OPEN ACCESS Freely available online

PLOS ONE

Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model

Wei-Zhong Lin[1,2], Jian-An Fang[1], Xuan Xiao[2,3]*, Kuo-Chen Chou[3]*

1 Information Science and technology School, Donghua University, Shanghai, China, **2** Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China, **3** Gordon Life Science Institute, San Diego, California, United States of America

Abstract

The malaria disease has become a cause of poverty and a major hindrance to economic development. The culprit of the disease is the parasite, which secretes an array of proteins within the host erythrocyte to facilitate its own survival. Accordingly, the secretory proteins of malaria parasite have become a logical target for drug design against malaria. Unfortunately, with the increasing resistance to the drugs thus developed, the situation has become more complicated. To cope with the drug resistance problem, one strategy is to timely identify the secreted proteins by malaria parasite, which can serve as potential drug targets. However, it is both expensive and time-consuming to identify the secretory proteins of malaria parasite by experiments alone. To expedite the process for developing effective drugs against malaria, a computational predictor called "**iSMP-Grey**" was developed that can be used to identify the secretory proteins of malaria parasite based on the protein sequence information alone. During the prediction process a protein sample was formulated with a 60D (dimensional) feature vector formed by incorporating the sequence evolution information into the general form of PseAAC (pseudo amino acid composition) via a grey system model, which is particularly useful for solving complicated problems that are lack of sufficient information or need to process uncertain information. It was observed by the jackknife test that **iSMP-Grey** achieved an overall success rate of 94.8%, remarkably higher than those by the existing predictors in this area. As a user-friendly web-server, **iSMP-Grey** is freely accessible to the public at http://www.jci-bioinfo.cn/iSMP-Grey. Moreover, for the convenience of most experimental scientists, a step-by-step guide is provided on how to use the web-server to get the desired results without the need to follow the complicated mathematical equations involved in this paper.

Citation: Lin W-Z, Fang J-A, Xiao X, Chou K-C (2012) Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model. PLoS ONE 7(11): e49040. doi:10.1371/journal.pone.0049040

Editor: Claudio Romero Farias Marinho, Instituto de Ciências Biomédicas/Universidade de São Paulo - USP, Brazil

Received August 22, 2012; **Accepted** October 3, 2012; **Published** November 26, 2012

Funding: This work was supported by the grants from the National Natural Science Foundation of China (No. 60961003, 31260273), the Key Project of Chinese Ministry of Education (No. 210116), and the department of education of Jiang-Xi Province (No. GJJ11557, GJJ12490), and the Jiangxi Provincial Foundation for Leaders of Disciplines in Science (No. 20113BCB22008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: xxiao@gordonlifescience.org (XX); kcchou@gordonlifescience.org (KCC)

Introduction

Malaria is a potentially fatal tropical disease caused by a parasite known as Plasmodium. Four distinct species of plasmodium that can produce the disease in different forms: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, and *Plasmodium malaria*. Of these four, *Plasmodium falciparum*, or *P. falciparum*, is the most widespread and dangerous. If not timely treated, it may lead to the fatal cerebral malaria, which remains one of the most devastating global health crises. Nearly half of the world's population is still at risk from its infection. According to the World Health Organization's 2010 World Malaria Report (http://www.who.int/malaria/world_malaria_report_2010/worldmalariareport2010.pdf), there are more than 225 million cases of malaria each year, killing around 781,000 people, corresponding to 2.23% of deaths worldwide. Malaria is more dangerous for women and children. It was stated in the World Health Organization's 2011 World Malaria Report (http://www.who.int/malaria/world_malaria_report_2011/9789241564403_eng.pdf) that 81% of cases and 91% of deaths occurred in the African Region, mostly involving children under five and women with pregnancy. Malaria was usually associated with poverty; actually it was a cause of poverty and a major hindrance for economic development. The situation has become even worse over the last few years with the increase in resistance to the drugs normally used to combat the parasites that cause the disease. Therefore, one strategy to deal with the growing malaria problem is to identify and characterize new and durable antimalarial drug targets, the majority of which are parasite proteins [1]. Parasite secretes an array of proteins within the host erythrocyte to facilitate its own survival within the host cell. These proteins can serve as potential drug or vaccine targets. However, it is difficult to experimentally identify the secretory proteins of *P. falciparum* owing to the complex nature of parasite. With the completion of *Plasmodium* genome sequence, it is both challenging and urgent to develop an automatic method or high throughput tool for identifying secretory proteins of *P. falciparum*.

