature set. (A) Feature distribution of the amino acid factor features in the optimal feature set. Codon diversity and polarity features were the most important features, and secondary structure, molecular volume and electrostatic charge were almost equally important features to the PCA modification site prediction. (B) Site-specific distribution of the amino acid factor features in the optimal feature set. Residues at site 10 have the most important effect on PCA modification site prediction. Residues at site 1–5 and site 8 had relatively more effect on PCA modification site prediction.
doi:10.1371/journal.pone.0028221.g004

In equation (1),x, y are vectors, $p(x,y)$ is their joint probabilistic density, and $p(x)$ and $p(y)$ are the marginal probabilistic densities.

Ω was used to denote the whole feature set. Ω_s was used to denote the already-selected feature set containing m features and Ω_t was used to denote the to-be-selected feature set containing n features. The relevance D between the feature f in Ω_t and the target c can be calculated by:

$$D = I(f,c) \tag{2}$$

The redundancy R between the feature f in Ω_t and all the features in Ω_s can be calculated by:

$$R = \frac{1}{m}\sum_{f_i \in \Omega_s} I(f,f_i) \tag{3}$$

To get the feature f_j in Ω_t with maximum relevance and minimum redundancy, the mRMR function combined equation (2) and equation (3) and are defined as below:

$$\max_{f_j \in \Omega_t}\left[I(f_j,c) - \frac{1}{m}\sum_{f_i \in \Omega_s} I(f_j,f_i)\right] (j = 1,2,\ldots,n) \tag{4}$$

The mRMR feature evaluation would be run N rounds when given a feature set with N (N = m+n) features. After the evaluation, we get an ordered feature set S:

$$S = \left\{f_1{'}, f_2{'}, \ldots, f_h{'}, \ldots, f_N{'}\right\} \tag{5}$$

In S, index h of a feature indicates at which round that the feature is selected. The smaller the index h is, the earlier the feature satisfied equation (4) and the better the feature is.

Nearest Neighbor Algorithm

Nearest Neighbor Algorithm (NNA) was used to predict PCA modification sites. NNA calculates the similarities between the test sample and all the training samples and by which makes its classification decision. In our study, the distance between vector p_x and p_y is defined as below [38,39]:

$$D(p_x,p_y) = 1 - \frac{p_x \cdot p_y}{||p_x|| \cdot ||p_y||} \tag{6}$$

In equation (6), $||p||$ is the module of vector p. $p_x \cdot p_y$ denotes the inner product of p_x and p_y. The smaller $D(p_x,p_y)$, the more similar p_x to p_y is.

In NNA, given a training set $P = \{p_1, p_2, \ldots, p_n, \ldots, p_N\}$ and a vector p_t, p_t will be designated to the same class of its nearest neighbor p_n in P, which is the vector having the smallest $D(p_n,p_t)$:

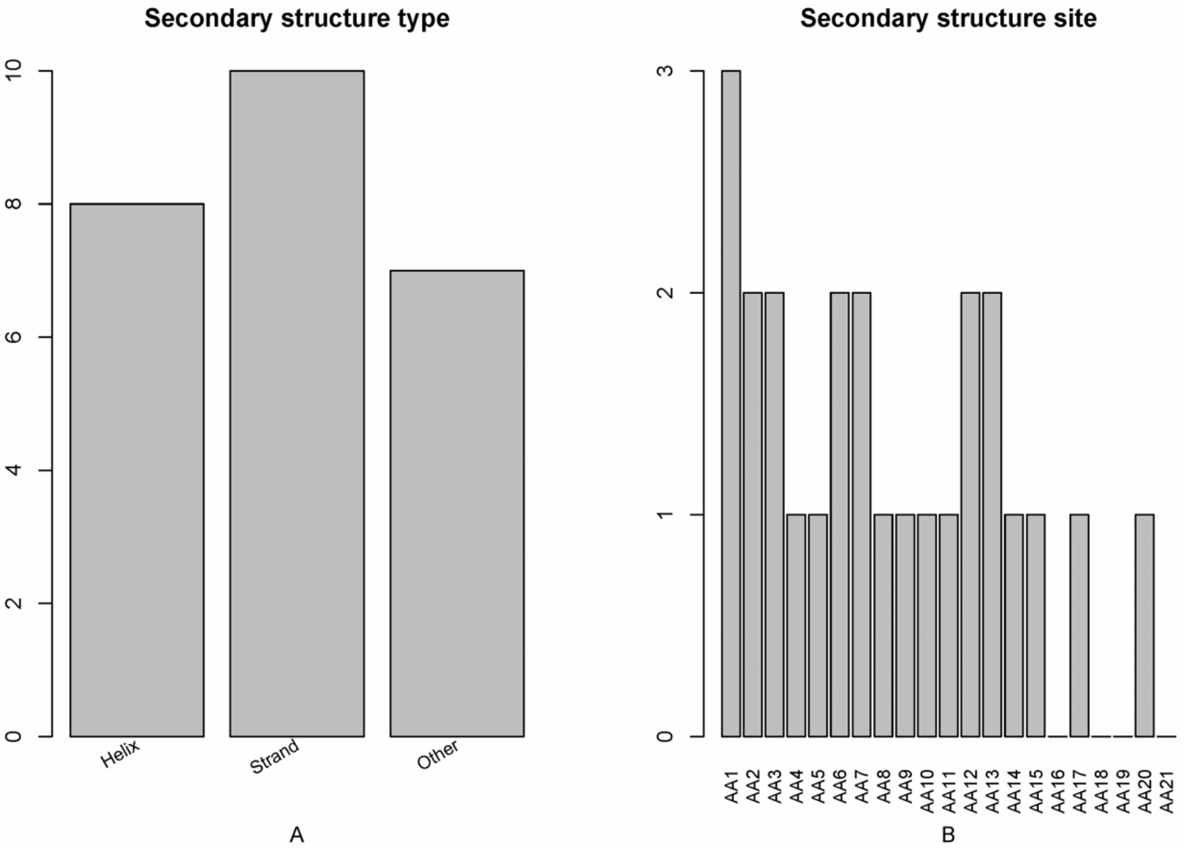

Figure 5. Feature and site specific distribution of the secondary structure features in the optimal feature set. (A) Feature distribution of the secondary structure features in the optimal feature set. All the three types of secondary structure features (helix, strand and other) can influence the PCA modification site determination. (B) Site-specific distribution of the secondary structure features in the optimal feature set. The secondary structure of site 1 may influence more to PCA modification site determination than other sites. The secondary structure features at site 2, 3, 6, 7, 12 and 13 may also influence more on PCA site determination.
doi:10.1371/journal.pone.0028221.g005

$$D(p_n,p_t)=\min\{D(p_1,p_t),D(p_2,p_t),...,D(p_z,p_t),...,D(p_N,p_t)\}(z\neq t) \quad (7)$$

Jackknife Cross-Validation Method

Jackknife Cross-Validation Method [40–42] (also called the Leave-one-out cross-validation, LOOCV) was used to evaluate the performance of a classifier. In Jackknife Cross-Validation Method, every sample is tested by the predictor that is trained with all the other samples. Let TP denotes true positive. TN denotes true negative. FP denotes false positive and FN denotes false negative. To evaluate the performance of our PCA modification site predictor, the prediction accuracy, specificity, sensitivity and MCC (Matthews's correlation coefficient) were calculated as below:

$$\begin{cases} accuracy=\dfrac{TP+TN}{TP+TN+FP+FN} \\ sensitivity=\dfrac{TP}{TP+FN} \\ specificity=\dfrac{TN}{TN+FP} \\ MCC=\dfrac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \quad (8)$$

Incremental Feature Selection (IFS)

Based on the ranked features derived from mRMR, we used Incremental Feature Selection (IFS) [39,43] to determine the optimal number of features.

During IFS procedure, features in the ranked feature set are added one by one from higher to lower rank. A new feature set is composed when one feature is added. Thus given N ranked features, N feature sets would be composed. The i-th feature set is:

$$S_i=\{f_1,f_2,...,f_i\}(1\leq i\leq N)$$

For each of the N feature sets, an NNA predictor was constructed and tested using Jackknife cross-validation test. With N prediction accuracy, sensitivity, specificity and MCC calculated, we obtain an IFS table with one column being the index i and the other columns to be the prediction accuracy, sensitivity, specificity and MCC. We then can get the optimal feature set (S_{optimal}), with which the predictor achieves the best prediction performance.

