iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC

Jianhua Jia , Xiaoyan Li , Wangren Qiu , Xuan Xiao ,
Kuo-Chen Chou

PII: S0022-5193(18)30497-1
DOI: https://doi.org/10.1016/j.jtbi.2018.10.021
Reference: YJTBI 9664

To appear in: *Journal of Theoretical Biology*

Received date: 5 August 2018
Revised date: 16 September 2018
Accepted date: 8 October 2018

Please cite this article as: Jianhua Jia , Xiaoyan Li , Wangren Qiu , Xuan Xiao , Kuo-Chen Chou , iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC, *Journal of Theoretical Biology* (2018), doi: https://doi.org/10.1016/j.jtbi.2018.10.021

Highlights

- Knowledge of protein–protein interactions (PPIs) may provide valuable insights into the inner workings of cells.

- A powerful predictor has been proposed to identify PPIs in a cell.

- A user-friendly web-server for the predictor has been established by which the majority of experimental scientists can easily get their desired results.

iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC

Jianhua Jia[1,3]*, Xiaoyan Li[1], Wangren Qiu[1], Xuan Xiao[1,3], Kuo-Chen Chou[2,3]

1 Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403 China; **2** Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China; **3** Gordon Life Science Institute, Boston, MA 02478, USA;

Authors' e-mail addresses
Jianhua Jia: jjia@gordonlifescience.org
Xiaoyan Li: 282086884@qq.com
Wangren Qiu: qiuone@163.com
Xuan Xiao: xxiao@gordonlifescience.org
Kuo-Chen Chou: kcchou@gordonlifescience.org

*Corresponding author

Running Title: Identify Protein-Protein Interactions

ABSTRACT

Investigation into the network of protein–protein interactions (PPIs) will provide valuable insights into the inner workings of cells. Accordingly, it is crucially important to develop an automated method or high-throughput tool that can efficiently predict the PPIs. In this study, a new predictor, called "iPPI-PseAAC(CGR)", was developed by incorporating the information of "chaos game representation" into the PseAAC (Pseudo Amino Acid Composition). The advantage by doing so is that some key sequence-order or sequence-pattern information can be more effectively incorporated during the treatment of the protein pair samples. The operation engine used in this predictor is the random forest algorithm. It has been observed via the cross-validations on the widely used benchmark datasets that the success rates achieved by the proposed predictor are remarkably higher than those by its existing counterparts. For the convenience of the most experimental scientists, a user-friendly web-server for the new predictor has been established at http://www.jci-bioinfo.cn/iPPI-PseAAC(CGR), by which users can easily get their desired results without the need to go through the detailed mathematics.

Keywords: Chaos game representation; Pseudo amino acid composition; Random forests; Cross validation; *Saccharomyces Cerevisiae*; *Helicobacter Pylori*; Web-server

1. INTRODUCTION

The smallest unit of life is a cell, which contains numerous protein molecules. Most of the functions critical to the cell's survival are performed via the protein-protein interactions (PPIs) therein. Therefore, it is indispensable to study PPIs in order to really understand the molecular underpinnings of life since they affect all the biological processes in a living cell.

Currently, the determination of PPIs through experiments is mainly by the three manners: (1) yeast two-hybrid assay, (2) protein chips, and (3) mass spectrometry of purified protein complexes. But it is expensive, time-consuming, and labor-intensive to determine PPIs purely based on the experimental methods. Facing the explosive growth of protein sequences occurring in the post-genomic age, we are challenged to develop computation method to identify PPIs based on the sequence information alone.

During the last decade or so, considerable efforts have been made in this regard (see, e.g., [1-8]). Although these methods did play important roles in stimulating the development of this area, further endeavor is needed to enhance the power of identifying PPIs.

The present study was initiated in an attempt to develop a new predictor called iPPI-PseAAC(CGR) to identify protein-protein interactions by using random forest algorithm [9] and incorporating "chaos game representation" [10, 11] into general PseAAC (Pseudo Amino Acid Composition) [12].

To make the presentation of this paper logically more clear and transparent, its reported results easier to be repeated by others, and its proposed method practically more useful, the 5-step rules [12] were followed, as done in a series of recent publications (see, e.g., [13-38]).

2. MATERIAL AND METHODS

2.1. Benchmark Datasets

The first step in the 5-step rules [12] is how to construct or select a valid benchmark dataset to train and test the predictor. Two benchmark datasets were used for the current study: one is called the S.C. dataset used for studying the PPIs in the cell of *Saccharomyces Cerevisiae*; while the other called the H.P. dataset for studying the PPIs in the cell of *Helicobacter Pylori*.

2.1.1. S.C. dataset. To obtain a high quality benchmark dataset, the proteins in *Saccharomyces Cerevisiae* [39] were collected according to the following criteria. (**1**) To avoid fragments, each of the included proteins must contain at least 50 residues. (**2**) To reduce the homology bias, none of the included proteins has ^{3} 40% pairwise sequence identity. Based on the 7,374 proteins thus obtained, the S.C. dataset, $\mathbb{S}_{\text{S.C.}}$, was constructed as formulated below

$$\mathbb{S}_{\text{S.C.}} = \mathbb{S}^{+}_{\text{S.C.}} \cup \mathbb{S}^{-}_{\text{S.C.}} \tag{1}$$

where $\mathbb{S}_{\text{S.C.}}$ contains 50,652 protein pairs, of which 17,505 are interactive belonging to the positive subset $\mathbb{S}^{+}_{\text{S.C.}}$, while 33,147 are non-interactive belonging to the negative subset $\mathbb{S}^{-}_{\text{S.C.}}$, and ∪ represents the union in the set theory. For the codes of these proteins and their detailed sequences, see Supporting Information S1.

2.1.2. H.P. dataset. For facilitating comparison later, the benchmark dataset for studying the PPIs in the cell of *Helicobacter Pylori* [40] was also considered since it was used by many investigators [40-45] to test their methods with the results well documented. Likewise, here such benchmark is formulated by

$$\mathbb{S}_{\text{H.P.}} = \mathbb{S}^{+}_{\text{H.P.}} \cup \mathbb{S}^{-}_{\text{H.P.}} \tag{2}$$

where $\mathbb{S}_{\text{H.P.}}$ contains 2,916 protein pairs, of which 1,458 are interactive belonging to the positive subset $\mathbb{S}^{+}_{\text{H.P.}}$, while 1,458 are non-interactive belonging to the negative subset $\mathbb{S}^{-}$. For the codes of these proteins and their detailed sequences, see Supporting Information S2.

