Detecting Human Phosphorylated Protein by Using Class Imbalance Learning and Ensemble Classifier

Xuan XIAO*, Shun-lu LIAO and Wang-ren QIU

Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403 China

*Corresponding author

Keywords: Wavelets transforms, Pseudo amino acid composition, Random forests, Protein phosphorylation, Ensemble classifier.

Abstract. Protein phosphorylation plays a critical role by altering the structural conformation of a protein, causing it to become activated, deactivated, or modifying its function. Encouraged by Qiu’s pioneer work, this paper has developed a new ensemble classifier for detecting human protein phosphorylation. In the predictor, a protein sample is formulated by incorporating the stationary wavelet features derived from the numerical series of protein chain and two types of pseudo amino acid composition (PseAAC). The operation engine to run the predictor is an ensemble classifier formed by fusing nine individual random forest engines via a voting system. It is demonstrated with a larger dataset obtained from Uniprot web. The approach may also have notable impact on prediction of the other PTMs, such as ubiquitination, crotonylation, methylation, and succinylation, among many others.

Introduction

As one of the most-studied post-translational modification, protein phosphorylation can alters the structural conformation of a protein, causing it to become activated, deactivated, or modifying its function. As regard to human protein phosphorylation, it was reported that the phosphorylation is only one of the more than 500 protein kinases encoded within the human genome. Human protein phosphorylation is a critical control for the regulation of cell growth, differentiation and apoptosis, while its dysregulation is implicated in many diseases.

In a phosphorylated protein of human, an amino acid residue is phosphorylated by a protein kinase by the addition of a covalently bound phosphate group. The most commonly phosphorylated amino acids are serine, threonine and tyrosine in addition to arginine, lysine, and cysteine which easily be missed. The commonly method for detecting the protein phosphorylation is investigating the protein kinases which usually accompany with some certain phosphorylated amino acids. These methods include mass spectroscopy, phosphor-specific antibody, radioisotope labelling, dual-mode field-effect devices and nanoplasmonic sensors, the sensitive and selective electrochemical and optical detection methodologies, etc. However, these techniques usually are timely, costly or needing professional technology processes.

For a given human protein sequence, it is an interesting and significant problem to determine it can be phosphorylated or not? Qiu posted this question and tried to answer it with a computational model, his pioneer work is impacting other related researchers. However, there are two major gaps need to be filled. The first one is that the samples used in reference \[1\] are only 1770, a larger benchmark will be more easy to be accepted by the truth. The second one is that all of the predicted results are less than 75%, therefore, as a binary classification problem, there still are much room to improve it. Here, stimulated by the successes of using ensemble classifiers, we proposed an ensemble classifier formed by fusing nine individual random forest engines via a voting system. Below, let us address the procedures one-by-one.
Material and Methods

Benchmark Dataset

In UniProtKB, there are four options, i.e. any assertion method, any manual assertion, any automatic assertion and any experimental assertion, for searching the given human protein sequence(s). However, with “any automatic assertion” as the searching key words, the obtained proteins are only 47 samples, it’s a meaningless dataset for the prediction. This study thus constructed the benchmark datasets with “any assertion method” as the key words. Then the dataset means it was or was not annotated by phosphorylation by any assertion method. Since Phosphoserine, Phosphothreonine and Phosphotyrosine are the most of phosphorylation, the positive dataset of phosphorylation proteins are at least have one of the three kinds modification site. The benchmark datasets were derived from the Swiss-Prot database (version 2014_05).

To obtain a high quality benchmark dataset, the source phosphorylation proteins dataset were collected according to the following criteria: (1) each of the included proteins must contain at least 50 residues in order to avoid fragments \(^2\), and no more than 5000 residues for the convenience of PSSM; (2) the results in sequence clusters with identity of \(50\%\) in order to reduce the homology bias. From the 7,946 source proteins thus obtained, the dataset \(S\) was established, as formulated below

\[
S = S^+ \cup S^- \quad (1)
\]

where, \(S\) contains 6,632 proteins, \(S^+\) is the positive subset that contains 5,224 phosphorylated proteins, \(S^-\) is the negative subset containing 1,408 non-phosphorylated proteins, and \(\cup\) represents the union in the set theory. The profile of the dataset has listed in Table 1.