Actually, some efforts have been made in this regard. In a pioneer study, Verma et al. [2] proposed a method for identifying proteins secreted by malaria parasite. In their prediction method, the operation engine was the Support Vector Machine (SVM)

while the protein samples were formulated with the amino acid composition, dipeptide composition, and position specific scoring matrix (PSSM) [3]. Subsequently, Zuo and Li [4] introduced the K-minimum increment of diversity (K-MID) approach to predict secretory proteins of malaria parasite based on grouping of amino acids. Meanwhile, various studies around this topic were also carried out [5,6,7,8,9].

In the past, various predictors for protein systems were developed by incorporating the evolutionary information via PSSM [10,11,12,13,14,15,16,17,18,19,20]. In the above papers, however, only the statistical information of PSSM [3] was utilized but the inner interactions among the constituent amino acid residues in a protein sample, or its sequence-order effects, were ignored.

To avoid completely lose the sequence-order information associated with PSSM, the concept of pseudo amino acid composition (PseAAC) [21,22] was utilized to incorporate the evolutionary information into the formulation of a protein sample, as done in predicting protein subcellular localization [23,24,25], predicting protein fold pattern [26], identifying membrane proteins and their types [27], predicting enzyme functional classes and subclasses [28], identifying protein quaternary structural attribute [29], predicting antibacterial peptides [30], predicting allergenic proteins [31], and identifying proteases and their types [32].

The present study was initiated in an attempt to develop a new and more powerful predictor for identifying the secretory proteins of malaria parasite by incorporating the sequence evolution information into PseAAC via a grey system model [33].

According to a recent review [34], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

Materials and Methods

1. Benchmark Dataset

The benchmark dataset $\mathbb{S}^{\text{Bench}}$ used in this study was taken from Verma et al. [2]. The dataset can be formulated as

$$\mathbb{S}^{\text{Bench}}=\mathbb{S}^{+}\bigcup\mathbb{S}^{-} \tag{1}$$

where $\mathbb{S}^{+}$ contains 252 secretory proteins of malaria parasite, $\mathbb{S}^{-}$ contains 252 non-secretory proteins of malaria parasite, and the symbol $\bigcup$ represents the union in the set theory. The same benchmark dataset was also used by Zuo and Li [4]. For reader's convenience, the sequences of the 252 secretory proteins in $\mathbb{S}^{+}$ and those in $\mathbb{S}^{-}$ are given in Supporting Information S1.

2. A Novel PseAAC Feature Vector by Incorporating Sequence Evolution Information via the Grey System Theory

To develop a powerful predictor for a protein system, one of the keys is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted [34]. To realize this, the pseudo amino acid composition (PseAAC) was proposed [21] to replace the simple amino acid composition (AAC) for representing the sample of a protein. Ever since the concept of PseAAC was introduced in 2001 [21], it has penetrated into almost all the fields of protein attribute predictions, such as predicting protein submitochondrial localization [35], predicting protein structural class [36], predicting DNA-binding proteins [37], identifying bacterial virulent proteins [38], predicting metalloproteinase family [39], predicting protein folding rate [40], predicting GABA(A) receptor proteins [41], predicting protein supersecondary structure [42], identifying protein quaternary structural attribute [43], predicting cyclin proteins [44], classifying amino acids [45], predicting enzyme family class [46], identifying risk type of human papillomaviruses [47], and discriminating outer membrane proteins [48], among many others (see a long list of references cited in [49]). Because it has been widely used, recently a powerful software called PseAAC-Builder [49] was proposed for generating various special modes of PseAAC, in addition to the web-server PseAAC [50] established in 2008.

According to a recent review [34], the general form of PseAAC for a protein $\mathbf{P}$ can be formulated as

$$\mathbf{P}=[\,\psi_1\quad\psi_2\quad\cdots\quad\psi_u\quad\cdots\quad\psi_\Omega\,]^{\mathbf{T}} \tag{2}$$

where $\mathbf{T}$ is a transpose operator, while 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 $\mathbf{P}$.

The form of **Eq.2** can cover almost all the various modes of PseAAC. Particularly, it can be used to reflect much more essential core features deeply hidden in complicated protein sequences, such as those for the functional domain (FunD) information [51,52,53] (cf. Eqs.9–10 of [34]), gene ontology (GO) information [54,55] (cf. Eqs.11–12 of [34]), and sequence evolution information [3] (cf. Eqs.13–14 of [34]).

In this study, we are to use a novel approach to define the Ω elements in **Eq.2**. As is well known, biology is a natural science with historic dimension. All biological species have developed starting out from a very limited number of ancestral species. It is true for protein sequence as well [56]. Their evolution involves changes of single residues, insertions and deletions of several residues [57], gene doubling, and gene fusion. With these changes accumulated for a long period of time, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share many common attributes, such as having basically the same biological function and residing at a same subcellular location. To incorporate this kind of sequence evolution information into the PseAAC of **Eq.2**, let us use the information of the PSSM (Position-Specific Scoring Matrix) [3], as described below.