2.2. Using Pseudo Amino Acid Composition to Represent Protein Pairs

The second step in the 5-step rules [12] is how to formulate the biological sequence samples with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms (such as “Covariance Discriminant” or “CD” algorithm [46, 47], “Nearest Neighbor” or “NN” algorithm [48, 49], “Support Vector Machine” or “SVM” algorithm [50, 51], and “Random Forest” or “RF” algorithm [52, 53]) can only handle vectors as elaborated in a comprehensive review [54]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [55] or PseAAC [56] was proposed. Ever since the concept of Chou’s PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics (see, e.g., [57-63] [13, 50, 64-95] as well as a long list of references cited in [96]). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called ‘PseAAC-Builder’ [97], ‘propy’ [98], and ‘PseAAC-General’ [99], were established: the former two are for generating various modes of Chou’s special PseAAC [100]; while the 3rd one for those of Chou’s general PseAAC [12], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as “Functional Domain” or “FunD” mode, “Gene Ontology” or “GO” mode, and “Sequential Evolution” or “PSSM” mode. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its idea and approach were extended to PseKNC (Pseudo K-tuple Nucleotide Composition) to generate various feature vectors for DNA/RNA sequences [101] that have proved very successful as well [15, 17, 19, 102-106]. Particularly, recently a very powerful web-server called ‘Pse-in-One’ [107] and its updated version ‘Pse-in-One2.0’ [108] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the users’ need or their own definition.” According to the general PseAAC [12], any protein sequence can be formulated as a PseAAC vector given by

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

where **T** is a transpose operator, while the subscript Ω is an integer parameter and its value as well as the components Ψ_u $(u = 1, 2, \cdots, \Omega)$ will depend on how to extract the desired information from the amino acid sequence of **P**, as elaborated below.

In this study, we are to use CGR (Chaos Game Representation) proposed by Fiser et al. [10] to define the pseudo components in Eq.3. To realize this, we first converted the protein sequences to the nucleotide sequences according to **Table 1**, which was proposed by Deschavanne and Tuffery [109]. The advantage of using such code-converting method is able to keep balanced base composition so as to maximize the differences between the amino acid codes. Particularly, it can also provide useful insights via intuitive plot or graph for facilitating analysis of complicated biological systems as demonstrated in a series of works by the previous investigators (see, e.g. [110-125]).

A brief description to generate the CGR via **Table 1** is as follows. In a [0,1]×[0,1] square, the four vertices of the defined square correspond to the four letters: **A**, **C**, **G**, and **T**. Then, the CGR-plot can be obtained via the following steps: (**1**) put an initial point at an random site of the square (we selected the center of it in this paper); (**2**) place the second point at the half way between the initial point and the vertex corresponding to the first letter of the nucleotide sequence; (**3**) place the *i*th point half way between the (*i*-1)th point and the vertex corresponding to the *i*th letter; (**4**) go to Step 3 until you reach the end of the nucleotide sequence.

Shown in **Fig.1** is the CGR thus generated for a given protein sequence. As we can see from the figure, many important features, which are hidden in a long and complicated biological sequence, can be clearly revealed via its CGR. Subsequently, we divided the GGR square to 4×4=16 sub-squares with each having the same size. By calculating the number of the points within each of the 16 sub-squares, plus the occurrence frequencies of the 20 native amino acids therein, we have a total of $(16 + 20) = 36$ components to define the components of Eq.3 and its Ω. In other words, the PseAAC vector for **P** in this study is a 36-D (dimensional) vector.

Accordingly, for a protein pair formed by $\mathbf{P}^{k1}$ and $\mathbf{P}^{k2}$, the corresponding PseAAC vector can be formulated by their orthogonal sum; i.e.,

$$\mathbf{P}^{k1} \oplus \mathbf{P}^{k2} = [\Psi_1^{k1} \quad \Psi_2^{k1} \quad \cdots \quad \Psi_{36}^{k1} \quad \Psi_1^{k2} \cdots \quad \Psi_{36}^{k2}]^{\mathbf{T}} \tag{4}$$

where $\mathbf{P}^{k1}$ and $\mathbf{P}^{k2}$ as well as their components have exactly the same meaning as those in Eq.3 except for that they are now referred to the $\mathbf{P}^{k1}$ or $\mathbf{P}^{k2}$, instead of **P**, and the symbol $\oplus$ represents the sign of orthogonal sum. Thus, a 72-D PseAAC vector is used to formulate the sample of a PP pair.

2.3. Random Forest and Ensemble Classifier

The third step in the 5-step rules [12] is how to introduce or develop a powerful algorithm (or engine) to operate the prediction. Here we used the random forest (RF) algorithm [9], which has been widely used in the area of computational biology (see, e.g., [7, 33, 52, 53, 106, 126-134]). The detailed procedures and formulation of RF have been very clearly described in [9], and hence there is no need to repeat here. The random forests algorithm usually produces a remarkable improvement in performance over the single decision tree classifier [135]. Moreover, we found that the random forests algorithm was not sensitive to the number of trees according to the aforementioned sample formulation. Thus, a total of 200 trees were used in order to alleviate computational cost and the over-fitting problem.

In this study, there are two benchmark datasets, one is the S.C. dataset (see Eq.1 and Supporting Information S1) and the other is the H.P. datasets (see Eq.2 and Supporting Information S2). The model trained by the former is for predicting the PPIs in *Saccharomyces Cerevisiae*; while the model trained by the latter is for predicting the PPIs in *Helicobacter Pylori*. In the S.C. dataset, the number of negative samples is much larger than that of the positive samples, meaning it is a very imbalanced or skewed dataset. But in the H.P. dataset there is no such a problem since its positive subset is the same in size as its negative subset. To alleviate the biased outcomes caused by the skewed dataset, there are some existing methods to add some theoretical or hypothetical samples into the smaller subsets, such as the “Monte Carlo samples expanding” approach [136, 137], “seed-propagation” approach [138], “LogiBoost” [139], “SMOTE” (synthetic minority over-sampling technique) approach [140-142], “Bootstrap” approach [8], and IHTS (Inserting Hypothetical Training Samples) treatment [36, 143-147]. But here we used a different approach to deal with the unbalanced problem, as described below.

From the 33,147(=$n_{\text{S.C.}}^{-}$) negative samples in $\mathbb{S}_{\text{S.C.}}^{-}$ of Eq.1, we randomly picked 17,505 (=$n_{\text{S.C.}}^{+}$) samples to form a negative sub-subset whose size is the same as the positive subset $\mathbb{S}_{\text{S.C.}}^{+}$. Repeat the above procedure for 7 times, generating an array of 7 negative sub-subsets. Using each of these negative sub-subsets and the positive subset $\mathbb{S}_{\text{S.C.}}^{+}$ (Eq.1) to train the Random Forest model, we obtained an array of 7 individual predictors $\text{RF}(k)$ $(k = 1,2,\cdots,7)$. Based on them, an ensemble classifier was formed via a voting system, as formulated by

$$\mathbb{RF}^{\text{E}} = \text{RF}(1)\forall\cdots\forall\text{RF}(7) = \forall_{k=1}^{7}\text{RF}(k) \tag{5}$$

where $\mathbb{RF}^{\text{E}}$ stands for the ensemble classifier, and the symbol ∀ for the fusing operator. For the detailed procedures of how to fuse the results from the seven individual predictors to reach a final outcome via the voting system, see Eqs.30-35 in [148], where a crystal clear and elegant derivation was elaborated and hence there is no need to repeat here. To provide an intuitive picture, a flowchart is given in **Fig.2** to illustrate how the seven individual RF predictors are fused into the ensemble classifier.

