Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201600085

Multi-iPPseEvo: A Multi-label Classifier for Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into Chou's General PseAAC via Grey System Theory

Wang-Ren Qiu,*[a, b, c] Quan-Shu Zheng,[a] Bi-Qian Sun,[a] and Xuan Xiao*[a, d]

Abstract: Predicting phosphorylation protein is a challenging problem, particularly when query proteins have multi-label features meaning that they may be phosphorylated at two or more different type amino acids. In fact, human protein usually be phosphorylated at serine, threonine and tyrosine. By introducing the "multi-label learning" approach, a novel predictor has been developed that can be used to deal with the systems containing both single- and multi-label phosphorylation protein. Here we proposed a predictor called Multi-iPPseEvo by (1) incorporating the protein sequence evolutionary information into the general pseudo amino acid composition (PseAAC) via the grey system theory, (2) balancing out the skewed training datasets by the asymmetric bootstrap approach, and (3) constructing an ensemble predictor by fusing an array of individual random forest classifiers thru a voting system. Rigorous cross-validations via a set of multi-label metrics indicate that the multi-label phosphorylation predictor is very promising and encouraging. The current approach represents a new strategy to deal with the multi-label biological problems, and the software is freely available for academic use at http://www.jci-bioinfo.cn/Multi-iPPseEvo.

Keywords: Multi-label learning · Random Forests · Protein phosphorylation · Ensemble classifier

1 Introduction

As a ubiquitous post-translational modification (PTM), protein phosphorylation controls a number of intracellular processes. Due to the importance of PTM to basic research and drug development, identification of PTM sites in proteins has become a very hot topic in bioinformatics.[1–6] It was estimated that at least one-third of the cellular proteins are modified by phosphorylation.[7] Actually, human phosphorylation also plays a critical role in the transmission of signals controlling a diverse array of cellular functions including cell growth, survival differentiation and metabolism.[8] Human protein phosphorylation thus is a critical control for the regulation of cell growth, differentiation and apoptosis,[9] and its dysregulation is implicated in many diseases.[10] The most commonly phosphorylated amino acids are serine, threonine and tyrosine in addition to arginine, lysine and cysteine.[11]

It is a critical work for researchers investigating the effective and fast method for detecting phosphorylation proteins. Most of phosphorylation sites are usually experimentally determined by mass spectrometry-based techniques.[12–13] The commonly method for detecting the protein is investigating the protein kinases which usually accompany with some certain phosphorylated amino acids. However, it is not a good method because it is hard to find the key kinases.[14] Furthermore, most of methods attempted the detection of protein phosphorylation by measuring the addition of negative charges changes of the protein after phosphorylation[8] or the release of protons in the reaction buffer upon phosphorylation of protein.[15] These methods noted as mass spectroscopy, phosphor-specific antibody,[16] radioisotope labelling,[17] the sensitive and selective electrochemical[18] and optical detection methodologies,[19] etc.

During the exploration of phosphorylation proteins, there is a great gap between what it really is and what has been known because of two causes. The first one is that

[a] *W.-R. Qiu, Q.-S. Zheng, B.-Q. Sun, X. Xiao*
Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403 China
**e-mail: qiuone@163.com*
zhengquanshu@163.com
Sunbiq@126.com
xxiao@gordonlifescience.org

[b] *W.-R. Qiu*
Department of Computer Science, University of Missouri, Columbia, MO, USA

[c] *W.-R. Qiu*
Bond Life Science Center, University of Missouri, Columbia, MO, USA

[d] *X. Xiao*
Gordon Life Science Institute, Boston, Massachusetts 02478, United States of America

most of techniques cannot be applied widely because they are time consuming, laborious, cost inefficient, or require the usage of chemical reagents. In the process of mass spectrometry, large investments and expertise are required.[20] The process of using phosphor-specific antibodies is costly and relies on the development of reliable target-specific antibodies.[21] The second one is that there is few effective predictor for detecting phosphorylation. Although we have developed a predictor for identifying human phosphorylated proteins,[22] to the best of our knowledge, no predictor can handle the problem. But these types of proteins are also very important because they may have some special biological significances. Thus, it is in the urgent need that the development of technologies enables efficient and convenient analyses of protein phosphorylation.

Considering the fact that several predictors which can deal with both single and multiple type problems have been established, much work should made in this area. For example, there are many computational methods which were proposed for predicting protein subcellular localization,[23–24] distinguishing functional genomics and text categorization;[25] sentiment classification;[26] classifying colon cancer;[27] recognizing protein function.[28–29] Encouraged by the success of multi-Label classification[30–41] and ensemble teachnique,[42–43] the present study developed a phosphorylation protein predictor, called Multi-iPPseEvo, specialized for human proteins by improving the aforementioned shortcomings. In this study, the benchmark datasets were derived from the Swiss-Prot database. The set of human phosphorylation proteins comprising of phosphoserine, phosphothreonine and phosphotyrosine proteins. A protein sample is formulated by incorporating evolutionary information into Chou's general PseAAC via grey system theory.