According to [3], the sequence evolution information of protein $\mathbf{P}$ with L amino acid residues can be expressed by a $20\times L$ matrix, as given by

$$\mathbf{P}^{(0)}_{\text{PSSM}}=\begin{bmatrix} m^{(0)}_{1,1} & m^{(0)}_{1,2} & \cdots & m^{(0)}_{1,20} \\ m^{(0)}_{2,1} & m^{(0)}_{2,2} & \cdots & m^{(0)}_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ m^{(0)}_{L,1} & m^{(0)}_{L,2} & \cdots & m^{(0)}_{L,20} \end{bmatrix} \tag{3}$$

where $m^{(0)}_{i,j}$ represents the original score of amino acid residue in

the i-th ($i=1,2,\cdots,L$) sequential position of the protein that is being changed to amino acid type j ($j=1,2,\cdots,20$) during the evolution process. Here, the numerical codes 1, 2, …, 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes [58]. The $20\times L$ scores in **Eq.3** were generated by using PSI-BLAST [3] to search the UniProtKB/Swiss-Prot database (Release 2010_04 of 23-Mar-2010) through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of the protein **P**. In order to make every element in **Eq.3** within the range of 0–1, a conversion was performed through the standard sigmoid function to make it become

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

where

$$\mathrm{m}^{(1)}_{\mathrm{i,j}}=\frac{1}{1+\mathrm{e}^{-\mathrm{m}^{(0)}_{\mathrm{i,j}}}} \qquad (1\le \mathrm{i}\le \mathrm{L},\quad 1\le \mathrm{j}\le 20) \tag{5}$$

Now, let us describe how to extract the useful information from **Eq.4** via a grey system model. According to the grey system theory [33], if the information of a system investigated is fully known, it is called a "white system"; if completely unknown, a "black system"; if partially known, a "grey system". The model developed based on such a theory is called "grey model", which is a kind of nonlinear and dynamic model formulated by a differential equation. The grey model is particularly useful for solving complicated problems that are lack of sufficient information, or need to process uncertain information and reduce random effects of acquired data. In the grey system theory, an important and generally used model is called GM(1,1) [33]. It is quite effective for monotonic series, with good simulating effect and small error, as reflected by the fact that using the GM(1,1) model has remarkably improved the success rates in predicting protein structural classes [59]. However, if the series concerned are not monotonic, the simulating effect of the GM(1,1) model would not be good and its error might be quite large. To overcome such a shortcoming, in this study we are to use a different grey system model called GM(2,1) [33], which can be effectively used to deal with the oscillation series.

To extract the serial information of **Eq.4**, let us consider the L components in its j-th column, i.e., $\left(m^{(1)}_{1,j}\quad m^{(1)}_{2,j}\quad \cdots\quad m^{(1)}_{L,j}\right)$, as an initial series. Obviously, the j-th column of the **Eq.4** is an oscillation series but not monotonic as in the case investigated in [59]. To deal with such a problem, instead of the GM(1,1), let us adopt the GM(2,1) model here. According to the GM(2,1) model [33], we have the following 2nd-order grey differential equation with one variable:

$$\begin{aligned} &\alpha^{(1)}m^{(1)}_{k,j}+a^j_1 m^{(1)}_{k,j}+a^j_2 z^{(1)}(k)=b^j \\ &(k=2,3,\cdots,L;\quad j=1,2,\ \cdots,20) \end{aligned} \tag{6}$$

where

$$\alpha^{(1)}m^{(1)}_{k,j}=m^{(1)}_{k,j}-m^{(1)}_{k-1,j} \tag{7}$$

and

$$z^{(1)}(k)=\sum_{i=1}^{k-1} m^{(1)}_{i,j}+0.5m^{(1)}_{k,j} \tag{8}$$

In **Eq.6**, the coefficients a^j_1 and a^j_2 are associated with the developing coefficients, and b^j the influence coefficient. Actually, a^j_1, a^j_2, and b^j can be expressed as the components of a 3D vector as given by

$$\mathbf{H}^j=\begin{bmatrix} a^j_1 & a^j_2 & b^j \end{bmatrix}^{\mathbf{T}} \quad (j=1,2,\cdots,20) \tag{9}$$

in which the components a^j_1, a^j_2, and b^j can be directly derived from the following equation

$$\mathbf{H}^j=(\mathbf{B}^{\mathbf{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathbf{T}}\mathbf{U} \tag{10}$$