But for predicting the PPIs in the *Helicobacter Pylor*, there is no need at all to go through the above ensemble-learning procedure. This is because the H.P. benchmark dataset is a balanced one; its negative and positive subsets are the same in size, namely $n_{\text{H.P.}}^{-} = n_{\text{H.P.}}^{+}$. Therefore, we can directly use $\mathbb{S}_{\text{H.P.}}$(Eq.2) to train the random forest model.

The final predictor thus obtained is called “**iPPI-PseAAC(CGR)**”, where “i” stands for “identify”, “PPI” for “protein-protein interaction”, and “PseAAC(CGR)” for “incorporating CGR (Chaos Game Representation) into general PseAAC (Pseudo Amino Acid Composition)”. In dealing with the *Saccharomyces Cerevisiae* system, the RF-ensemble engine is on; in dealing with the *Helicobacter Pylor* system, however, the RF-ensemble engine will be replaced by RF only.

2.4. Cross-Validation

The fourth step in the 5-step rules [12] is how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. To address this problem, we need to consider the following two sub-problems. (**1**) What metrics should be used to quantitatively measure the predictor's quality? (**2**) What test approach should be adopted to score the metrics?

2.4.1. A set of four intuitive metrics

For examining the performance of a predictor in identifying whether two proteins are in interaction with each other, four metrics [149] are often used in literature; they are (**1**) overall accuracy or Acc, (**2**) Mathew's correlation coefficient or MCC, (**3**) sensitivity or Sn, and (**4**) specificity or Sp. But their conventional formulations directly copied from math books are difficult to understand for most experimental scientists, particularly the one for MCC. Fortunately, by using the symbols introduced by Chou [150, 151] in studying the signal peptide cleavage sites, a set of intuitive metrics were derived [51, 152, 153], as given below

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \ \leq \mathrm{Sn} \leq 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \leq \mathrm{Sp} \leq 1 \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \leq \mathrm{Acc} \leq 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \leq \mathrm{MCC} \leq 1 \end{cases} \tag{5}$$

where N^+ represents the total number of interactive protein pairs investigated whereas N_-^+ the number of true interactive pairs incorrectly predicted as of non-interactive pair; N^- the total number of the non-interactive protein pairs investigated whereas N_+^- the number of non-interactive protein pairs incorrectly predicted as of interactive pair. Because of its intuitiveness and ease to be understood, the set of metrics has been increasingly and widely used in computational biology (see, e.g., [8, 15, 17, 19, 20, 22, 24-26, 28-31, 33, 35, 53, 72, 102, 104-106, 129, 130, 132-134, 141-144, 154-179]. It is instructive to point out that both the original four metrics [149] in math books and the intuitive ones in Eq.5 are valid only for the single-label systems (where each sample belongs to one and only one class or attribute). For the multi-label systems (where a sample may simultaneously belong to several classes or attributes), whose existence has become more frequent in system biology [14, 16, 18, 23, 32, 34, 36, 145, 180], system medicine [21, 181] and biomedicine [131], a completely different set of metrics as defined in [182] is absolutely needed.

2.4.2. Cross-validation

In statistical prediction, the following three cross-validation methods are often used to check the performance of a predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [183]. Of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in [12] and demonstrated by Eqs.28-30 therein. However, to reduce the computational time, in this study we adopted the 5-fold and 10-fold cross validation tests, as done by many investigators with random forests algorithm as the operation engine.

2.5. Web-Server for iPPI-PseAAC(CGR)

The last but not the least important step of the 5-step rules [12] is how to establish a user-friendly web-server for the predictor that is accessible to the public. As pointed out in [184] and demonstrated in a series of recent publications (see, e.g., [14-21, 23, 26, 27, 29, 32, 34, 105, 134, 164, 174, 175, 177, 178, 180, 185-188]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have significantly increased the impacts of bioinformatics on medical science [54], driving medicinal chemistry into an unprecedented revolution [96]. In view of this, the web-server for the IPP-PseAAC (CGR) predictor has also been established at http://www.jci-bioinfo.cn/iPPI-PseAAC(CGR). Shown in **Fig.3** is the semi-screenshot of its top page. Note: when using the web-server, click the S.C. button if you want to predict the PPIs in *Saccharomyces Cerevisiae*; click on H.P. button if you want to predict the PPIs in *Helicobacter Pylori*.

3. RESULTS AND DISCUSSION

Listed in **Table 2** are the scores of the four metrics (cf. Eq.5) obtained by iPPI-PseAAC(CGR) on the S.C. benchmark dataset (Supporting Information S1) via the 5-fold cross-validation. For facilitating comparison, listed there are also the corresponding rates obtained by the existing state-of-the-art method [7]. As we can see from the table that, the success rates for Acc and MCC achieved by the proposed predictor iPPI-PseAAC(CGR) are higher than those by iPPI-Esml [7], the existing state-of-the-art predictor. Although the rate of Sn by iPPI-Esml is about 6% higher than that by iPPI-PseAAC(CGR), the rate of Sp by iPPI-Esml is about 8% lower than that by the iPPI-PseAAC(CGR). Actually, among the four metrics in Eq.5, the most important are the Acc and MCC: the former reflects the overall accuracy of a predictor; while the latte, its stability in practical applications. Sn and Sp are actually constrained with each other [137]. Therefore, it is meaningless to use only one of the two for comparing the quality of two predictors. In other words, a meaningful comparison in this regard should use the rates of both Sn and Sp, or even better use their combination that is none but MCC, for which the rate achieved by the iPPI-PseAAC(CGR) is about 6% higher than that by the iPPI-Esml predictor.

Listed in **Table 3** are the success rates obtained by the proposed predictor on the H.P. benchmark dataset (Supporting Information S2) via the 10-fold cross-validation. For facilitating comparison, listed there are also the corresponding rates obtained by the other six prediction methods [40-45]. It is clearly shown from there that the newly proposed predictor iPPI-PseAAC(CGR) is remarkably superior to its six cohorts. Besides, none of them [40-45] has web-server as iPPI-PseAAC(CGR) does, and hence their practical application value by the majority of experimental scientists is very limited.

4. CONCLUSION

iPPI-PseAAC(CGR) is a powerful predictor for identifying the protein-protein interactions in cell according to the protein sequence information alone. In the predictor, each protein is formulated by a PseAAC vector formed by 36 components, of which 20 are the occurrence frequencies of the 20 native amino acid residues in the protein, and the remaining 16 components are derived from the chaos game representation. Thus, each protein pair is denoted by a $36 \times 2 = 72$-D PseAAC vector. The learning machine implemented in the new predictor is random forests and their ensemble. Its success rates have been examined by two stringent benchmark datasets: one for *Saccharomyces Cerevisiae*, and one for *Helicobacter Pylori*, indicating the new predictor is superior to its counterparts. A public-accessible web-server for iPPI-PseAAC(CGR) has been established. We anticipate that it will become a very useful high throughput toll for identifying PPIs in any related areas.