Results

mRMR result

We get the ranked mRMR feature list of 727 features using mRMR method. A smaller index of a feature suggests that it is more important for the prediction of PCA modification site. This

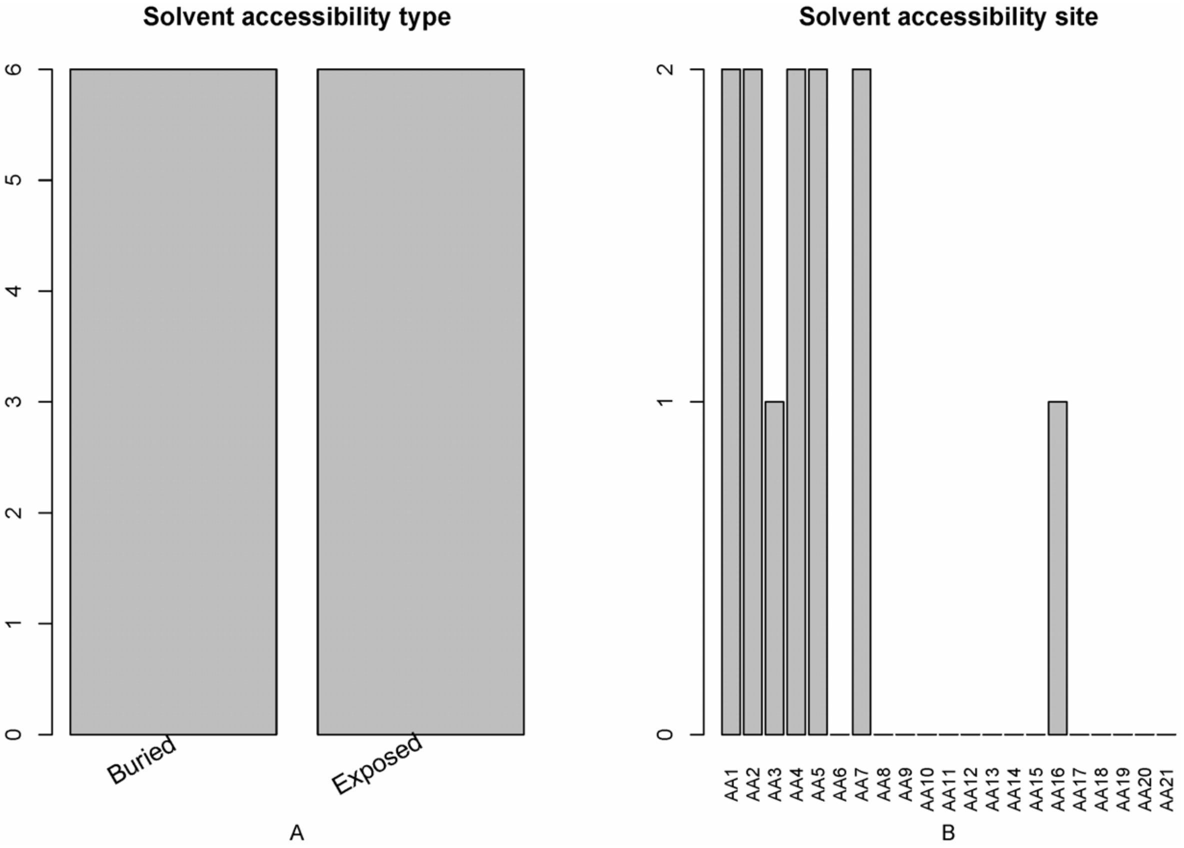

Figure 6. Feature and site specific distribution of the solvent accessibility features in the optimal feature set. (A) Feature distribution of the solvent accessibility features in the optimal feature set. The number of the two types of solvent accessibility features (buried and exposed) is equal (6), indicating that features derived from solvent accessibility are equally important for PCA site determination. (B) Site-specific distribution of the solvent accessibility features in the optimal feature set. Solvent accessibility features at site 1, 2, 4, 5, 7, 3 and 16 influences more on PCA site determination.

doi:10.1371/journal.pone.0028221.g006

ranked 727-feature list was used in IFS procedure for the selection of optimal feature set.