where

$$\mathbf{B}=\begin{bmatrix} -m^{(1)}_{2,j} & -z^{(1)}(2) & 1 \\ -m^{(1)}_{3,j} & -z^{(1)}(3) & 1 \\ \vdots & \vdots & \vdots \\ -m^{(1)}_{L,j} & -z^{(1)}(L) & 1 \end{bmatrix} \tag{11}$$

and

$$\mathbf{U}=\begin{bmatrix} \alpha^{(1)}m^{(1)}_{2,j} \\ \alpha^{(1)}m^{(1)}_{3,j} \\ \vdots \\ \alpha^{(1)}m^{(1)}_{L,j} \end{bmatrix} \tag{12}$$

Accordingly, the Ω elements in **Eq.2** are given by

$$\begin{cases} \psi_{3j-2}=a^j_1 f_j w_1 \\ \psi_{3j-1}=a^j_2 f_j w_2 \quad (j=1,2,\cdots,20) \\ \psi_{3j}\ \ =b^j f_j w_3 \end{cases} \tag{13}$$

where f_i ($i=1,2,\cdots,20$) are the occurrence frequencies of the 20 different types of amino acids in the protein sample concerned, and w_1, w_2, and w_3 are the weight factors that will be determined by optimizing the performance of the predictor, and their concrete values will be explicitly given in the footnote of **Table 1**. Substituting **Eq.13** into **Eq.2**, we immediately obtain a feature vector with $\Omega=3\times 20=60$ components. The 60D feature vector thus derived will be used to represent the samples of protein sequences for further study.

3. The SVM Operation Engine

In this study, the Support Vector Machine (SVM) algorithm was adopted to perform the prediction. The SVM software was implemented from the LIBSVM package [60]. The software thus obtained provided a simple interface by which the users can easily

Table 1. A comparison between iSMP-Grey and K-MID by the jackknife test.

Predictor	Sn (%)	Sp (%)	Acc (%)	MCC
iSMP-Grey[a]	93.25	96.46.	94.84	0.90
K-MID[b]	81.75	99.60	90.67	0.83

[a]The parameters used: $w_1 = 800$, $w_2 = 35$, and $w_3 = 800$ for **Eq.14**; $c = 8$ and $g = 0.00012$ for the LIBSVM operation engine.
[b]From ref.[4].
doi:10.1371/journal.pone.0049040.t001

perform classification prediction by properly selecting the built-in parameters c and g. In this study we searched the optimal parameters c and g by the grid arithmetic built in the LIBSVM software, and their optimal values are also explicitly given in the footnote of **Table 1**. Meanwhile, the MATLAB windows were adopted in developing the classifier.

The predictor thus established is called **iSMP-Grey**, which can be used to identify whether a protein of malaria parasite is secretory or non-secretory according to its sequence information alone.

4. Web-Server and User Guide

To enhance the value of its practical applications, a web-server for **iSMP-Grey** was established. Moreover, for the convenience of the vast majority of experimental scientists, here let us provide a step-by-step guide to show how the users can easily get the desired result by means of the web-server without the need to follow the above mathematical equations for its development and integrity.

Step 1. Open the web server at the site http://www.jci-bioinfo.cn/iSMP-Grey and you will see the top page of the predictor on your computer screen, as shown in **Fig. 1**. Click on the Read Me button to see a brief introduction about **iSMP-Grey** predictor and the caveat when using it.

Step 2. Either type or copy and paste the query protein sequence into the input box at the center of **Fig. 1**. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol ("$>$") in the first column, followed by lines of sequence data. The words right after the "$>$" symbol in the single initial line are optional and only used for the purpose of identification and description. The sequence ends if another line starting with a "$>$" appears; this indicates the start of another sequence. The example sequences in FASTA format can be seen by clicking on the Example button right above the input box. The maximum number of query protein sequences allowed for each submission is 10.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the two query peptide sequences in the Example window as the input, about 2–3 minutes after clicking the Submit button, you will see on your screen that the 1st query protein is a "**Secretory Protein of Malaria Parasite**", and that the 2nd query protein 2 is "**Non-Secretory Protein of Malaria parasite**". All these results are fully consistent with the experimental observations.

Step 4. Click on the Citation button to find the relevant paper that documents the detailed development and algorithm of **iSMP-Grey**.

Step 5. Click on the Data button to download the benchmark dataset used to train and test the **iSMP-Grey** predictor.

Step 6. The program is also available by clicking the button download on the lower panel of **Fig. 1**.