2 Materials and Methods

2.1 Benchmark Dataset

The protein data set was taken from the UniprotKB/Swiss-Prot database at http://www.ebi.ac.uk/uniprot/ released in September 2015 (version 2014_05). The detailed procedures are as follows: (1) open the Web site at http://www.uniprot.org/; (2) click the button "Advanced", select "PTM/Processing" and "Modified residue [FT]" for "Fields", type in "Phosphoserine" (or "Phosphothreonine", "Phosphotyrosine") for "Term", and select "Any experimental assertion" for "Evidence"; (3) click the button "Add & Search", and repeat step 2 with different Terms. (4) Only those sequences annotated with "human" in the ID field were collected because the current study was focused on human proteins only, so the number is reduced to 2,076. Sequences annotated with "fragment" were excluded; sequences with less than 50 and more than 5000 amino acid residues were removed for the convenience of constructing Position Specific Scoring Matrix (PSSM). (5) To reduce the redundancy and homology bias, the program CD-HIT was utilized to remove those proteins that had $>50\%$ pairwise sequence identity to any other in the same subset. After strictly following the aforementioned procedures, we finally obtained the dataset denoted as $\mathbb{S}$.

The protein samples considered in this study can be expressed as

$$\mathbf{P} = R_1R_2R_3R_4R_5R_6R_7 \ldots R_L \rhd \{\text{S; T; or Y}\} \quad (1)$$

where R_1 represents the 1st residue of the protein P, R_2 the 2nd residue, R_3 the 3rd residue, and so forth; the symbol $\rhd$ means that, of the L amino acid residue, at least one must be S (Ser), or T (Thr), or Y (Tyr). In other words, there is at least one $i \in \{1, 2, \ldots L\}$ subject to $R_i \in \{S, T, Y\}$.

After refined the data with the five steps, we obtained 1,507 human proteins, of which 845 occur in phosphoserine phosphorylation subset, noted as S_1, 386 in phosphothreonine phosphorylation subset, noted as S_2, 249 in phosphotyrosine phosphorylation subset, noted as S_3, and 375 in non-phosphorylation subset, noted as S_4. The proteins benchmark dataset $\mathbb{S}$ covers 4 different subsets, and can be formulated as Eq. 2. The profile of the dataset was listed in Table 1 and 2. See Online Supporting Information for the details of these proteins and their ID information.

Table 1. The profile of the dataset (1)

Subset	Number of Samples
S_1	845
S_2	386
S_3	249
S_4	375
Total different locative proteins	1855

Table 2. The profile of the dataset (2)

Number of phosphorylation	Number of Samples
One	820
Two	276
Three	36
zero	375
Total different proteins	1507

$$\mathbb{S} = S_1 \cup S_2 \cup S_3 \cup S_4 \quad (2)$$

2.2 Protein Sample Formulation with General PseAAC

Since all the existing machine-learning algorithms, such as Neural Network, Support Vector Machine, K nearest Neighbor, Random Forest, can only handle vector but not sequence samples, as elaborated in the review,[1] it is very im-

portant for researchers to formulate a biological sequence with a discrete model or a vector, and make the represent(s) considerably keep the sequence pattern or inherent characteristic. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To overcome such a problem, the pseudo amino acid composition or PseAAC was proposed. Ever since the concept of pseudo amino acid composition[44–46] or Chou's PseAAC[47–48] was proposed, it has penetrated into nearly all the areas of computational proteomics (see, e.g.,[49–54]) as well as a long list of references cited in [8,28]). Because it has been widely and increasingly used, recently several powerful open access softwares (see, e.g.,[47,55]) were established for extracting various Chou's general PseAAC, including various higher level feature vectors such as "Functional Domain" mode, "Gene Ontology" mode, and "Sequential Evolution" or "PSSM" mode. Encouraged by the successes of using Chou's PseAAC to deal with protein/peptide sequences, its concept was extended to generate various feature vectors for DNA/RNA sequences as well. In addition, a very powerful web-server called Pse-in-One[56] was established by which users can generate most of existing feature vectors for DNA/RNA and protein/peptide sequences (see a long list of references cited in[57]), or some personalized feature vectors proposed by themselves. These works are very helpful for the development of computational biology.

According to the idea of Chou's general PseAAC, a protein **P** can be formulated as

$$\mathbf{P} = [\Psi_1\ \Psi_2\ \cdots\ \Psi_u\ \cdots\ \Psi_\Omega]^{\mathrm{T}} \tag{3}$$

where T is the transpose operator, the subscript Ω is an integer, and its value as well as the components Ψ_1, Ψ_2, ..., will depend on how to extract the desired information from the amino acid sequence of **P** (cf. Eq. 1). Below, let us describe how to extract the core and essential features from a protein sequence to define the components in Eq. 1.

From the viewpoint of historic dimension, all the protein sequences have developed beginning from a very limited number of ancestral species. Their evolution involves changes of single residues, insertions and deletions of several residues, gene doubling, and gene fusion. With these changes accumulated for a long period of time, many similarities between original and evoluted amino acid sequences have been gradually disappeared, but they may still share many common attributes, such as belonging to the same type of protein, residing in a same subcellular location, or having basically the same biological function. To extract the sequential evolutionary information and incorporate it into the definition of the components in Eq. 3, one of the feasible approaches is via the PSSM, as described below.