ACKNOWLEDGMENTS

This work was partially supported by the National Nature Science Foundation of China (No. 61761023, 31760315, 31560316, 61462047), the Natural Science Foundation of Jiangxi Province, China (No.20171BAB202020, 20171ACB20023, 20161BAB202047), China Postdoctoral Science Foundation Funded Project (No.2017M612949). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Table 1. Reverse encoding for the amino acids used in this study

A=GCT	G=GGT	M=ATG	S=TCA	C=TGC	H=CAC	N=AAC	T=ACT	D=GAC	I=ATT
P=CCA	V=GTG	E=GAG	K=AAG	Q=CAG	W=TGG	F=TTC	L=CTA	R=CGA	Y=TAC

The encoding approach was according to Deschavanne and Tuffery [109].

Table 2. The results obtained by the 5-fold cross-validation on the S.C. benchmark dataset (Online Supporting Information S1).

Method	Acc (%)[a]	MCC [a]	Sn (%)[a]	Sp (%)[a]
iPPI-Esml[b]	84.29	0.7063	97.64	77.43
iPPI-PseAAC(CGR)[c]	88.01	0.7624	91.09	85.37

[a] See Eq.5 for the metrics definition.
[b] Proposed in [7].
[c]Proposed in this paper.

Table 3. Compared with the other seven methods via the 10-cross-validation on the *H. P.* dataset [40] (Online Supporting Information S2).

Method	Acc (%)	MCC	Sn (%)	Sp (%)	Web-server
Bock and Gough [a]	75.80	N/A	69.80	80.20	No
Guo et al. [b]	80.96	0.5577	78.65	83.20	No
Martin [c]	83.40	N/A	79.90	85.70	No
Nanni [d]	83.00	N/A	80.60	85.10	No
Nanni and Lumini [e]	86.60	N/A	86.70	85.00	No
Xia et al. [f]	88.40	N/A	88.20	89.20	No
iPPI-PseAAC(CGR)[g]	**92.95**	**0.8505**	**97.61**	**88.00**	**Yes**

[a] Results reported by Bock et al. [41].
[b] Results reported by Guo et al. [42].
[c] Results reported by Martin et al. [40].
[d] Results reported by Nanni. [43].
[e] Results reported by Nanni et al. [44].
[f] Results reported by Xia et al. [45].
[g] Proposed in this paper.

FIGURE LEGENDS

Figure 1. A CGR-plot for one protein sequence. See the text in Section 2.2 for further explanation.

Figure 2. The flowchart to show the procedure of the ensemble approach. See the text in Section 2.3 for further explanation.

Figure3. A semi-screenshot to show the top-page of the iPPI-PseAAC(GPR) web-server at http://www.jci-bioinfo.cn/iPPI-PseAAC(CGR). See the text in Section 2.5 for further explanation.