5. Performance Evaluation

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling (K-fold cross-validation) test, and jackknife test. However, as elaborated by a recent review [34] and demonstrated by Eqs.28–32 therein, among the three cross-validation methods, the jackknife test is deemed the least arbitrary and most objective because it can always yield a unique result for a given benchmark dataset, and hence has been widely recognized and increasingly used by investigators for examining the accuracy of various predictors (see, e.g., [36,38,39,41,44,47,61,62,63,64,65,66]). Accordingly, the jackknife test was also adopted in this study to examine the anticipated success rates of the current predictor.

Also, to use a more intuitive and easier-to-understand method to measure the prediction quality, the rates of correct predictions for the secretory proteins of malaria parasite in dataset $\sum^+$ and the non-secretory proteins of malaria parasite in dataset $\sum^-$ are respectively defined by [67]

$$\begin{cases} \Lambda^+ = \dfrac{N^+ - m^+}{N^+}, & \text{for the secretory proteins} \\ \Lambda^- = \dfrac{N^- - m^-}{N^-}, & \text{for the non-secretory proteins} \end{cases} \tag{14}$$

where N^+ is the total number of the secreted proteins investigated and m^+ the number of the secreted proteins missed in the predicted result; N^- the total number of the non-secreted proteins investigated and m^- the number of the non-secreted proteins missed in the predicted result. The overall success prediction rate is given by [68]

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{m^+ + m^-}{N^+ + N^-} \tag{15}$$

It is clear from **Eqs.14–15** that, if and only if none of the secreted proteins and non-secreted proteins are mispredicted, i.e., $m^+ = m^- = 0$ and $\Lambda^+ = \Lambda^- = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be smaller than 1. It is instructive to point out that the following equation is often used in literatures for examining the performance quality of a predictor

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{MCC} = \dfrac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{cases} \tag{16}$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient.

The relations between the symbols in **Eq.15** and those in **Eq.16** are given by

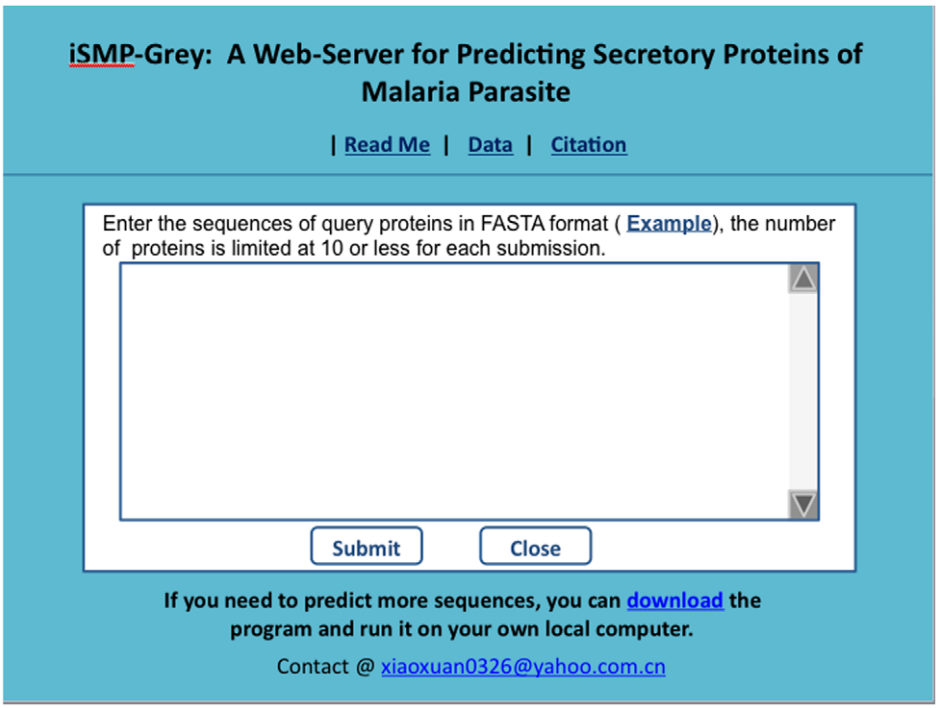

Figure 1. A semi-screenshot to show the top page of the iSMP-Grey web-server. Its web-site address is at http://www.jci-bioinfo.cn/iSMP-Grey.
doi:10.1371/journal.pone.0049040.g001

$$\begin{cases} \mathrm{TP} = N^+ - m^+ \\ \mathrm{TN} = N^- - m^- \\ \mathrm{FP} = m^- \\ \mathrm{FN} = m^+ \end{cases} \tag{17}$$