According to the reference[58], the sequence evolution information of protein **P** with L amino acid residues can be expressed by a $L\times 20$ matrix, as given by

$$\mathbf{P}^{(0)}_{\mathrm{PSSM}} = \begin{bmatrix} m^{(0)}_{1,1} & m^{(0)}_{1,2} & \cdots & m^{(0)}_{1,20} \\ m^{(0)}_{2,1} & m^{(0)}_{2,2} & \cdots & m^{(0)}_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ m^{(0)}_{\mathrm{L},1} & m^{(0)}_{\mathrm{L},2} & \cdots & m^{(0)}_{\mathrm{L},20} \end{bmatrix} \tag{4}$$

where $m^{(0)}_{i,j}$ represents the original score of amino acid residue in the i-th ($i=1,2,\ldots,L$) sequential position of the protein that is being changed to amino acid type j ($j=1,2,\ldots,20$) during the evolution process. Here, the numerical codes 1, 2, ..., 20 are used to denote the 20 native amino acids according to the alphabetical order of their single character codes. The $L\times 20$ scores in Eq. 4 were generated by using PSI-BLAST to search the UniProtKB/Swiss-Prot database through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of **P**. To make every element in Eq. 4 within the range of 0–1, a conversion was performed through the standard sigmoid function.

$$\mathbf{P}^{(1)}_{\mathrm{PSSM}} = \begin{bmatrix} m^{(1)}_{1,1} & m^{(1)}_{1,2} & \cdots & m^{(1)}_{1,20} \\ m^{(1)}_{2,1} & m^{(1)}_{2,2} & \cdots & m^{(1)}_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ m^{(1)}_{\mathrm{L},1} & m^{(1)}_{\mathrm{L},2} & \cdots & m^{(1)}_{\mathrm{L},20} \end{bmatrix} \tag{5}$$

where

$$m^{(1)}_{i,j} = \frac{1}{1+e^{-m^{(0)}_{i,j}}}. \tag{6}$$

Now, we can use the grey system theory to extract useful information from Eq. 6 to define the components of Eq. 3. According to,[59–62] we can extract the following information from the j-th column of Eq. 5

$$\left[a^j_1, a^j_2, b^j\right]^T = \left(\mathrm{B}^T_j \mathrm{B}_j\right)^{-1} \mathrm{B}^T_j \mathrm{U}_j \qquad (j = 1, 2, \cdots 20) \tag{7}$$

where

$$\mathbf{B}_j = \begin{bmatrix} -m^{(1)}_{2,j} & -m^{(1)}_{1,j} & -0.5m^{(1)}_{2,j} & 1 \\ -m^{(1)}_{3,j} & -\sum_{i=1}^{2} m^{(1)}_{2,j} & -0.5m^{(1)}_{3,j} & 1 \\ \vdots & & \vdots & \vdots \\ -m^{(1)}_{k,j} & \sum_{i=1}^{k-1} m^{(1)}_{i,j} & -0.5m^{(1)}_{k,j} & 1 \\ \vdots & & \vdots & \vdots \\ -m^{(1)}_{L,j} & -\sum_{i=1}^{L-1} m^{(1)}_{i,j} & -0.5m^{(1)}_{L,j} & 1 \end{bmatrix} \tag{8}$$

and

$$\mathbf{U}_j = \left[m_{2,j}^{(1)} - m_{1,j}^{(1)}, \quad \cdots, m_{k,j}^{(1)} - m_{k-1,j}^{(1)}, \quad \cdots, m_{L,j}^{(1)} - m_{L-1,j}^{(1)}\right]^T. \tag{9}$$

Therefore, when using the grey model approach to extract the protein sequence evolution information via the PSSM of Eq. 5, we can extract a total of $\Omega = 3 \times 20 = 60$ quantities. Thus, Eq. 3 can be quantitatively converted to

$$\mathbf{P}_{\text{EVO-Grey}} = \left[\Psi_1^{\text{EG}} \ \Psi_2^{\text{EG}} \ \cdots \ \Psi_u^{\text{EG}} \ \cdots \ \Psi_{60}^{\text{EG}}\right]^T \tag{10}$$

where

$$\begin{cases} \Psi_{3j-2}^{\text{EG}} = a_1^j f_j w_1 \\ \Psi_{3j-1}^{\text{EG}} = a_2^j f_j w_2 \\ \Psi_{3j}^{\text{EG}} = b^j f_j w_1 \quad (j = 1, 2, \cdots, 20) \end{cases} \tag{11}$$

where f'_j is the occurrence frequency of the j-th amino acid in the protein concerned, and w_1, w_2, and w_3 are the weight factors, which were all set to 1 in the current study.

2.3 Ensemble Random Forest Operation Engine

As we can see from Fig. 1, the predictor was combined by three sub-predictors, named as $\mathbb{RF}_1^E$, $\mathbb{RF}_2^E$ and $\mathbb{RF}_3^E$, which are constructed for identifying phosphoserine, phosphothreonine and phosphotyrosine respectively. At these three sub-steps, the number of samples in positive dataset $\mathbb{S}_1^+$ is 845, the number of samples in negative dataset $\mathbb{S}_1^-$ is 662; the number of samples in positive dataset $\mathbb{S}_2^+$ is 386, the number of samples in the negative dataset $\mathbb{S}_2^-$ is 1121; and the number of samples in positive dataset $\mathbb{S}_3^+$ is 249, the number of samples in the negative dataset $\mathbb{S}_3^+$ is 1258. In the last two data, the number of negative samples are much larger than the number of positive samples, i.e., $N^- \gg N^+$. Although this might reflect the real situation about protein phosphorylation, a predictor trained by such a highly skewed benchmark dataset would inevitably have the bias consequence that many non-phosphorylation proteins might be mispredicted as phosphorylation ones. Therefore, it is important to find an effective approach to minimize this kind of bias consequence. Here we use the ensemble random forest approach to deal with this problem.