<table>
<thead>
<tr>
<th>Number(^a) of Data samples</th>
<th>7946</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number(^b) of Data samples</td>
<td>6632</td>
</tr>
<tr>
<td>Positive vs Negative</td>
<td>5224:1408</td>
</tr>
</tbody>
</table>

\(^a\) The human protein with phosphorylated residue(s).
\(^b\) The number of proteins which has \(\leq 50\%\) pairwise sequence identity to any other and the length of protein resides sequence ranges from 50 to 5000.

Balancing the Training Dataset

More and more evidences show that class imbalances have been reported to make great compromise on the performance of most standard classifiers, such as Naive Bayes, Support Vector Machine and Neural Networks. Stimulated by the successes of using evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, this study will present a new method for enhancing the imbalanced training data.

The protein sequence is composed of 20 different kinds of native amino acids. Let \(\overline{R} = \{\overline{R_1}, \ldots, \overline{R_{20}}\}\), and \(\overline{R_1}\) represents A, \(\overline{R_2}\) represents C, \(\ldots\), \(\overline{R_{20}}\) represents Y, respectively. Given a protein sequence \(P\) with \(L\) amino acid residues,

\[
P = R_1R_2R_3R_4R_5R_6\cdots R_i\cdots R_L
\]

where \(R_i\) represents the \(i^{th}\) amino acid residue of the protein \(P\) and \(R_i \in \{\overline{R_1}, \ldots, \overline{R_{20}}\}\).

Conservation evolutionary information is an important characteristic of protein for the reason that the conserved residues at specific sequence sites are under strong selective pressure. To obtain this information, protein amino acid sequence was used as a seed to search the homogenous sequences from the SWISSPROT protein database by using the PSIBLAST program with parameters \(h\) and \(j\) (which is 0.001 and 3 respectively). The version of the UniProtKB/Swiss-Prot database was 2010_04 of 23-Mar-2010. These aligned sequences are further converted into position-specific scoring matrices (PSSMs). According to \(^3\), the sequence evolution information of protein \(P\) can be expressed by a \(L \times 20\) matrix, as given by
\[ P_{PSSM} = \begin{bmatrix}
  m_{1,1} & \cdots & m_{1,20} \\
  \vdots & \ddots & \vdots \\
  m_{L,1} & \cdots & m_{L,20}
\end{bmatrix} \quad (3) \]

where \( m_{i,j} \) represents the original score of amino acid residue in the \( i \) (\( i = 1, 2, \ldots, L \)) sequential position of the protein \( P \) that is being changed to amino acid type \( j \) \( (j = 1, 2, \ldots, 20) \) during the evolution process. Here, the numerical codes 1, 2, \ldots, 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes.

Then the protein \( P \) may produce a new protein \( P' = R'_1 \cdots R'_j \cdots R'_L \) with \( R'_i \) may be replaced by \( R'_i \) with probability no more than \( \mu \). In this study, \( \mu \) equals to 5\%. I.e. \( R_i \leftarrow R'_i \) with probability \( \leq \mu \). Here \( R_i, R'_i \in \{R_1, \ldots, R_{20}\} \) and \( R'_i \) is the location of max-value of the \( i \)th row of \( P_{PSSM} \). After balancing the training dataset, the number of negative samples is the same as that of positive samples. And then, it is possible that the conventional algorithms based on balanced dataset can be used here.