It follows by substituting **Eq.17** into **Eq.16** and noting **Eq.15**

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{m^+}{N^+} \\ \mathrm{Sp} = 1 - \dfrac{m^-}{N^-} \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{m^+ + m^-}{N^+ + N^-} \\ \mathrm{MCC} = \dfrac{1 - \left(\frac{m^+}{N^+} + \frac{m^-}{N^-}\right)}{\sqrt{\left(1 + \frac{m^- - m^+}{N^+}\right)\left(1 + \frac{m^+ - m^-}{N^-}\right)}} \end{cases} \tag{18}$$

As can be obviously seen from the above equation, when $m^+ = 0$ meaning none of the secreted proteins was missed in prediction, we have the sensitivity $\mathrm{Sn} = 1$; while $m^+ = N^+$ meaning all the secreted proteins were missed in prediction, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $m^- = 0$ meaning none of the non-secreted proteins was incorrectly predicted as secreted protein, we have the specificity $\mathrm{Sp} = 1$; while $m^- = N^-$ meaning all the non-secreted proteins were incorrectly predicted as secreted proteins, we have the specificity $\mathrm{Sp} = 0$. When $m^+ = m^- = 0$ meaning that none of the secreted proteins in the dataset $\mathbb{S}^+$ and non of non-secreted proteins in $\mathbb{S}^-$ was incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = \Lambda = 1$; while $m^+ = N^+$ and $m^- = N^-$ meaning that all the secreted proteins in the dataset $\mathbb{S}^+$ and all the non-secreted proteins in $\mathbb{S}^-$ were incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = \Lambda = 0$. The MCC correlation coefficient is usually used for measuring the quality of binary (two-class) classifications. When $m^+ = m^- = 0$ meaning that none of the secreted proteins in the dataset $\mathbb{S}^+$ and none of the non-secreted proteins in $\mathbb{S}^-$ was incorrectly predicted, we have $\mathrm{Mcc} = 1$; when $m^+ = N^+/2$ and $m^- = N^-/2$ we have $\mathrm{Mcc} = 0$ meaning no better than random prediction; when $m^+ = N^+$ and $m^- = N^-$ we have $\mathrm{MCC} = -1$ meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier-to-understand when using **Eq.18** to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient.

Results and Discussion

The results obtained with **iSMP-Grey** on the benchmark dataset $\mathbb{S}^{\mathrm{Bench}}$ of **Eq.1** by the jackknife test are given in **Table 1**, where for facilitating comparison the results obtained by the **K-MID** predictor [4] on the same benchmark dataset with the same test method are also given. As we can see from **Table 1**, the overall success rate by **iSMP-Grey** was 94.84% with $\mathrm{MCC} = 0.90$, which are remarkably higher than those by the K-MID predictor [4].

Moreover, a comparison was also made with the **PSEApred** predictor [2]. Although the results by **PSEApred** as reported by Verma et al. [2] were also based on the same benchmark dataset $\sum^{\mathrm{Bench}}$ of **Eq.1**, the test method used by these authors for **PSEApred** was 5-fold cross-validation. As elaborated in [34], this would make the test without a unique result as demonstrated below. For the current case, $\mathbb{S}^{\mathrm{Bench}}$ consists of $\mathbb{S}^+$ and $\mathbb{S}^-$, where

$\mathbb{S}^+$ contains 252 secretory proteins of malaria parasite, and $\mathbb{S}^-$ contains 252 non-secretory proteins of malaria parasite. Substituting these data into Eqs.28–29 of [34] with $M=2$ (number of groups for classification) and $\Gamma=5$ (number of folds for cross-validation), we obtain

$$\Pi=\frac{252!}{[252-\mathrm{Int}(252/5)]!\mathrm{Int}(252/5)!}\cdot\frac{252!}{[252-\mathrm{Int}(252/5)]!\mathrm{Int}(252/5)!}=\left[\frac{252!}{(252-50)!50!}\right]^2>9.25\times10^{128} \quad (19)$$