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 very clearly described in,[63] and hence there is no need to repeat here. Most machine-learning classification algorithms (including RF) work properly if they are trained by a balanced benchmark dataset. If they are trained by a skewed dataset like the current one, the outcome might be problematic. To overcome such a problem, here we resort to the asymmetric bootstrap approach. The concrete procedures are listed as follows.

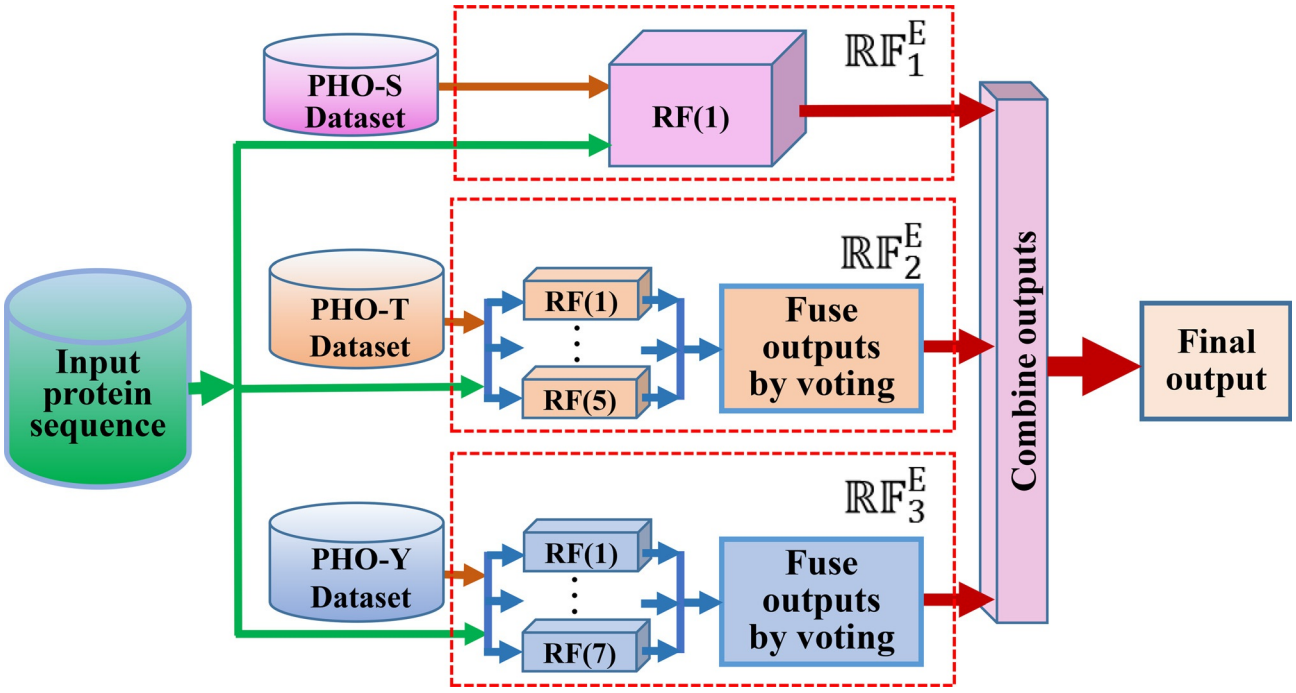

Figure 1. A flowchart to show the prediction process. PHO-S, PHO-T and PHO-Y represent phosphoserine, phosphothreonine and phosphotyrosine, respectively. See the text for further explanation.

First of all, to make the working benchmark dataset become a balanced one by using classifier 2 and classifier 3, we randomly extract N^- samples from the negative subset. The working positive subset thus obtained is denoted by $\mathbb{S}_1^-()$, whose size is exactly the same as $\mathbb{S}_l^+$ We repeated the above procedure for m_j times, generating an array of positive working subsets $\mathbb{S}_l^-(k)$ ($k=1,2,\cdots,m_l$. Accordingly, we also have an array of working benchmark datasets denoted by

$$\mathbb{S}_l(k) = \mathbb{S}_l^+ \cup \mathbb{S}_l^-(k) \quad (k = 1, 2, \cdots, \quad m_l; l \in \{2, 3\}). \qquad (12)$$