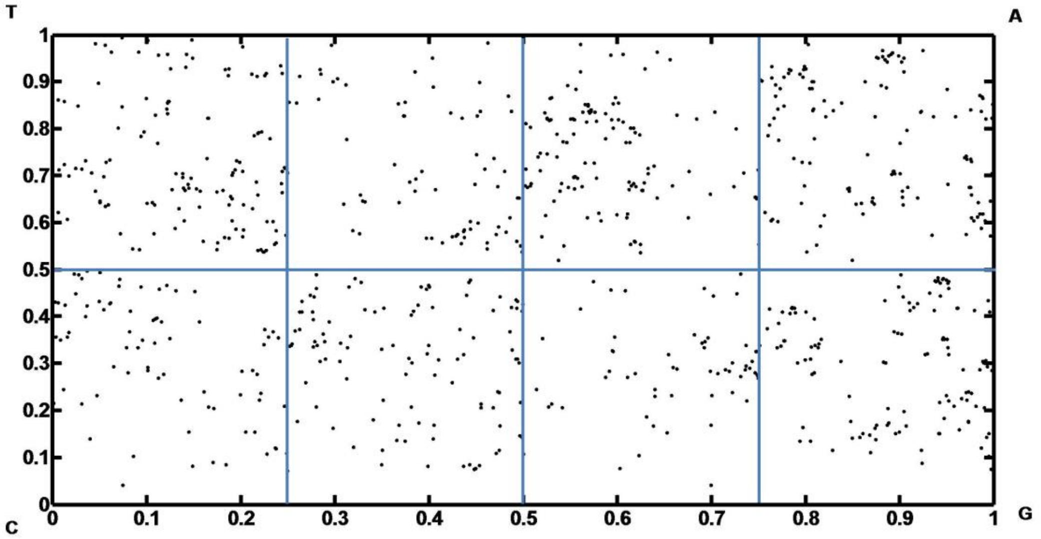


Figure 1

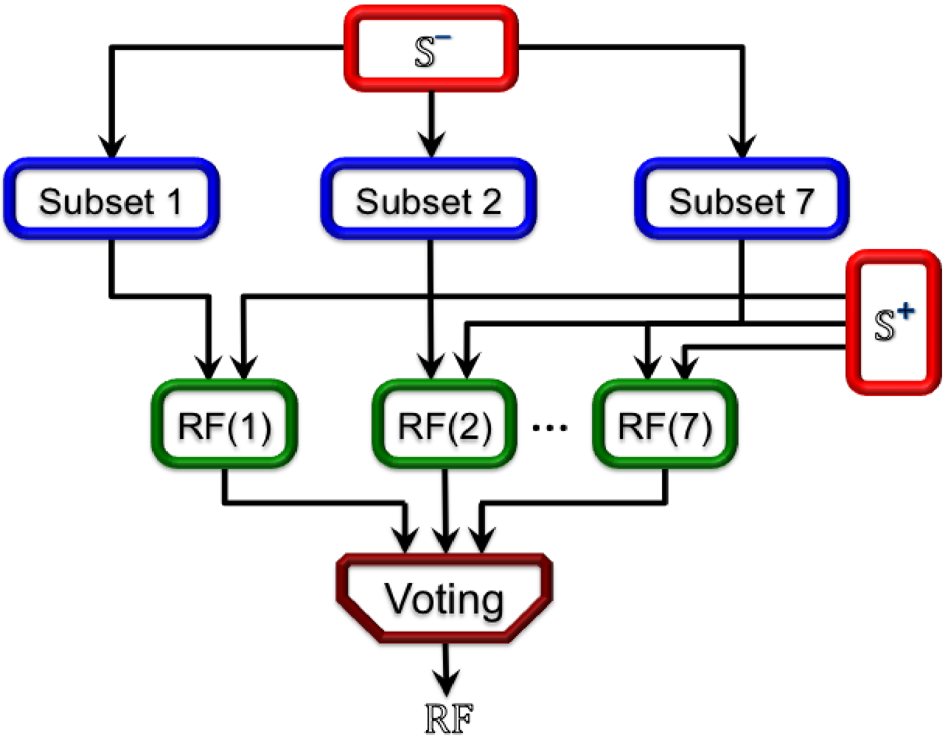
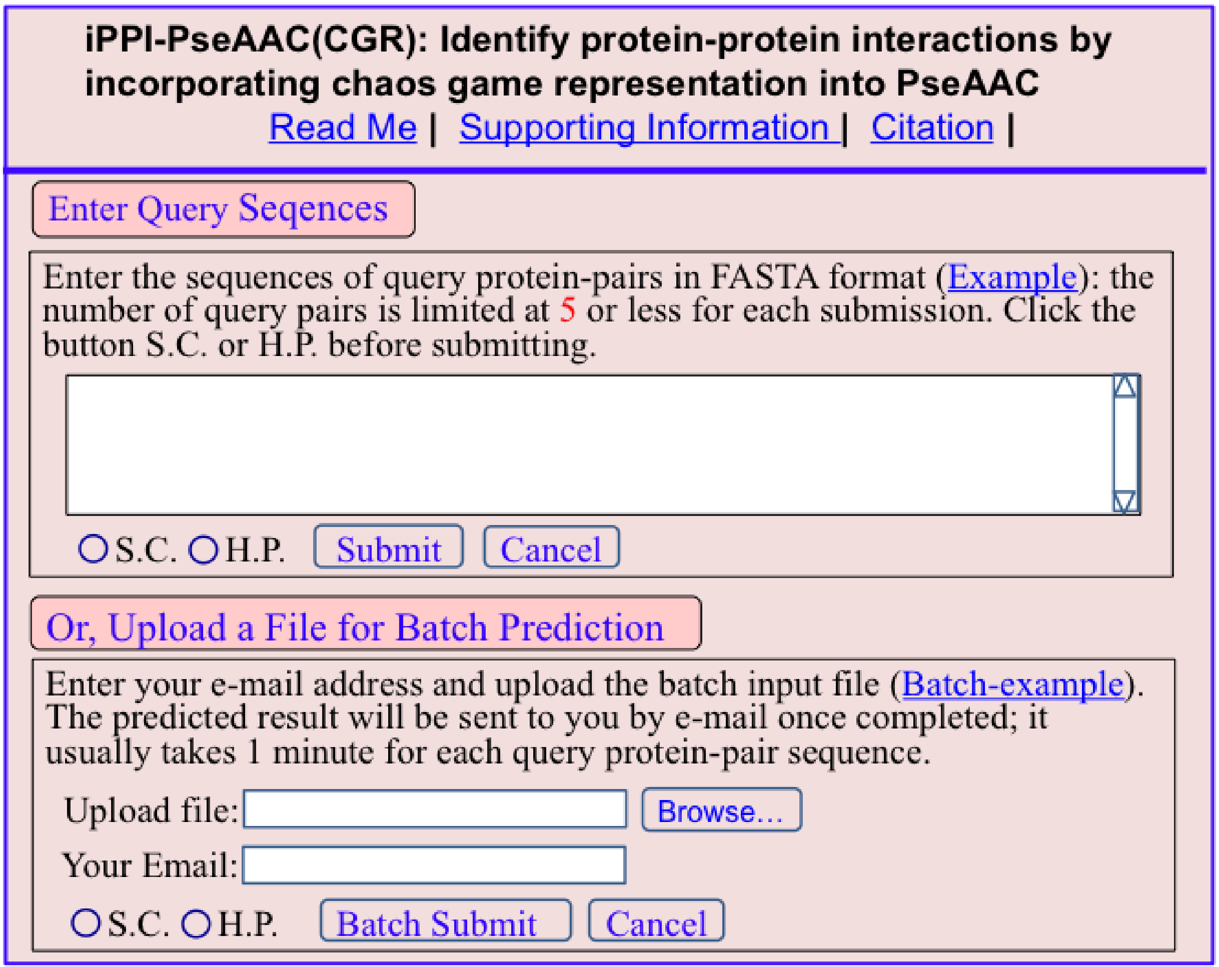


Figure 2



iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC

Read Me | Supporting Information | Citation |

Enter Query Seqences

Enter the sequences of query protein-pairs in FASTA format (Example): the number of query pairs is limited at 5 or less for each submission. Click the button S.C. or H.P. before submitting.

Or, Upload a File for Batch Prediction

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 1 minute for each query protein-pair sequence.

Upload file:

Your Email:

Figure 3

REFERENCES

[1] K.C. Chou, Y.D. Cai, Predicting protein-protein interactions from sequences in a hybridization space. Journal of Proteome Research 5 (2006) 316-322.
[2] L. Hu, T. Huang, X. Shi, W.C. Lu, Y.D. Cai, K.C. Chou, Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties PLoS ONE 6 (2011) e14556.
[3] T. Huang, L. Chen, Y.D. Cai, Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS ONE 6 (2011) e25297.
[4] L.H. Ren, Y.Z. Shen, Y.S. Ding, Bio-entity network for analysis of protein-protein interaction networks. Asian Journal of Control 13 (2011) 726-737.
[5] L.L. Hu, K.Y. Feng, Y.D. Cai, Using Protein-protein Interaction Network Information to Predict the Subcellular Locations of Proteins in Budding Yeast. Protein & Peptide Letters 19 (2012) 644-651.
[6] B.Q. Li, T. Huang, L. Liu, Y.D. Cai, Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS ONE 7 (2012) e33393.
[7] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol 377 (2015) 47-56.
[8] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC). J Biomol Struct Dyn (JBSD) 34 (2016) 1946-1961.
[9] L. Breiman, Random forests. Machine learning 45 (2001) 5-32.
[10] A. Fiser, G.E. Tusnady, I. Simon, Chaos game representation of protein structures. Journal of molecular graphics 12 (1994) 302-304.
[11] H.J. Jeffrey, Chaos game representation of gene structure. Nucleic Acids Research 18 (1990) 2163-2170.
[12] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273 (2011) 236-247.
[13] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci Rep 7 (2017) 42362.
[14] X. Cheng, X. Xiao, pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC. Molecular BioSystems 13 (2017) 1722-1727.
[15] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Molecular Therapy - Nucleic Acids 7 (2017) 155-163.
[16] X. Cheng, X. Xiao, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene (Erratum: ibid., 2018, Vol.644, 156-156) 628 (2017) 315-321.