where the symbol Int is the integer-truncating operator meaning to take the integer part for the number in the bracket right after it. The result of **Eq.19** indicates that the number of possible combinations of taking one-fifth proteins from each of the two subsets, $\mathbb{S}^+$ and $\mathbb{S}^-$, for conducting the 5-fold cross-validation will be greater than 9.25×10^{128}, which is an astronomical figure, too large to be practically feasible. Actually, in their study [2], Verma et al. only randomly picked 100 different combinations from the possible 9.25×10^{128} combinations (cf. **Eq.19**) to perform the 5-fold cross-validation, yielding 100 different results located within a certain region. Therefore, in their report, rather than a single figure but a figures region was used to show their test result. For example, according to their report (**Table 2**), $\mathrm{Acc}=71.03\sim92.66\%$, meaning that the lowest one of the 100 overall success rates obtained by the **PSEApred** predictor [2] was 71.03%, while the highest one was 92.66%. To make the comparison of **iSMP-Grey** with **PSEApred** [2] under the same condition with the same test method, we also randomly picked 100 different combinations as done by Verma et al. [2] to perform the 5-fold cross-validation test with **iSMP-Grey**, and the corresponding results thus obtained are given in **Table 2** as well. As we can see from the table, not only the average rates obtained by the **iSMP-Grey** predictor are remarkably higher than those by the **PSEApred** predictor [2], but the corresponding region widths by the former are also significantly narrower than those by the latter, indicating the success rates by the **iSMP-Grey** are not only higher but also more stable than those by the **PSEApred** predictor [2].

Table 2. A comparison between iSMP-Grey and PSEApred by 5-fold cross-validation test.

Predictor	Sn (%)[c]	Sp (%)[c]	Acc (%)[c]	MCC[c]
iSMP-Grey[a]	90.48∼92.46	94.05∼98.02	92.86∼94.84	0.87∼0.90
PSEApred[b]	73.41∼97.22	44.84∼100	71.03∼92.66	0.49∼0.86

[a]See footnote a of **Table 1**.
[b]From ref. [2].
[c]See the discussion in the text and **Eq.19** for why the results obtained by the 5-fold cross-validation test were not unique.
doi:10.1371/journal.pone.0049040.t002

All the above results have indicated that the novel pseudo amino acid composition formulated via the grey system model GM(2,1) can more effectively incorporate the protein sequence evolution information so as to remarkably enhance the success rates of the **iSMP-Grey** predictor in identifying the secretory proteins of malaria parasite. It is anticipated that **iSMP-Grey** may become a useful high throughput tool for both basic research and drug development in the relevant areas.

Supporting Information

Supporting Information S1 The benchmark dataset $\mathbb{S}^{\mathrm{Bench}}$ includes 504 proteins, classified into 252 secretory proteins of malaria parasite and 252 non-secretory proteins.
(PDF)

Acknowledgments

The authors wish to thank the two anonymous Reviewers, whose constructive comments were very helpful for strengthening the presentation of this paper.

Author Contributions

Conceived and designed the experiments: WZL XX. Performed the experiments: WZL JAF. Analyzed the data: WZL XX KCC. Contributed reagents/materials/analysis tools: XX. Wrote the paper: WZL KCC.

References

1. Birkholtz LM, Blatch G, Coetzer TL, Hoppe HC, Human E, et al. (2008) Heterologous expression of plasmodial proteins for structural studies and functional annotation. Malaria Journal 7: 197.
2. Verma R, Tiwari A, Kaur S, Varshney GC, Raghava GP (2008) Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. BMC Bioinformatics 9: 201.
3. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994–3005.
4. Zuo YC, Li QZ (2010) Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino Acids 38: 859–867.
5. Zhang VM, Chavchich M, Waters NC (2012) Targeting protein kinases in the malaria parasite: update of an antimalarial drug target. Curr Top Med Chem 12: 456–472.
6. Hayakawa T, Arisue N, Udono T, Hirai H, Sattabongkot J, et al. (2009) Identification of Plasmodium malariae, a human malaria parasite, in imported chimpanzees. PLoS ONE 4: e7412.
7. Oyelade J, Ewejobi I, Brors B, Eils R, Adebiyi E (2011) Computational identification of signalling pathways in Plasmodium falciparum. Infect Genet Evol 11: 755–764.
8. Tedder PM, Bradford JR, Needham CJ, McConkey GA, Bulpitt AJ, et al. (2010) Gene function prediction using semantic similarity clustering and enrichment analysis in the malaria parasite Plasmodium falciparum. Bioinformatics 26: 2431–2437.
9. Tonkin CJ, Kalanon M, McFadden GI (2008) Protein targeting to the malaria parasite plastid. Traffic 9: 166–175.
10. Nguyen MN, Rajapakse JC (2006) Two-stage support vector regression approach for predicting accessible surface areas of amino acids. Proteins 63: 542–550.
11. Chang DT, Huang HY, Syu YT, Wu CP (2008) Real value prediction of protein solvent accessibility using enhanced PSSM features. BMC Bioinformatics 9 Suppl 12: S12.
12. Kumar M, Gromiha MM, Raghava GP (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 8: 463.
13. Kumar KK, Pugalenthi G, Suganthan PN (2009) DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. J Biomol Struct Dyn 26: 679–686.
14. Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 6: -.
15. Hwang S, Gou ZK, Kuznetsov IB (2007) DP-Bind: a Web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23: 634–636.
16. Wang L, Yang MQ, Yang JY (2009) Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics 10 Suppl 1: S1.
17. Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recognition Letters 28: 1610–1615.