With any of the above m_l working benchmark datasets, we can establish an individual random forest classifier denoted by $\mathbb{RIF}_i(k)$. Stimulated by the reports that using the ensemble classifier formed by fusing many individual classifiers can remarkably enhance the success rates in predicting protein fold pattern,[64–66] predicting signal peptides,[67] protein subcellular localization,[38, 68–69] membrane proteins and their types,[68,70–71] in this study we are also to develop an ensemble classifier by fusing the m_i individual predictors $\mathbb{RIF}_i(k)$ through a voting system, as formulated by

$$\begin{aligned} \mathbb{RIF}_i^E &= \mathbb{RIF}_i(1) \forall \mathbb{RIF}_i(2) \forall \cdots \forall \mathbb{RIF}_i(m_l) \\ &= \forall_{k=1}^{m_l} \mathbb{RIF}_i(k) \quad (i \in \{2, 3\}) \end{aligned} \qquad (13)$$

where $\mathbb{RIF}_i^E$ represents the ensemble classifier I, and the symbol $\forall$ denotes the fusing operator. For the detailed procedures of how to fuse the results from the m individual predictors to reach a final decision via the voting system, see Eqs. 30–35 in,[69] where a crystal clear and elegant derivation was elaborated and hence there is no need to repeat here. A flowchart given in reference,[22] the Figure 1, can be used to illustrate how the m_l individual random forest predictors are fused into the ensemble classifier. In the current study, the number of bagging times was set at 5 and 7, i.e., $m_2=5$, $m_3=7$. Also, 50 trees were used for each of the individual predictors.

2.4 Evaluation Metrics and Validation Method

To evaluate the performance of the classfication, the following nine popular multilabel evaluation metrics[25,33,72–74] are used: (1) Hamming Loss; (2) One-Error; (3) Ranking Loss; (4) Coverage; (5) Subset-Accuracy; (6) Accuracy; (7) Average Precision; (8) Precision and (9) Recall. In reference,[34] Chou rewrote some of popular multi-label evaluation metrics,[25,73] and the improved metrics are more intuitive and easier to be understood for most biologists. However, there are only 5 enhanced metrics, i.e. Aiming, Coverage, Accuracy, Absolute-True and Absolute-False, which corresponding to Precision, Recall, Accuracy, Subset-Accuracy and Hamming-loss, respectively. To make in-deep evaluation of this novel research, we selected all of the nine popular metrics.

In all, the first three evaluation metrics are often used as measures for errors resulted from the multi-label classifier. The fourth metric measure the performance of a system for all the possible labels of samples. It can evaluate the number of steps needed, on the average, to move down the label list in order to cover all the proper labels attached to an instance. The last five ones can be used to evaluate the match level of the predicted results and the actual set of labels. Note that, for the first four metrics, the smaller the better, for the last five ones, the larger the better the performance. As shown in the Table 3, ↓ represents the smaller the better and ↑ represents the larger the better.

Table 3. The metrics obtained on the Dataset

Metrics	results
Hamming loss↓	0.2520
One-Error↓	0.2363
Coverage↓	1.2770
Ranking-Loss↓	0.4081
Subset-Accuracy↑	0.4056
Accuracy↑	0.5761
Average-Precision↑	0.6976
Precision↑	0.5998
Recall↑	0.7335

In statistical prediction, 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.[75] Of the three methods, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in[76] and demonstrated by Eqs. 28–32 therein. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors.[43,77–79] However, to reduce the computational time, as done by most investigators with random forests algorithm, the 5-fold cross-validation was applied to evaluate the prediction engine in this study.

2.5 Web-Server and User Guide

For the convenience of the vast majority of biological scientists, a web server for the predictor has been established at http://www.jci-bioinfo.cn/Multi-iPPseEvo. Here, let us provide a step by-step guide on how to use the web server.

Step 1. Opening the web-server at http://www.jci-bioinfo.cn/Multi-iPPseEvo, you will see the top page of Multi-iPPseEvo on your computer screen, as shown in Fig. 2. Click on the Read Me button to see a brief introduction about the predictor.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of Fig. 2. The input sequence should be in the FASTA format. For the ex-

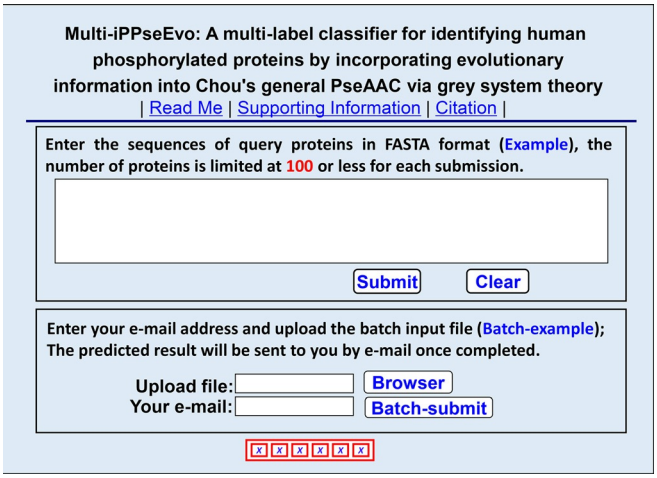

Figure 2. A semi-screenshot to show the top-page.

amples of sequences in FASTA format, click the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. After the results appeared on the screen, you will see the following shown on the screen of your computer (see Fig. 3): (1) Your Testing time is dd-mm-year HH:MM:SS;