**Sample Representation**

According to a recent review \([4]\), the general form of encoding PseAAC for a protein sequence fragment can be formulated as: \( P = [\Psi_1, \Psi_2, \ldots, \Psi_u, \ldots, \Psi_\Omega]^T \). Here, \( T \) is a transpose operator, while the subscript \( \Omega \) is an integer to reflect the vector’s dimension. The value of \( \Omega \) as well as the components \( \Psi_u (u = 1, 2, \ldots, \Omega) \) will depend on how to extract the desired information from the protein sequence. In this paper, the protein would be represented by three kinds of representations, i.e. Discrete Transform (DWT) \([5]\), Chou’s Pseudo Amino Acid Composition (PAAC) and PAAC of Complexity measure factor based on cellular automata image (PCAI).

**Discrete Wavelet Transform (DWT)**

As a multi-resolution analysis tool, Wavelet Transform (WT) is quite popular for analyzing, de-noising and compressing signals and images. This approach can overcome the shortcoming of Fourier analysis, which is based on the functions that are localized in frequency domain but not in time domain. To take advantage of the WT, this study considered representing the protein sequence with numerical series extracted from its physicochemical properties. The considered seven physicochemical properties are (1) hydrophobicity, (2) hydrophilicity, (3) side-chain volume, (4) polarity, (5) polarizability, (6) solvent-accessible surface area (SASA) and (7) side-chain net charge index (NCI). In the same way of \([6]\), with the given physicochemical properties, the protein \( P \) then can be represented by a 20 dimension vector.

**Pseudo Amino Acid Composition of Chou (PAAC)**

In developing a statistical method for predicting the attribute of peptides in proteins, the most important procedures was to formulate the peptide samples with an effective mathematical expression that could truly reflect the intrinsic correlation with the desired target. To realize this, various feature vectors were proposed to express peptides by extracting their different features into the pseudo amino acid composition (PseAAC) \([7]\), the method for PseAAC used here can be found in \([7]\), and the \( \lambda \) also is 5 in this study.

**Pseudo Amino Acid Composition of Complexity measure factor based on cellular automata image (PCAI)**

To address the representation of a given protein, especially for a protein with long sequence, Xiao \([8]\) had proposed a novel method by using complexity measure factor of cellular automata image. The representation contains the lost information of order effects for that the images are derived from the amino acid sequence through the space-time evolution of cellular automata. See \([8]\) to refer the detail process.

**Random Forest and Ensemble Classifier**

The random forests (RF) algorithm is a powerful algorithm and has been used in many areas of computational biology. The detailed procedures and formulation of RF have been described by Breiman in \([9]\), and hence we will not repeat it.
As shown in the represents of a given protein sequence, a protein can be formulated with different PseAAC forms, each of which can be used to train the RF predictor. Accordingly, we have a total of nine individual predictors for detecting human protein phosphorylation, as formulated by

\[ \text{Individual predictor} = \text{RF}(\xi) \ (\xi = 1, 2, \ldots, 9) \]  

where \( \text{RF}(\xi) \) represents the RF based on \( P_{\text{DWT}}^{(1)}, \ldots, P_{\text{DWT}}^{(7)}, P_{\text{PACC}} \) and \( P_{\text{PCA1}} \).

Now, the problem is how to combine the results from the nine individual predictors to maximize the prediction quality. Encouraged by the previous investigators’ studies, here we are also to develop an ensemble classifier by fusing the nine individual predictors \( \text{RF}(\xi) \ (\xi = 1, 2, \ldots, 9) \) through a voting system, as formulated by

\[ \text{RF}^E = \text{RF}(1) \lor \cdots \lor \text{RF}(9) = \lor_{k=1}^{9} \text{RF}(k) \]  

where \( \text{RF}^E \) represents the ensemble classifier, and the symbol \( \lor \) denotes the fusing operator. For the detailed procedures of how to fuse the results from the nine individual predictors to reach a final outcome via the voting system, see Eqs.30-35 in [10].

**Evaluation Metrics and Validation Method**

In literature, the following four metrics are often used for examining the performance quality of a predictor. They are (1) overall accuracy or Acc, (2) Mathew’s correlation coefficient or MCC, (3) sensitivity or Sn, and (4) specificity or Sp. We used them here also.