[17] B. Liu, S. Wang, R. Long, iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics 33 (2017) 35-41.
[18] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics 33 (2017) 3524-3531.
[19] B. Liu, F. Yang, 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Molecular Therapy - Nucleic Acids 7 (2017) 267-277.
[20] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, J.H. Jia, iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 110 (2018) 239-246.
[21] X. Cheng, S.G. Zhao, X. Xiao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics (Corrigendum, ibid., 2017, Vol.33, 2610) 33 (2017) 341-346.
[22] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites. Molecular Therapy: Nucleic Acid 11 (2018) 468-474.
[23] X. Cheng, X. Xiao, pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics 110 (2018) 50-58.
[24] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics doi:10.1016/j.ygeno.2018.01.005 (2018).
[25] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. Analytical Biochemistry 550 (2018) 109-116.
[26] F. Li, C. Li, T.T. Marquez-Lago, A. Leier, T. Akutsu, A.W. Purcell, A.I. Smith, T. Lightow, R.J. Daly, J. Song, Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics doi: 10.1093/bioinformatics/bty522 (2018).
[27] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, G.I. Webb, PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework. Journal of Theoretical Biology 443 (2018) 125-137.
[28] B. Liu, K. Li, D.S. Huang, iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach. Bioinformatics doi:10.1093/bioinformatics/bty458 (2018).
[29] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Briefings in Bioinformatics doi: 10.1093/bib/bby028 (2018).
[30] Z.D. Su, Y. Huang, Z.Y. Zhang, Y.W. Zhao, D. Wang, W. Chen, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics doi:10.1093/bioinformatics/bty508 (2018).
[31] B. Liu, F. Weng, D.S. Huang, iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC. Bioinformatics doi: 10.1093/bioinformatics/bty312/4978052 (2018).
[32] X. Cheng, X. Xiao, pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics 110 (2018) 231-239.
[33] B. Liu, F. Yang, D.S. Huang, iPromoter-2L: a two-layer predictor for identifying

promoters and their types by multi-window-based PseKNC. Bioinformatics 34 (2018) 33-40.
[34] X. Cheng, X. Xiao, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics 34 (2018) 1448-1456.
[35] H. Yang, W.R. Qiu, G. Liu, F.B. Guo, W. Chen, H. Lin, iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC International Journal of Biological Sciences 14 (2018) 883-891.
[36] X. Xuao, X. Cheng, G. Chen, Q. Mao, pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics doi:10.1016/j.ygeno.2018.05.017 (2018).
[37] L. Cai, T. Huang, J. Su, X. Zhang, W. Chen, F. Zhang, L. He, Implications of newly identified brain eQTL genes and their interactors in Schizophrenia. Molecular Therapy - Nucleic Acids 12 (2018) 433-442.
[38] W. Chen, H. Ding, X. Zhou, H. Lin, iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition. Analytical Biochemistry doi:10.1016/j.ab.2018.09.002 (2018).
[39] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.-M. Kim, D. Eisenberg, DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research 30 (2002) 303-305.
[40] S. Martin, D. Roe, J.L. Faulon, Predicting protein-protein interactions using signature products. Bioinformatics 21 (2005) 218-226.
[41] J.R. Bock, D.A. Gough, Whole-proteome interaction mining. Bioinformatics 19 (2003) 125-134.
[42] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Research 36 (2008) 3025-3030.
[43] L. Nanni, Hyperplanes for predicting protein-protein interactions. Neurocomputing 69 (2005) 257-263.
[44] L. Nanni, A. Lumini, An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics 22 (2006) 1207-1210.
[45] J.-F. Xia, K. Han, D.-S. Huang, Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. Protein and Peptide Letters 17 (2010) 137-145.
[46] K.C. Chou, D.W. Elrod, Bioinformatical analysis of G-protein-coupled receptors. Journal of Proteome Research 1 (2002) 429-433.
[47] W. Chen, H. Lin, P.M. Feng, C. Ding, Y.C. Zuo, iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. PLoS ONE 7 (2012) e47843.
[48] Y.D. Cai, Predicting subcellular localization of proteins in a hybridization space. Bioinformatics 20 (2004) 1151-6.
[49] K.C. Chou, Y.D. Cai, Prediction of protease types in a hybridization space. Biochem Biophys Res Comm (BBRC) 339 (2006) 1015-1020.
[50] P.M. Feng, W. Chen, H. Lin, iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical Biochemistry 442

(2013) 118-25.
[51] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition Nucleic Acids Research 41 (2013) e68.
[52] W.Z. Lin, J.A. Fang, X. Xiao, iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE 6 (2011) e24756.
[53] J. Jia, Z. Liu, X. Xiao, B. Liu, pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of Theoretical Biology 394 (2016) 223-230.
[54] K.C. Chou, Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry 11 (2015) 218-234.
[55] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43 (2001) 246-255.
[56] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21 (2005) 10-19.
[57] K.C. Chou, Y.D. Cai, A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochemical and Biophysical Research Communications (BBRC) 311 (2003) 743-747.
[58] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology 248 (2007) 546–551.
[59] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34 (2008) 653-660.
[60] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. Journal of Theoretical Biology 257 (2009) 17-26.
[61] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263 (2010) 203-209.
[62] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17 (2010) 1207-1214.
[63] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 34 (2010) 320-327.
[64] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo amino acid composition and support vector machine. Journal of Theoretical Biology 281 (2011) 18-23.
[65] B.M. Mohammad, M. Behjati, H. Mohabatkar, Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. Journal of Structural and Functional Genomics 12 (2011) 191-197.
[66] M. Hayat, A. Khan, Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. Protein & Peptide Letters 19 (2012) 411-421.
[67] S. Mei, Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. Journal of Theoretical

Biology 310 (2012) 80-87.
[68] L. Nanni, S. Brahnam, A. Lumini, Wavelet images and Chou's pseudo amino acid composition for protein classification. Amino Acids 43 (2012) 657-65.
[69] M.K. Gupta, R. Niyogi, M. Misra, An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition. SAR QSAR Environ Res 24 (2013) 597-609.
[70] M. Khosravian, F.K. Faramarzi, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting Antibacterial Peptides by the Concept of Chou's Pseudo amino Acid Composition and Machine Learning Methods. Protein & Peptide Letters 20 (2013) 180-186.
[71] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology 341 (2014) 34-40.
[72] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, N.Y. Deng, iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE 9 (2014) e105018.
[73] M. Hayat, N. Iqbal, Discriminating protein structure classes by incorporating Pseudo Average Chemical Shift to Chou's general PseAAC and Support Vector Machine. Comput Methods Programs Biomed 116 (2014) 184-92.
[74] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J Theor Biol 356 (2014) 30-5.
[75] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. J Theor Biol 360 (2014) 109-116.
[76] S. Ahmad, M. Kabir, M. Hayat, Identification of Heat Shock Protein families and J-protein types by incorporating Dipeptide Composition into Chou's general PseAAC. Comput Methods Programs Biomed 122 (2015) 165-74.
[77] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo amino acid composition and support vector machine. J Theor Biol 365 (2015) 96-103.
[78] M. Behbahani, H. Mohabatkar, M. Nosrati, Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. J Theor Biol 411 (2016) 1-5.
[79] M. Rahimi, M.R. Bakhtiarizadeh, A. Mohammadi-Sangcheshmeh, OOgenesis_Pred: A sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition. J Theor Biol 414 (2017) 128-136.
[80] M. Tahir, M. Hayat, M. Kabir, Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition. Comput Methods Programs Biomed 146 (2017) 69-75.
[81] P. Tripathi, P.N. Pandey, A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. J Theor Biol 424 (2017) 49-54.
[82] S. Akbar, M. Hayat, iMethyl-STTNC: Identification of N(6)-methyladenosine sites by extending the Idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol 455 (2018) 205-211.
[83] M.A. Al Maruf, S. Shatabda, iRSpot-SF: Prediction of recombination hotspots by