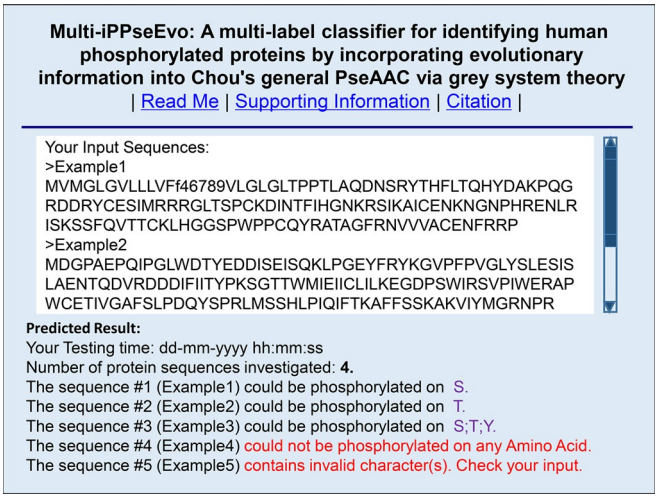

Figure 3. A semi-screenshot to show the output.

(2) The number of Protein sequences investigated is 4; (3) The detail predicted results:

The sequence #1 could be phosphorylated on Amina Acid(s) S .

The sequence #2 could be phosphorylated on Amina Acid(s) S; T; Y .

Since the fourth input contains number, it is illegal for a protein amino acid sequence, the fourth output is *"the Sequence #4 contains invalid character(s). Check your input. "*.

It may take about 30 seconds for each proposed sequence before the predicted results appear on the computer screen; the higher the number of query proteins and the longer each sequence, the more time will be taken.

Step 4. As shown on the lower panel of Fig. 2, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the "Browse" button. To see the sample of batch input file, click on the button Batch-example.

Step 5. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of Multi-iPPseEvo.

Step 6. Click the Supporting Information button to download the benchmark datasets used to train and test the current predictor.

Caveat: To obtain the predicted result with the anticipated success rate, the entire sequence of the query protein rather than its fragment should be used as an input.

3 Results and Discussion

The proposed predictor was tested by the benchmark dataset $\mathbb{S}$ in Eq. 1, and shown as Table 1 and 2, which contains 375 non-phosphorylated proteins and 1,132 phosphorylated proteins. From Table 2, we can see that of the 1,132 proteins, 820 proteins have only one type of phosphorylation, 276 proteins have two types and 36 proteins have all of the three phosphorylations. Table 3 provides the test results of the nine metrics obtained by Multi-iPPseEvo via the 5-fold cross-validation on $\mathbb{S}$.

Considered the diversity of the data set, listed as Table 1 and 2, the experimental results shrieked a good news for the phosphorylation prediction. Of the two cross-validation experiments, Hamming_loss is 0.2520 and it means that most of instance label pair, about 75%, is classified correctly. The advisable sign of One-Error, which gains 0.2363, means most of the top-ranked label is in the set of proper labels of the instance. From the Coverage result, we know it is easy to go down the list of labels in order to cover all the proper labels. The Average-Precision is 0.6976 and the Recall is 0.7335, the Precision is 0.5998. However, it does not perform well in the metrics of Subset-Accuracy and Ranking-Loss. It may need much effort to improve the predictor in the future.

The metrics of Hamming_loss, One-Error, Accuracy, Average-Precision have shown positive sign for the predictor. Ranking-Loss and Subset-Accuracy have shown the negative sign. However, the Subset Accuracy is 0.4081, this may be due to the great diversity of the datum. As shown in the Table 2, the ratio between all concerned proteins and proteins with three types of phosphorylations is 1,507 to 36, it means the set of proteins with all of the three types is such a tiny part of the training dataset.

Above all, the experiments bring good news for human phosphorylation detection, especially for predicting the simultaneous phosphorylation proteins. The proposed predictor is a good choice for human phosphorylation detec-

tion. Particularly, the current proposed predictor has provided a use-friendly web-server that is no doubt very useful for the majority of experimental scientists in this or related areas.

4 Conclusions

In this work, we developed a method for detecting the phosphorylation proteins. In the proposed predictor, a query protein is formulated into the general pseudo amino acid composition (PseAAC) via the grey system theory, and the operation engine to run the Multi-iPPseEvo prediction is an ensemble classifier formed via a voting system to fuse different random forest classifiers.

The prediction model achieved a promising performance indicated by the 5-fold rigorous cross-validations. The detailed feature analysis in this study might help understand the human phosphorylation. The ensemble classifier Multi-iPPseEvo is a bioinformatics tool aimed to discriminate the proteins phosphorylation types. This is a first predictor ever developed for such a purpose. It is anticipated that Multi-iPPseEvo will become a very useful high throughout tool for identifying phosphorylated proteins in cells, stimulating a series of interesting follow-up researches in this and related areas.

Conflict of Interest

None declared.

Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (No. 61261027, 61262038, 31260273, 61202313), the Natural Science Foundation of Jiangxi Province, China (No. 20122BAB211033, 20122BAB201044, 20132BAB201053), the scholarship under the State Scholarship Fund (No. 201508360047). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] K. C. Chou, *Med. Chem.* **2015**, *11*, 218 – 234.
[2] Y. Xu, K. C. Chou, *Curr. Med. Chem.* **2016**, *16*, 591 – 603.
[3] Y. Xu, J. Ding, L. Y. Wu, K. C. Chou, *PloS one* **2013**, *8*.
[4] Y. Xu, X. J. Shao, L. Y. Wu, N. Y. Deng, K. C. Chou, *PeerJ.* **2013**, *1*.
[5] Y. Xu, X. Wen, X. J. Shao, N. Y. Deng, K. C. Chou, *Int. J. Mol. Sci.* **2014**, *15*, 7594 – 7610.
[6] Y. Xu, X. Wen, L. S. Wen, L. Y. Wu, N. Y. Deng, K. C. Chou, *PloS one* **2014**, *9*.
[7] P. Blume-Jensen, T. Hunter, *Nature* **2001**, *411*, 355 – 365.
[8] K. Kerman, H. Song, J. S. Duncan, D. W. Litchfield, H. B. Kraatz, *Anal. Chem.* **2008**, *80*, 9395 – 9401.
[9] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, *Science* **2002**, *298*, 1912-+.
[10] P. Cohen, *Eur. J. Biochem.* **2001**, *268*, 5001 – 5010.
[11] J. Ciesla, T. Fraczyk, W. Rode, *Acta Biochim. Pol.* **2011**, *58*, 137 – 148.
[12] R. Aebersold, M. Mann, *Nature* **2003**, *422*, 198 – 207.
[13] D. Zhang, C. Ortiz, Y. Xie, V. J. Davisson, D. Ben-Amotz, *Spectrochim. Acta, Part A* **2005**, *61*, 471 – 475.
[14] G. Burnett, E. P. Kennedy, *J. Biol. Chem.* **1954**, *211*, 969 – 980.
[15] N. Bhalla, M. Di Lorenzo, G. Pula, P. Estrela, *Biosens. Bioelectron.* **2014**, *54*, 109 – 114.
[16] H. Kaufmann, J. E. Bailey, M. Fussenegger, *Proteomics.* **2001**, *1*, 194 – 199.
[17] P. A. Weernink, G. Rijksen, *J. Biochem. Biophys. Methods* **1996**, *31*, 49 – 57.
[18] S. Martic, S. Beheshti, H. B. Kraatz, D. W. Litchfield, *Chem. Biodiversity* **2012**, *9*, 1693 – 1702.
[19] Z. Wang, J. Lee, A. R. Cossins, M. Brust, *Anal. Chem.* **2005**, *77*, 5770 – 5774.
[20] S. Yoon, K. Y. Han, H. S. Nam, V. Nga le, Y. S. Yoo, *J. Chromatogr. A* **2004**, *1056*, 237 – 242.
[21] K. L. Huss, P. E. Blonigen, R. M. Campbell, *J. Biomol. Screening* **2007**, *12*, 578 – 584.
[22] W.-R. Qiu, B.-Q. Sun, X. Xiao, D. Xu, K.-C. Chou, *Mol. Inform.* **2016**, n/a-n/a.
[23] W. Z. Lin, J. A. Fang, X. Xiao, K. C. Chou, *Mol. BioSyst.* **2013**, *9*, 634 – 644.
[24] L. Zhu, J. Yang, H. B. Shen, *Protein J.* **2009**, *28*, 384 – 390.
[25] M. L. Zhang, Z. H. Zhou, *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1338 – 1351.
[26] S. M. Liu, J. H. Chen, *Expert Syst Appl.* **2015**, *42*, 1083 – 1093.
[27] Y. Xu, L. Jiao, S. Wang, J. Wei, Y. Fan, M. Lai, E. I. Chang, *Microsc. Res. Tech.* **2013**, *76*, 1266 – 1277.
[28] H. B. Borges, J. C. Nievola, *80 ed., WASET.* **2013**, 85 – 89.
[29] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, Z. Yu, *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **2013**, *10*, 1 – 1.
[30] I. Pillai, G. Fumera, F. Roli, *Pattern Recogn.* **2013**, *46*, 2256 – 2266.
[31] G. Tsoumakas, I. Katakis, I. Vlahavas, in *Data Mining and Knowledge Discovery Handbook* (Eds.: O. Maimon, L. Rokach), Springer US, **2010**, 667 – 685.
[32] J. Read, B. Pfahringer, G. Holmes, E. Frank, *Lect. Notes Comput. Sci.* **2009**, *5782*, 254 – 269.
[33] G. Tsoumakas, I. Katakis, *int. J. Data Warehous* **2007**, *3*, 13.
[34] K. C. Chou, *Mol. BioSyst.* **2013**, *9*, 1092 – 1100.
[35] K. C. Chou, H. B. Shen, *J. Proteome Res.* **2007**, *6*, 1728 – 1734.
[36] K. C. Chou, H. B. Shen, *Nat. Protoc.* **2008**, *3*, 153 – 162.
[37] K. C. Chou, H. B. Shen, *PloS one* **2010**, *5*.
[38] H. B. Shen, K. C. Chou, *Biochem. Biophys. Res. Commun.* **2007**, *355*, 1006 – 1011.
[39] H. B. Shen, K. C. Chou, *Protein Pept. Lett.* **2009**, *16*, 1478 – 1484.
[40] H. B. Shen, K. C. Chou, *J. Theor. Biol.* **2010**, *264*, 326 – 333.
[41] H. B. Shen, K. C. Chou, *J. Biomol. Struct. Dyn.* **2010**, *28*, 175 – 186.
[42] H. B. Shen, K. C. Chou, *J. Proteome Res.* **2009**, *8*, 1577 – 1584.
[43] H. B. Shen, K. C. Chou, *Amino acids* **2007**, *32*, 483 – 488.
[44] K. C. Chou, *Proteins: Struct., Funct., and Genet.* **2001**, *44*, 60 – 60.
[45] K. C. Chou, *Proteins: Struct., Funct., and Genet.* **2001**, *43*, 246 – 255.
[46] K. C. Chou, *Bioinformatics* **2005**, *21*, 10 – 19.
[47] D. S. Cao, Q. S. Xu, Y. Z. Liang, *Bioinformatics* **2013**, *29*, 960 – 962.
[48] S. X. Lin, J. Lapointe, *JBiSE* **2013**, *6*, 435–442.