IFS result

By adding the ranked features one by one, we built 727 individual predictors for the 727 sub-feature sets to predict PCA modification sites. We then tested the prediction performance of each of the 727 predictors and get the IFS results (given in Table S1). Figure 1 shows the IFS curve plotted based on Table S1. The maximum MCC is 0.7812 when 244 features are included. These 244 features were considered as the optimal feature set of our classifier. Using which, the predictive sensitivity, specificity and accuracy were 0.8523, 0.9528, and 0.9355 respectively. The 244 optimal features were given in Table S2 and the top 20 features in Table 1.

Feature analysis of optimal feature set

The distribution of the number of each type of features in the optimal feature set was investigated and shown in Figure 2A. Among the optimized 244 features, there were 149 features of PSSM conservation score, 34 features of amino acid factor, 1 feature of disorder, 25 features of secondary structure, 12 features of accessibility, 13 features of propensity on surface or interface, 5 features of gain/loss during evolution and 5 features of deviation of side chain carbon atom number.

The site-specific distribution of the optimal feature set (Figure 2B) revealed that site 10 played the most important role in the determination of PCA modification sites. And site 1, 2, 3 and 5 played more important role than the remaining sites.

Feature analysis of PSSM conservation score

Among the optimized 244 features, there were 149 features of PSSM conservation score, the greatest proportion of the optimized features. We investigated the number of each kind of amino acids of the PSSM features (Figure 3A) and found that the conservation against mutations to the 20 amino acids influents differently on the prediction of PCA modification site. Mutations to amino acid C, M A, N and T influence more on PCA modification determination than the mutations to other amino acids. We also investigated the number of PSSM features at each site (Figure 3B). The conservation status of "AA10", "AA1" and "AA2" sites were most important for the prediction of PCA modification site, as shown in Figure 3B.

Feature analysis of amino acid factor

The number of each type of amino acid factor features (Figure 4A) and the number of amino acid factor features at each site (Figure 4B) were analyzed. It was found that codon diversity and polarity were the most important features, and secondary structure, molecular