incorporating sequence based features into Chou's Pseudo components. Genomics doi:10.1016/j.ygeno.2018.06.003 (2018).
[84] M. Arif, M. Hayat, Z. Jan, iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition. J Theor Biol 442 (2018) 11-21.
[85] E. Contreras-Torres, Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC. J Theor Biol 454 (2018) 139-145.
[86] J. Mei, J. Zhao, Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. Sci Rep 8 (2018) 2359.
[87] M. Mousavizadegan, H. Mohabatkar, Computational prediction of antifungal peptides via Chou's PseAAC and SVM. J Bioinform Comput Biol (2018) 1850016.
[88] J. Mei, J. Zhao, Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features. J Theor Biol 427 (2018) 147-153.
[89] Z. Ju, S.Y. Wang, Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. Gene 664 (2018) 78-83.
[90] F. Javed, M. Hayat, Predicting subcellular localizations of multi-label proteins by incorporating the sequence features into Chou's PseAAC. Genomics doi:10.1016/j.ygeno.2018.09.004 (2018).
[91] M.S. Krishnan, Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. J Theor Biol 445 (2018) 62-74.
[92] L. Zhang, L. Kong, iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components. J Theor Biol 441 (2018) 1-8.
[93] S. Zhang, X. Duan, Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J Theor Biol 437 (2018) 239-250.
[94] S.M. Rahman, S. Shatabda, S. Saha, M. Kaykobad, M. Sohel Rahman, DPP-PseAAC: A DNA-binding Protein Prediction model using Chou's general PseAAC. J Theor Biol 452 (2018) 22-34.
[95] A. Srivastava, R. Kumar, M. Kumar, BlaPred: predicting and classifying beta-lactamase using a 3-tier prediction system via Chou's general PseAAC. J Theor Biol 10.1016/j.jtbi.2018.08.030 (2018).
[96] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Current Topics in Medicinal Chemistry 17 (2017) 2337-2358.
[97] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions. Analytical Biochemistry 425 (2012) 117-119.
[98] D.S. Cao, Q.S. Xu, Y.Z. Liang, propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29 (2013) 960-962.
[99] P. Du, S. Gu, Y. Jiao, PseAAC-General: Fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets. International Journal of Molecular Sciences 15 (2014) 3495-3506.
[100] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics,

proteomics and system biology. Current Proteomics 6 (2009) 262-274.
[101] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, K.C. Chou, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Analytical Biochemistry 456 (2014) 53-60.
[102] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Research International (BMRI) 2014 (2014) 623149.
[103] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst 11 (2015) 2620-2634.
[104] W. Chen, H. Tang, J. Ye, H. Lin, iRNA-PseU: Identifying RNA pseudouridine sites Molecular Therapy - Nucleic Acids 5 (2016) e332.
[105] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32 (2016) 362-369.
[106] B. Liu, R. Long, iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics 32 (2016) 2411-2418.
[107] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research 43 (2015) W65-W71.
[108] B. Liu, H. Wu, Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein Sequences. Natural Science 9 (2017) 67-91.
[109] P. Deschavanne, P. Tuffery, Exploring an alignment free approach for protein classification and structural class prediction. Biochimie 90 (2008) 615-625.
[110] K.C. Chou, S.P. Jiang, W.M. Liu, C.H. Fee, Graph theory of enzyme kinetics: 1. Steady-state reaction system. Scientia Sinica 22 (1979) 341-358.
[111] K.C. Chou, S. Forsen, Graphical rules for enzyme-catalyzed rate laws. Biochemical Journal 187 (1980) 829-835.
[112] K.C. Chou, Graphic rules in steady and non-steady enzyme kinetics. Journal of Biological Chemistry 264 (1989) 12074-12079.
[113] G.P. Zhou, M.H. Deng, An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochemical Journal 222 (1984) 169-176.
[114] K.C. Chou, Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry 35 (1990) 1-24.
[115] K.C. Chou, H.B. Shen, FoldRate: A web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal 3 (2009) 31-50
[116] I.W. Althaus, J.J. Chou, A.J. Gonzales, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. Journal of Biological Chemistry 268 (1993) 6119-6124.
[117] I.W. Althaus, A.J. Gonzales, J.J. Chou, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. Journal of Biological Chemistry 268 (1993) 14875-14880.
[118] I.W. Althaus, J.J. Chou, A.J. Gonzales, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff,

W.G. Tarpley, F. Reusser, Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32 (1993) 6548-6554.
[119] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, Using cellular automata to generate Image representation for biological sequences. Amino Acids 28 (2005) 29-35.
[120] K.C. Chou, Graphic rule for drug metabolism systems. Current Drug Metabolism 11 (2010) 369-378.
[121] X. Xiao, S.H. Shao, A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Comm (BBRC) 342 (2006) 605-610.
[122] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation. Journal of Theoretical Biology 235 (2005) 555-565.
[123] Z.C. Wu, X. Xiao, 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. Journal of Theoretical Biology 267 (2010) 29-34.
[124] G.P. Zhou, The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. Journal of Theoretical Biology 284 (2011) 142-148.
[125] G.P. Zhou, R.B. Huang, The pH-Triggered Conversion of the PrP(c) to PrP(sc.). Curr Top Med Chem 13 (2013) 1152-63.
[126] K.K. Kandaswamy, T. Martinetz, S. Moller, P.N. Suganthan, S. Sridharan, G. Pugalenthi, AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology 270 (2011) 56-62.
[127] G. Pugalenthi, K.K. Kandaswamy, S. Vivekanandan, P. Kolatkar, RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method. Protein & Peptide Letters 19 (2012) 50-56.
[128] Y. Xu, J. Ding, L.Y. Wu, iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition PLoS ONE 8 (2013) e55844.
[129] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget 7 (2016) 44310-44321.
[130] J. Jia, Z. Liu, X. Xiao, B. Liu, iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget 7 (2016) 34558-34570.
[131] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32 (2016) 3116-3123.
[132] X. Xiao, H.X. Ye, Z. Liu, J.H. Jia, iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget 7 (2016) 34180-34189.
[133] W.R. Qiu, X. Xiao, Z.C. Xu, iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget 7 (2016) 51270-51283.
[134] W.R. Qiu, B.Q. Sun, X. Xiao, D. Xu, iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Molecular Informatics 36 (2017) UNSP 1600010.