18. Mei S, Fei W (2010) Amino acid classification based spectrum kernel fusion for protein subnuclear localization. BMC Bioinformatics 11 Suppl 1: S17.
19. Kumar M, Gromiha MM, Raghava GPS (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. Journal of Molecular Recognition 24: 303–313.
20. Ramana J, Gupta D (2010) Machine Learning Methods for Prediction of CDK-Inhibitors. PLoS ONE 5: -.
21. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol44, 60) 43: 246–255.
22. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19.
23. Wu ZC, Xiao X, Chou KC (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Molecular BioSystems 7: 3287–3297.
24. Wu ZC, Xiao X, Chou KC (2012) iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins. Protein & Peptide Letters 19: 4–14.
25. Xiao X, Wu ZC, Chou KC (2011) iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. Journal of Theoretical Biology 284: 42–51.
26. Shen HB, Chou KC (2009) Predicting protein fold pattern with functional domain and sequential evolution information. Journal of Theoretical Biology 256: 441–446.
27. Chou KC, Shen HB (2007) MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm 360: 339–345.
28. Shen HB, Chou KC (2007) EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm 364: 53–59.
29. Shen HB, Chou KC (2009) QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. Journal of Proteome Research 8: 1577–1584.
30. Khosravian M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H (2012) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett.
31. Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2012) Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. Med Chem: MC-EPUB-20120817-20120811 [pii].
32. Chou KC, Shen HB (2008) ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Comm 376: 321–325.
33. Deng JL (1989) Introduction to Grey System Theory. The Journal of Grey System: 1–24.
34. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273: 236–247.
35. Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34: 653–660.
36. Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 34: 320–327.
37. Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34: 103–109.
38. Nanni L, Lumini A, Gupta D, Garg A (2012) Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information. IEEE/ACM Trans Comput Biol Bioinform 9: 467–475.
39. Mohammad Beigi M, Behjati M, Mohabatkar H (2011) 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: 191–197.
40. Guo J, Rao N, Liu G, Yang Y, Wang G (2011) Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. Journal of Computational Chemistry 32: 1612–1617.
41. Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) 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: 18–23.
42. Zou D, He Z, He J, Xia Y (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. Journal of Computational Chemistry 32: 271–278.
43. Sun XY, Shi SP, Qiu JD, Suo SB, Huang SY, et al. (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. Mol Biosyst: 10.1039/c1032mb25280e.
44. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
45. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) 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: 17–26.
46. Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology 248: 546–551.
47. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263: 203–209.
48. Hayat M, Khan A (2012) Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. Protein & Peptide Letters 19: 411–421.
49. Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Analytical Biochemistry 425: 117–119.
50. Shen HB, Chou KC (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry 373: 386–388.
51. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
52. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, et al. (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34: D257–260.
53. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, et al. (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35: D237–240.
54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–29.
55. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258–261.
56. Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11: 2105–2134.
57. Chou KC (1995) The convergence-divergence duality in lectin domains of the selectin family and its implications. FEBS Letters 363: 123–126.
58. Chou KC, Wu ZC, Xiao X (2012) iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Molecular Biosystems 8: 629–641.
59. Xiao X, Wang P, Chou KC (2008) Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. Journal of Theoretical Biology 254: 691–696.
60. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 27:21–27.
61. Hayat M, Khan A (2012) MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM. J ournal of Theoretical Biology 292: 93–102.
62. Zakeri P, Moshiri B, Sadeghi M (2011) Prediction of protein submitochondria locations based on data fusion of various features of sequences. Journal of Theoretical Biology 269: 208–216.
63. Chen C, Shen ZB, Zou XY (2012) Dual-Layer Wavelet SVM for Predicting Protein Structural Class Via the General Form of Chou's Pseudo Amino Acid Composition. Protein & Peptide Letters 19: 422–429.
64. Chou KC, Wu ZC, Xiao X (2011) iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. PLoS One 6: e18258.
65. Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. Journal of Theoretical Biology 271: 10–17.
66. Saffari B, Mohabatkar H, Mohsenzadeh S (2008) T and B-cell Epitopes Prediction of Iranian Saffron (Crocus sativus) Profilin by Bioinformatics Tools. Protein Pept Lett 15: 280–285.
67. Chou KC (2001) Using subsite coupling to predict signal peptides. Protein Engineering 14: 75–79.
68. Chou KC (2001) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979.