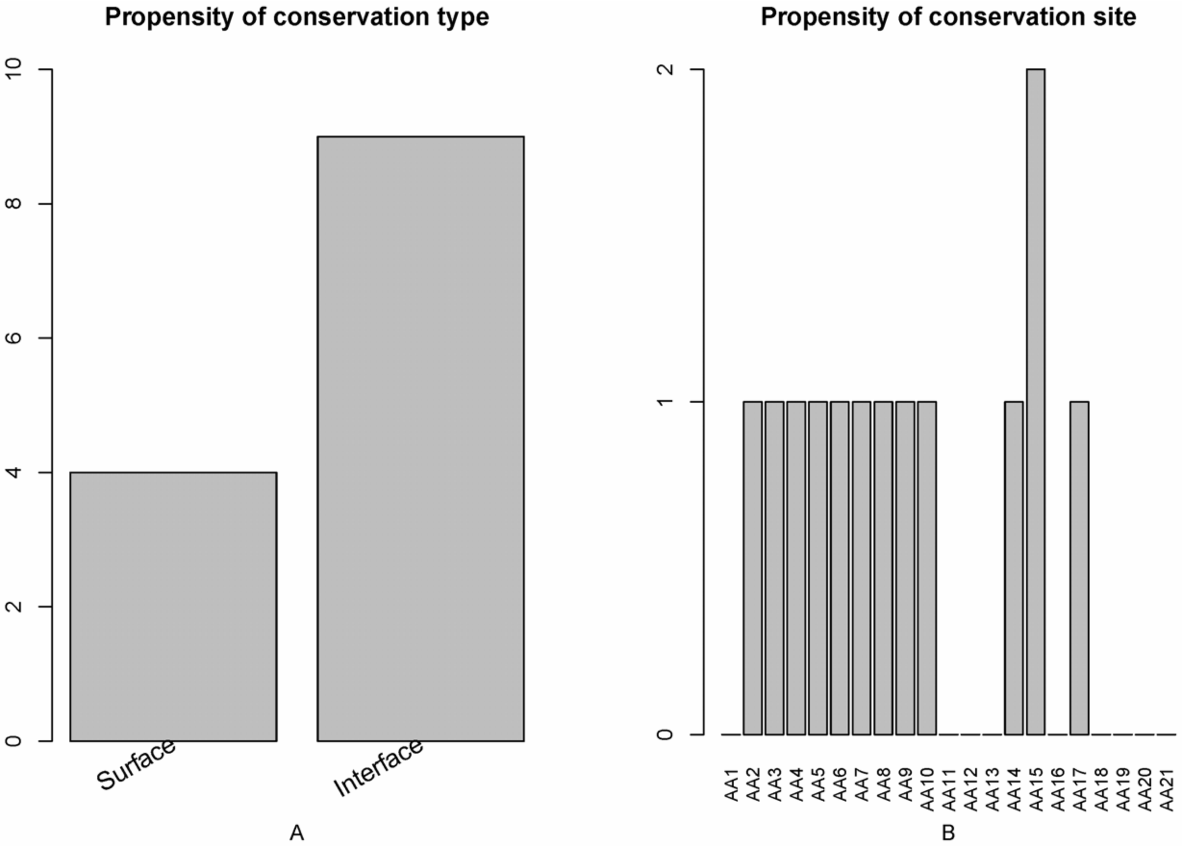

Figure 7. Feature and site specific distribution of the propensity of amino acid to be conserved at protein-protein interface and protein surface features in the optimal feature set. (A) Feature distribution of these features in the optimal feature set. Propensity of amino acid to be conserved at protein-protein interface influence more on PCA modification site determination and propensity of amino acid to be conserved at protein surface also can influence PCA modification site determination. (B) Site-specific distribution of these features in the optimal feature set. Propensity of amino acid to be conserved at protein-protein interface and protein surface features at site 15, 2–10, 14 and 17 influence more on PCA modification site determination.
doi:10.1371/journal.pone.0028221.g007

volume and electrostatic charge were almost equally important features. In Figure 4B, residues at site 10 contribute most to the prediction of PCA modification site, followed by residues at site 1–5 and site 8 which contribute less, and then the residues at the remaining sites which contribute least to the prediction.

Feature analysis of disorder score

Among the optimal feature set, only 1 disorder feature was selected, the disorder feature at site 14, having an index of 77.

Feature analysis of secondary structure and solvent accessibility

The feature- and site-specific distribution of the secondary structures in the optimal feature set was shown in Figure 5. From Figure 5A, we can see that all three types of secondary structures (helix, strand and other) influence the PCA modification site determination. From 5B, we can see that secondary structure of site 1 may influence more to the determination of PCA modification site than other sites. The secondary structures at site 2, 3, 6, 7, 12 and 13 may also influence more on PCA site determination.

We also investigated the 12 features of solvent accessibility in the optimal feature set (Figure 6). As shown in Figure 6A, the number of the two types of solvent accessibility features (buried and exposed) is equal (6) to each other, indicating that both types of solvent accessibility are important for PCA site determination. Figure 6B showed that solvent accessibility features at site 1, 2, 4, 5, 7 3 and 16 influence more on PCA site determination.

Features analysis of propensity of amino acid to be conserved at protein-protein interface and protein surface

There were 13 features of propensity of amino acid to be conserved at protein-protein interface and protein surface in the optimal feature set. As shown in Figure 7A, propensity of amino acid to be conserved at protein-protein interface influence more on PCA modification site determination and propensity of amino acid to be conserved at protein surface also influence PCA modification site determination.

As shown in Figure 7B, the Propensity of amino acid to be conserved at protein-protein interface and protein surface features at site 15, 2–10, 14 and 17 influence more on PCA modification site determination.

Features analysis of gain/loss of amino acids during evolution

There were 5 features of gain/loss of amino acids during evolution in the optimal feature set, which located at site 13, 10, 6, 18 and 20.

Feature analysis of deviation of side chain carbon atom number

There were 5 features of deviation of side chain carbon atom number in the optimal feature set which located at site 8, 9, 13, 10 and 14, indicating that these sites may influence more on PCA modification site determination than other sites.