[49] Z. U. Khan, M. Hayat, M. A. Khan, *J. Theor. Biol.* **2015**, *365*, 197 – 203.
[50] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, *J. Theor. Biol.* **2015**, *364*, 284 – 294.
[51] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, *J. Theor. Biol.* **2015**, *365*, 96 – 103.
[52] K. C. Chou, *Curr. Proteomics* **2009**, *6*, 262 – 274.
[53] B. Liu, X. L. Wang, Q. Zou, Q. W. Dong, Q. C. Chen, *Mol. Inform.* **2013**, *32*, 775 – 782.
[54] B. Liu, J. H. Xu, S. X. Fan, R. F. Xu, J. Y. Zhou, X. L. Wang, *Mol. Inform.* **2015**, *34*, 8 – 17.
[55] P. F. Du, S. W. Gu, Y. S. Jiao, *Int. J. Mol. Sci.* **2014**, *15*, 3495 – 3506.
[56] B. Liu, F. L. Liu, X. L. Wang, J. J. Chen, L. Y. Fang, K. C. Chou, *Nucleic Acids Research* **2015**, *43*, W65-W71.
[57] W. Chen, H. Lin, K. C. Chou, *Molecular bioSystems* **2015**, *11*, 2620 – 2634.
[58] A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, S. F. Altschul, *Nucleic Acids Research* **2001**, *29*, 2994 – 3005.
[59] X. Xiao, J. L. Min, P. Wang, K. C. Chou, *PloS one* **2013**, *8*.
[60] Z. C. Wu, X. Xiao, K. C. Chou, *Mol. BioSyst.* **2011**, *7*, 3287 – 3297.
[61] W. Z. Lin, J. A. Fang, X. Xiao, K. C. Chou, *Mol. BioSyst.* **2013**, *9*, 634 – 644.
[62] W. Z. Lin, J. A. Fang, X. Xiao, K. C. Chou, *PloS one* **2012**, *7*.
[63] L. Breiman, *Mach Learn.* **2001**, *45*, 5 – 32.
[64] W. Chmielnicki, K. Stapor, *Hybrid Artificial Intelligence Systems, Pt 1* **2010**, *6076*, 162 – 169.
[65] H. B. Hashemi, A. Shakery, M. P. Naeini, *2009 International Conference of Soft Computing and Pattern Recognition* **2009**, 436–441.
[66] H. B. Shen, K. C. Chou, *Bioinformatics* **2006**, *22*, 1717 – 1722.
[67] H. B. Shen, K. C. Chou, *Biochem. Biophys. Res. Commun.* **2007**, *363*, 297 – 303.
[68] D. J. Yu, X. W. Wu, H. B. Shen, J. Yang, Z. M. Tang, Y. Qi, J. Y. Yang, *IEEE Trans. Nanobiosci.* **2012**, *11*, 375 – 385.
[69] K. C. Chou, H. B. Shen, *Anal. Biochem.* **2007**, *370*, 1 – 16.
[70] M. Hayat, A. Khan, M. Yeasin, *Amino acids* **2012**, *42*, 2447 – 2460.
[71] K. C. Chou, H. B. Shen, *J. Cell. Biochem.* **2006**, *99*, 517 – 527.
[72] M. L. Zhang, *Neural Process Lett.* **2009**, *29*, 61 – 74.
[73] M. A. Tahir, J. Kittler, A. Bouridane, *Pattern Recognit. Lett.* **2012**, *33*, 513 – 523.
[74] A. Veloso, W. Meira, M. Goncalves, M. Zaki, *Lect. Notes Artif. Int.* **2007**, *4702*, 605 – 612.
[75] K. C. Chou, C. T. Zhang, *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275 – 349.
[76] K. C. Chou, *J. Theor. Biol.* **2011**, *273*, 236 – 247.
[77] S. Mondal, P. P. Pai, *J. Theor. Biol.* **2014**, *356*, 30 – 35.
[78] G. P. Zhou, *J. Protein Chem.* **1998**, *17*, 729 – 738.
[79] G. P. Zhou, K. Doctor, *Proteins* **2003**, *50*, 44 – 48.

Received: June 17, 2016
Accepted: September 7, 2016
Published online: September 29, 2016