In statistical prediction process, the following three cross-validation methods are often used to derive the metrics values for predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test. Of these three methods, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in [4]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors. However, to reduce the computational time, in this study we adopted the 5-fold cross-validation, as done by most investigators with random forests algorithm as the prediction engine.

**Results and Discussion**

The proposed predictor was tested by the benchmark dataset in Eq.1, which contains 5,224 phosphorylated proteins and 1,408 non-phosphorylated proteins listed in Table 1. Listed in Table 2 are the predicted metrics rates via the 5-fold cross-validation.

<table>
<thead>
<tr>
<th>Balance method</th>
<th>Original method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc</td>
<td>Mcc</td>
</tr>
<tr>
<td>RF(1)</td>
<td>77.20</td>
</tr>
<tr>
<td>RF(2)</td>
<td>78.31</td>
</tr>
<tr>
<td>RF(3)</td>
<td>75.19</td>
</tr>
<tr>
<td>RF(4)</td>
<td>79.07</td>
</tr>
<tr>
<td>RF(5)</td>
<td>74.36</td>
</tr>
<tr>
<td>RF(6)</td>
<td>74.21</td>
</tr>
<tr>
<td>RF(7)</td>
<td>76.62</td>
</tr>
<tr>
<td>RF(8)</td>
<td>79.07</td>
</tr>
<tr>
<td>RF(9)</td>
<td>78.80</td>
</tr>
<tr>
<td>RF(^E)</td>
<td>80.58</td>
</tr>
</tbody>
</table>

For facilitating comparison, listed in table 2 are also the corresponding rates obtained from the same dataset without balancing technique. It can be clearly seen from Table 2, the proposed predictor is very effective and can achieve good predictions. As we can see from the Table 2, the Acc, Mcc, Sn and Sp achieved by balancing predictor was 80.58%, 0.5887, 92.18% and 77.45%. Meanwhile, those
metrics rates obtained on unbalanced dataset was 79.95%, 0.2139, 11.39% and 98.43%. The balanced predictor is superior to un-balanced one in three of the four metrics as defined. Especially on the MCC, which means the proposed model can greatly detecting the non- phosphorylated proteins. With these results, researchers may saving much time to do other experiments for protein phosphorylation.

Why the proposed method is so powerfully in detecting human phosphorylated proteins? The reason maybe is because that there are much important information can be refined from many key features including the amino acids occurrence frequencies, the complexity measure factors, the correlation factor of its physicochemical properties. Just like in dealing with the extremely complicated internal motions of proteins, it is the key to grasp the low-frequency collective motion in-depth understanding or revealing the dynamic mechanisms of their various important biological functions, such as cooperative effects, allosteric transition, assembly of microtubules, and switch between active and inactive states.

Conclusion

In the proposed predictor, a query protein is formulated into three types represents: PseAAC, PCAI and Discrete Wavelet Transform. PseAAC and PCAI are used in most of above-mentioned works. The last represent is formulated by a general form of PseAAC, and the components are defined via the following procedures: (1) a protein sequence is converted into a numerical series via the physicochemical properties of amino acids; (2) the numerical series is subsequently converted into a 20 dimensional feature vector by means of the DWT technique.

The operation engine to run the prediction is an ensemble classifier formed via a voting system to fuse nine different random forest classifiers based on three types represents which comprising of amino acids occurrence frequencies, the complexity measure factors and the correlation factor of its physicochemical properties. Furthermore, the rigorous cross-validations have indicated that the new predictor established with the above procedures is very powerful and promising.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (No. 31560316, 61261027, 61262038, 31260273, 61202313), the Natural Science Foundation of Jiangxi Province, China (No. 20132BAB201053), the Scientific Research plan of the Department of Education of JiangXi Province(GJJ14640), The Young Teacher Development Plan of Visiting Scholars Program in the University of Jiangxi Province. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References