[135] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Machine learning: An artificial intelligence approach, Springer Science & Business Media, 2013.
[136] C.T. Zhang, Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophysical Journal 63 (1992) 1523-1529.
[137] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry 268 (1993) 16938-16948.
[138] C.T. Zhang, An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. Journal of Protein Chemistry 14 (1995) 583-593.
[139] Y.D. Cai, K.Y. Feng, W.C. Lu, Using LogitBoost classifier to predict protein structural classes. Journal of Theoretical Biology 238 (2006) 172-176.
[140] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16 (2011) 321-357.
[141] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J Biomol Struct Dyn (JBSD) 33 (2015) 2221-2233.
[142] Z. Liu, X. Xiao, W.R. Qiu, iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical Biochemistry 474 (2015) 69-77.
[143] J. Jia, Z. Liu, X. Xiao, B. Liu, iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem 497 (2016) 48-56.
[144] J. Jia, Z. Liu, X. Xiao, B. Liu, iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules 21 (2016) E95.
[145] X. Cheng, W.Z. Lin, X. Xiao, pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinformatics doi:10.1093/bioinformatics/bty628 (2018).
[146] K.C. Chou, X. Cheng, X. Xiao, pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset Genomics doi:10.1016/j.ygeno.2018.08.007 (2018).
[147] X. Cheng, X. Xiao, pLoc_bal-mGneg: predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC. Journal of Theoretical Biology doi:10.1016/j/jtbi.2018.09.005 (2018).
[148] K.C. Chou, H.B. Shen, Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370 (2007) 1-16.
[149] J. Chen, H. Liu, J. Yang, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33 (2007) 423-428.
[150] K.C. Chou, Using subsite coupling to predict signal peptides. Protein Engineering 14 (2001) 75-79.
[151] K.C. Chou, Prediction of signal peptides using scaled window. Peptides 22 (2001) 1973-1979.
[152] Y. Xu, X.J. Shao, L.Y. Wu, N.Y. Deng, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1 (2013) e171.
[153] H. Lin, E.Z. Deng, H. Ding, W. Chen, K.C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide

composition. Nucleic Acids Research 42 (2014) 12961-12972.
[154] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, H. Lin, W. Chen, iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Research International (BMRI) 2014 (2014) 286419.
[155] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS ONE 9 (2014) e106691.
[156] Y. Xu, X. Wen, X.J. Shao, N.Y. Deng, iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. International Journal of Molecular Sciences (IJMS) 15 (2014) 7594-7610.
[157] W.R. Qiu, X. Xiao, W.Z. Lin, iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. Biomed Res Int (BMRI) 2014 (2014) 947416.
[158] Y.N. Fan, X. Xiao, J.L. Min, iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. Intenational Journal of Molecular Sciences (IJMS) 15 (2014) 4915-4937.
[159] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30 (2014) 1522-1529.
[160] W.R. Qiu, X. Xiao, iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci (IJMS) 15 (2014) 1746-1766.
[161] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Analytical Biochemistry 462 (2014) 76-83.
[162] W.R. Qiu, X. Xiao, W.Z. Lin, iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model Journal of Biomolecular Structure and Dynamics (JBSD) 33 (2015) 1731-1742.
[163] W. Chen, P. Feng, H. Ding, H. Lin, iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Analytical Biochemistry 490 (2015) 26-33.
[164] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE 10 (2015) e0121501.
[165] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. Journal of Theoretical Biology 385 (2015) 153-159.
[166] R. Xu, J. Zhou, B. Liu, Y.A. He, Q. Zou, X. Wang, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. Journal of Biomolecular Structure & Dynamics (JBSD) 33 (2015) 1720-1730.
[167] W. Chen, P. Feng, H. Ding, H. Lin, Using deformation energy to analyze nucleosome positioning in genomes. Genomics 107 (2016) 69-75.
[168] J. Jia, L. Zhang, Z. Liu, X. Xiao, pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics 32 (2016) 3133-3141.
[169] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, pRNAm-PC: Predicting N-methyladenosine sites

in RNA sequences via physical-chemical properties. Anal Biochem 497 (2016) 60-67.
[170] C.J. Zhang, H. Tang, W.C. Li, H. Lin, W. Chen, iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget 7 (2016) 69783-69793.
[171] L. Cai, W. Yuan, Z. Zhang, L. He, In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data Scientific Reports 6 (2016) 36540.
[172] B. Liu, L. Fang, F. Liu, X. Wang, iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J Biomol Struct Dyn (JBSD) 34 (2016) 223-35.
[173] W. Chen, H. Ding, P. Feng, H. Lin, iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 7 (2016) 16895-16909.
[174] W.R. Qiu, S.Y. Jiang, Z.C. Xu, X. Xiao, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget 8 (2017) 41178-41188.
[175] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget 8 (2017) 4208-4217.
[176] W.R. Qiu, S.Y. Jiang, B.Q. Sun, X. Xiao, X. Cheng, iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Medicinal Chemistry 13 (2017) 734-743.
[177] Y. Xu, C. Li, iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med Chem 13 (2017) 544-551.
[178] L.M. Liu, Y. Xu, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med Chem 13 (2017) 552-559.
[179] A. Ehsan, K. Mahmood, Y.D. Khan, S.A. Khan, A Novel Modeling in Mathematical Biology for Classification of Signal Peptides. Scientific Reports 8 (2018) 1039.
[180] X. Xiao, X. Cheng, S. Su, Q. Nao, pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Natural Science 9 (2017) 331-349.
[181] X. Cheng, S.G. Zhao, X. Xiao, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget 8 (2017) 58494-58503.
[182] K.C. Chou, Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. Molecular Biosystems 9 (2013) 1092-1100.
[183] K.C. Chou, C.T. Zhang, Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30 (1995) 275-349.
[184] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes. Natural Science 1 (2009) 63-92
[185] J. Wang, B. Yang, J. Revote, A. Leier, T.T. Marquez-Lago, G. Webb, J. Song, T. Lithgow, POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics 33 (2017) 2756-2758.
[186] Z. Chen, P.Y. Zhao, F. Li, Leier A, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, J. Song, iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34 (2018) 2499-2502.

[187] J. Song, F. Li, A. Leier, T.T. Marquez-Lago, T. Akutsu, G. Haffari, G.I. Webb, R.N. Pike, PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics 34 (2018) 684-687.
[188] J. Wang, B. Yang, A. Leier, T.T. Marquez-Lago, M. Hayashida, A. Rocker, Z. Yanju, T. Akutsu, R.A. Strugnell, J. Song, T. Lithgow, Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. Bioinformatics 34 (2018) 2546-2555.