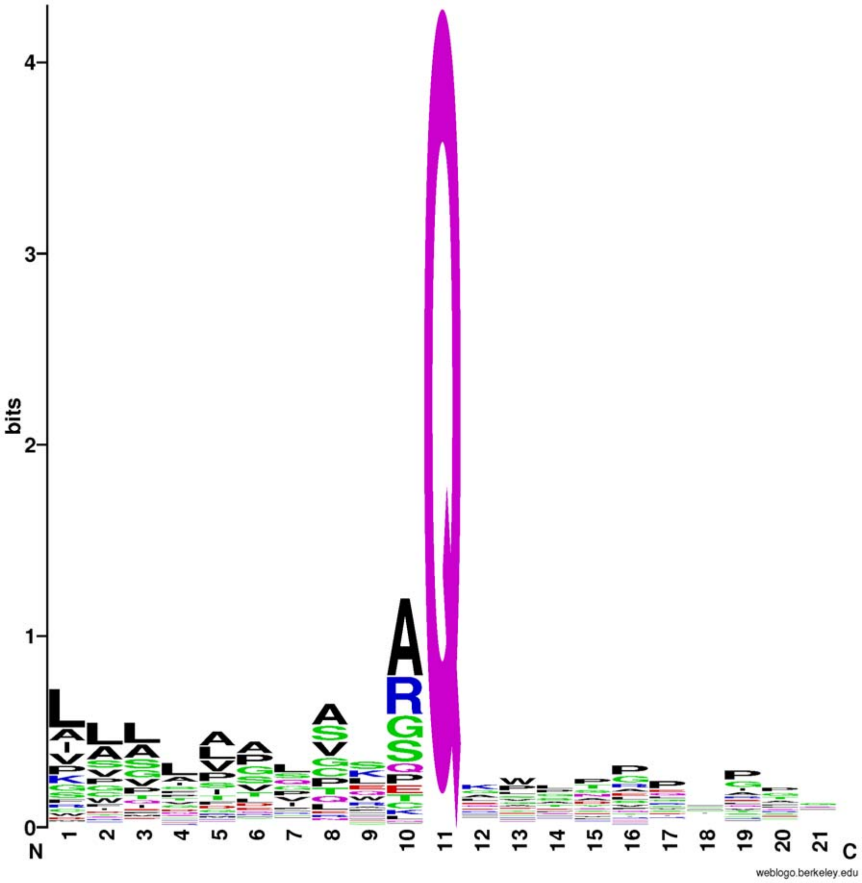

Figure 8. The sequence logo of 370 21-residue peptides [49]. In the upstream of Q site, sites 1, 2, 3, 5, 8 and 10 show some degree of conservation.
doi:10.1371/journal.pone.0028221.g008

3.4 Directions for experimental validation

We investigated the top 20 features within the optimal feature set. Among which there are 14 features of PSSM conservation, 3 features of AAFactor, 2 features of secondary structures and 1 feature of deviation of side chain carbon atom number. The top 20 features indicate that conservation may play the main role in PCA modification site prediction.

Discussion

Due to biological importance, PTM sites and regions surrounding the sites often display some degree of conservation, such as phosphosites and their −5 and +5 residues [44]. Evolutionary analysis has also showed evidences of PTM positions' higher purifying selection in 70% of the phosphorylated proteins [45]. Obviously, PCA sites have the similar conservative property. Even they are so extremely conservative that nearly all of the sites are restricted to Glutamine (Q), implying the significance of Q in PCA modification. Mutations to these PTM sites or surrounding regions may cause diseases [46]. However, as Figure 8 shows that, there is no remarkable conservative pattern found in the downstream of the PCA modification site Q. In contrast, in the upstream of the Q site, there is some evidence to exemplify the conservation in the surrounding region. For example, sites 1, 2, 3, 5, 8 and 10 are more likely to be amino acids L, L, L, A and A . This is also to some extent consistent with our previous results. Of the optimized 224 features, based on which we could obtain the best prediction accuracy, more than a half are features of PSSM conservation score, indirectly demonstrating there is some conservation around PCA sites. In particular, for the prediction, it is not equally important for all positions to influence the accuracy. Our results displayed that site 10 played the most important role in the prediction, and then sites 1, 2, 3 and 5 also played more important role than the other remaining sites.

In order to improve the prediction accuracy of PCA modification sites, a wide range of different types and of sources of information have been combined. These include biophysical information like amino acid physicochemical and biochemical properties, gain/loss of amino acid during evolution, deviation of side chain carbon atom number, or structural information like secondary structure, solvent accessibility and disorder score. For example, it has been noted that many PTM sites have a relative tendency to occur in regions lacking secondary structure, such as regions of intrinsic disorder [47,48]. One reason is that increased flexibility of intrinsic disorder regions allow them to fold, and therefore the amino acid side chains would fit into a modifying enzyme's catalytic site easily. Our results revealed that this feature of disorder score is not very important for the prediction, because we could find from Table 1, features of secondary structure played more important role. Nevertheless, Figure 5A indicated that three types of secondary structure features (helix, strand and other) influence the determination of PCA modification site nearly equally, a little inconsistent with others'. dbPTM and Pang et al. have both shown that PCA sites are more likely to be within coiled regions [17,48]. This may be because our peptides are 21-residues long, much longer than theirs. In the article, Pang et al. have also manifested PCA modifications did not show any strong preferences for surface accessibility, completely compatible with ours. From Figure 6A, buried and exposed have the same influence for the prediction, suggesting this feature we selected was used correctly.

At present, our method is limited to PCA modification occurring in the internal region of the protein sequences. One pivotal reason is that, most (222/399) of the protein sequences containing PCA modified Q sites at the N-terminus have less than 21 residues. Consequently, for those at the N-terminal of sequences, it is necessary to establish a novel method, to predict the PCA modification which occurs at the amino-terminus of proteins or both by integrating our model.

In this study, we developed a method for the prediction of PCA modification sites using totally 727 features. Our method achieved an overall MCC of 78.12% using the optimal feature set (244 features). Further detailed feature analysis may provide clues for understanding the PCA modification mechanism. The selected optimal feature set, especially the top features may provide important clues for further experimental researches in this area.

Supporting Information

Dataset S1 Training dataset used in the study.
(XLS)

Table S1 IFS results.
(XLS)

Table S2 Optimal feature set.
(XLS)

Author Contributions

Conceived and designed the experiments: LLZ SN YL YDC. Performed the experiments: LLZ SN. Analyzed the data: SN YDC. Contributed reagents/materials/analysis tools: LLZ SN YDC. Wrote the paper: LLZ SN KF PH. Gave suggestions for revision of the manuscript: PH YL YDC.

References

1. Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21: 255–261.
2. Cohen P (2000) The regulation of protein function by multisite phosphorylation–a 25 year update. Trends Biochem Sci 25: 596–601.
3. Awade AC, Cleuziat P, Gonzales T, Robertbaudouy J (1994) Pyrrolidone Carboxyl Peptidase (Pcp) - an Enzyme That Removes Pyroglutamic Acid (Pglu) from Pglu-Peptides and Pglu-Proteins. Proteins-Structure Function and Bioinformatics 20: 34–51.
4. Dimarchi RD, Tam JP, Kent SBH, Merrifield RB (1982) Weak Acid-Catalyzed Pyrrolidone Carboxylic-Acid Formation from Glutamine during Solid-Phase Peptide-Synthesis - Minimization by Rapid Coupling. International Journal of Peptide and Protein Research 19: 88–93.
5. Abraham GN, Podell DN (1981) Pyroglutamic Acid - Non-Metabolic Formation, Function in Proteins and Peptides, and Characteristics of the Enzymes Effecting Its Removal. Molecular and Cellular Biochemistry 38: 181–190.
6. Cummins PM, O'Connor B (1998) Pyroglutamyl peptidase: an overview of the three known enzymatic forms. Biochim Biophys Acta 1429: 1–17.
7. Fernandez Garcia A, Butz P, Trierweiler B, Zoller H, Starke J, et al. (2003) Pressure/temperature combined treatments of precursors yield hormone-like peptides with pyroglutamate at the N terminus. J Agric Food Chem 51: 8093–8097.
8. Hinkle PM, Tashjian AH, Jr. (1973) Receptors for thyrotropin-releasing hormone in prolactin producing rat pituitary cells in culture. J Biol Chem 248: 6180–6186.
9. Dannies PS, Tashjian AR, Jr. (1973) Effects of thyrotropin-releasing hormone and hydrocortisone on synthesis and degradation of prolactin in a rat pituitary cell strain. J Biol Chem 248: 6174–6179.
10. Pellegrini-Giampietro DE, Moroni F, Pistelli A, Palmerani B, Zorn AM, et al. (1989) Pyrrolidone carboxylic acid in acute and chronic alcoholism. Preclinical and clinical studies. Recenti Prog Med 80: 160–164.
11. Silva AR, Silva CG, Ruschel C, Helegda C, Wyse AT, et al. (2001) L-pyroglutamic acid inhibits energy production and lipid synthesis in cerebral cortex of young rats in vitro. Neurochem Res 26: 1277–1283.
12. Fenves AZ, Kirkpatrick HM, Patel VV, Sweetman L, Emmett M (2006) Increased anion gap metabolic acidosis as a result of 5-oxoproline (pyroglutamic acid): A role for acetaminophen. Clinical Journal of the American Society of Nephrology 1: 441–447.
13. Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, et al. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. Journal of Molecular Biology 289: 645–657.
14. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, et al. (2006) Dynamic spectrum quality assessment and iterative computational analysis of

shotgun proteomic data - Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Molecular & Cellular Proteomics 5: 652–670.

15. Mandal AK, Balaram P (2007) Mass spectrometric identification of pyroglutamic acid in peptides following selective hydrolysis. Anal Biochem 370: 118–120.
16. Koenig T, Menze BH, Kirchner M, Monigatti F, Parker KC, et al. (2008) Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. Journal of Proteome Research 7: 3708–3717.
17. Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, et al. (2006) dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res 34: D622–627.
18. Huang HD, Lee TY, Tzeng SW, Horng JT (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Research 33: W226–W229.
19. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, et al. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research 38: D142–D148.
20. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10: 136.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
22. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28: 374.
23. Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 102: 6395–6400.
24. Torkamani A, Schork NJ (2007) Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 23: 2918–2925.
25. Rubinstein ND, Mayrose I, Pupko T (2009) A machine-learning approach for predicting B-cell epitopes. Molecular Immunology 46: 840–847.
26. Marsella L, Sirocco F, Trovato A, Seno F, Tosatto SCE (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. Bioinformatics 25: i289–i295.
27. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of Molecular Biology 293: 321–331.
28. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41: 6573–6582.
29. Liu J, Tan H, Rost B (2002) Loopy proteins appear conserved in evolution. Journal of Molecular Biology 322: 53–64.
30. Tompa P (2002) Intrinsically unstructured proteins. Trends in Biochemical Sciences 27: 527–533.
31. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7: 208.
32. Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Research 33: W72–W76.
33. Goldschmidt L, Teng PK, Riek R, Eisenberg D (2010) Identifying the amylome, proteins capable of forming amyloid-like fibrils. Proceedings of the National Academy of Sciences of the United States of America 107: 3487–3492.
34. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, et al. (2005) A universal trend of amino acid gain and loss in protein evolution. Nature 433: 633–638.
35. Ma BY, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proceedings of the National Academy of Sciences of the United States of America 100: 5772–5777.
36. Popelier PL, Aicken FM (2003) Atomic properties of selected biomolecules: quantum topological atom types of carbon occurring in natural amino acids and derived molecules. J Am Chem Soc 125: 1284–1292.
37. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27: 1226–1238.
38. Qian Z, Cai YD, Li Y (2006) A novel computational method to predict transcription factor DNA binding preference. Biochem Biophys Res Commun 348: 1034–1037.
39. Huang T, Cui W, Hu L, Feng K, Li YX, et al. (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS ONE 4: e8126.
40. Liu MC, Yasuda S, Idell S (2007) Sulfation of nitrotyrosine: biochemistry and functional implications. IUBMB Life 59: 622–627.
41. Cai Y, He J, Li X, Lu L, Yang X, et al. (2009) A novel computational approach to predict transcription factor DNA binding preference. J Proteome Res 8: 999–1003.
42. Huang T, Tu K, Shyr Y, Wei CC, Xie L, et al. (2008) The prediction of interferon treatment effects based on time series microarray gene expression profiles. J Transl Med 6: 44.
43. Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 5: e10972.
44. Gnad F, Ren S, Cox J, Olsen JV, Macek B, et al. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 8: R250.
45. Gray VE, Kumar S (2011) Rampant purifying selection conserves positions with posttranslational modifications in human proteins. Mol Biol Evol 28: 1565–1568.
46. Elemans CPH, Mead AF, Jakobsen L, Ratcliffe JM (2011) Superfast Muscles Set Maximum Call Rate in Echolocating Bats. pp 1885–1888.
47. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, et al. (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res 6: 1917–1932.
48. Pang CN, Hayen A, Wilkins MR (2007) Surface accessibility of protein post-translational modifications. J Proteome Res 6: 1833–1845.
49. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.