Imbalanced Multi-Modal Multi-Label Learning for Subcellular Localization Prediction of Human Proteins with Both Single and Multiple Sites

Jianjun He, Hong Gu*, Wenqi Liu

School of Control Science and Engineering, Dalian University of Technology, Dalian, Liaoning, China

Abstract

It is well known that an important step toward understanding the functions of a protein is to determine its subcellular location. Although numerous prediction algorithms have been developed, most of them typically focused on the proteins with only one location. In recent years, researchers have begun to pay attention to the subcellular localization prediction of the proteins with multiple sites. However, almost all the existing approaches have failed to take into account the correlations among the locations caused by the proteins with multiple sites, which may be the important information for improving the prediction accuracy of the proteins with multiple sites. In this paper, a new algorithm which can effectively exploit the correlations among the locations is proposed by using Gaussian process model. Besides, the algorithm also can realize optimal linear combination of various feature extraction technologies and could be robust to the imbalanced data set. Experimental results on a human protein data set show that the proposed algorithm is valid and can achieve better performance than the existing approaches.

Citation: He J, Gu H, Liu W (2012) Imbalanced Multi-Modal Multi-Label Learning for Subcellular Localization Prediction of Human Proteins with Both Single and Multiple Sites. PLoS ONE 7(6): e37155. doi:10.1371/journal.pone.0037155

Editor: Franca Fraternali, King's College London, United Kingdom

Received November 24, 2011; **Accepted** April 14, 2012; **Published** June 8, 2012

Funding: This work is supported by National Major Project during the 11st Five-Year Plan Period of China (Grant No. 2009ZX09306-005, 2009ZX10004-701), and Liaoning Provincial Natural Science Foundation of China (Grant No. 20102025). 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: guhong@dlut.edu.cn

Introduction

Over the past years, the research on determining the subcellular locations of proteins has attracted more attention from academia due to its important roles in understanding protein functions, identifying drug targets, annotating genomes and so on. The approaches for determining the subcellular locations of proteins can be divided into two categories: experimental and computational methods [1]. Experimental methods such as cell fractionation, electron microscopy and fluorescence microscopy usually are time consuming, expensive and laborious [2]. These limitations have made the experimental methods unable to cope with the situation that a large number of protein sequences continue to emerge from the genome sequencing projects, and have encouraged the ongoing efforts to develop computational methods. It is well known that the information on the final subcellular location of a protein is basically encoded as a part of its amino acid sequence and such a sequence is thought to be recognized by a specific receptor protein as a protein sorting signal. Thus, it would be possible, at least in principle, for us to predict the subcellular location of a protein from its amino acid sequence by using computational methods [3]. In addition, many studies in other related areas have indicated that sequence-based prediction approaches, such as those for predicting drug-target interaction networks [4], predicting transcriptional activity of multiple site p53 mutants [5], prediction of body fluids [6], predicting protein metabolic stability [7], predicting antimicrobial peptides [8], identifying DNA binding proteins [9], identifying regulatory pathways [10], predicting signal peptides [11], predicting HIV cleavage sites in proteins [12,13], predicting the network of substrate-enzyme-product triads [14], predicting protein pathway networks [15], predicting proteases and their types [16], and predicting membrane proteins and their types [17], can generate many useful data for which it would be time-consuming and costly to obtain by experiments alone, and can timely provide very useful insights for both basic research and application by being combined with the information derived from the structural bioinformatics tools (see, e.g., [18]). In view of this, computationally predicting the subcellular locations of proteins from their amino acid sequences may become a useful complement to the experimental methods.

Since the pioneering efforts were provided [19,20], a number of sequence-based computational methods had been developed for predicting the subcellular locations of proteins. For example, based on N-terminal sequence information only, a neural network-based tool called TargetP was developed in [21] for large-scale subcellular localization prediction. Support vector machine (SVM) was introduced to predict the subcellular locations of proteins from their amino acid composition [22] and functional domain composition [23], respectively. In [24,25], the subcellular localization prediction problem of apoptosis proteins was studied. In order to avoid losing the sequence order information, Chou [26] proposed a concept of pseudo amino acid composition (PseAA composition) to represent the protein samples. Soon afterwards, many different prediction methods were proposed based on PseAA composition [27–34]. Text mining approach was used to improve the prediction results of protein subcellular localization by Lu et al. [35] for both prokaryote and eukaryote. MultiLoc, a SVM-based approach, was proposed in [36] through integrating N-terminal targeting sequences, amino acid composition and protein sequence motifs. A package of web servers named Cell-PLoc was developed by Chou and Shen [37] for predicting the subcellular locations of proteins in various organisms. A wider view of some other published protein subcellular localization prediction methods may be found in [2,3].

As mentioned above, through the continuing efforts of researchers, many computational methods which can achieve superior performance have been developed. However, all these studies [1–3,19–39], except for [2] and [37], focused only on mono-locational proteins, i.e., they assume that each protein exists in only one cellular compartment. This is not always the case. In fact, recent evidences [40,41] indicate that a mass of proteins have multiple sites in the cell. For addressing this problem, Scott et al. [42] established a Bayesian network predictor based on the combination of InterPro motifs and specific membrane domains in human proteins. By hybridizing three feature extraction techniques including gene ontology, functional domain and pseudo amino acid composition, Chou and Cai [43] developed a nearest neighbor algorithm for predicting the subcellular locations of proteins with multiple sites in budding yeast. In 2007, based on a feature representation frame of hybridizing gene ontology and amphiphilic pseudo amino acid composition and an ensemble k-nearest neighbor classifier, two algorithms called Euk-mPLoc [44] and Hum-mPLoc [45] were developed by Chou and Shen to deal with the eukaryotic and human proteins with both single and multiple sites, respectively. Later, they presented an improved feature representation frame by hybridizing the gene ontology, functional domain, and sequential evolutionary information, and several new approaches such as Euk-mPLoc 2.0 [46], Hum-mPLoc 2.0 [47], Plant-mPLoc [48] and Virus-mPLoc [49] were proposed. Lee et al. [50] developed a PLPD algorithm by using a density-induced support vector data description (D-SVDD) approach. In [51], Briesemeister et al. presented an algorithm named YLoc by using the simple naive Bayes classifier. Lin et al. [52] proposed a knowledge based approach by using the local

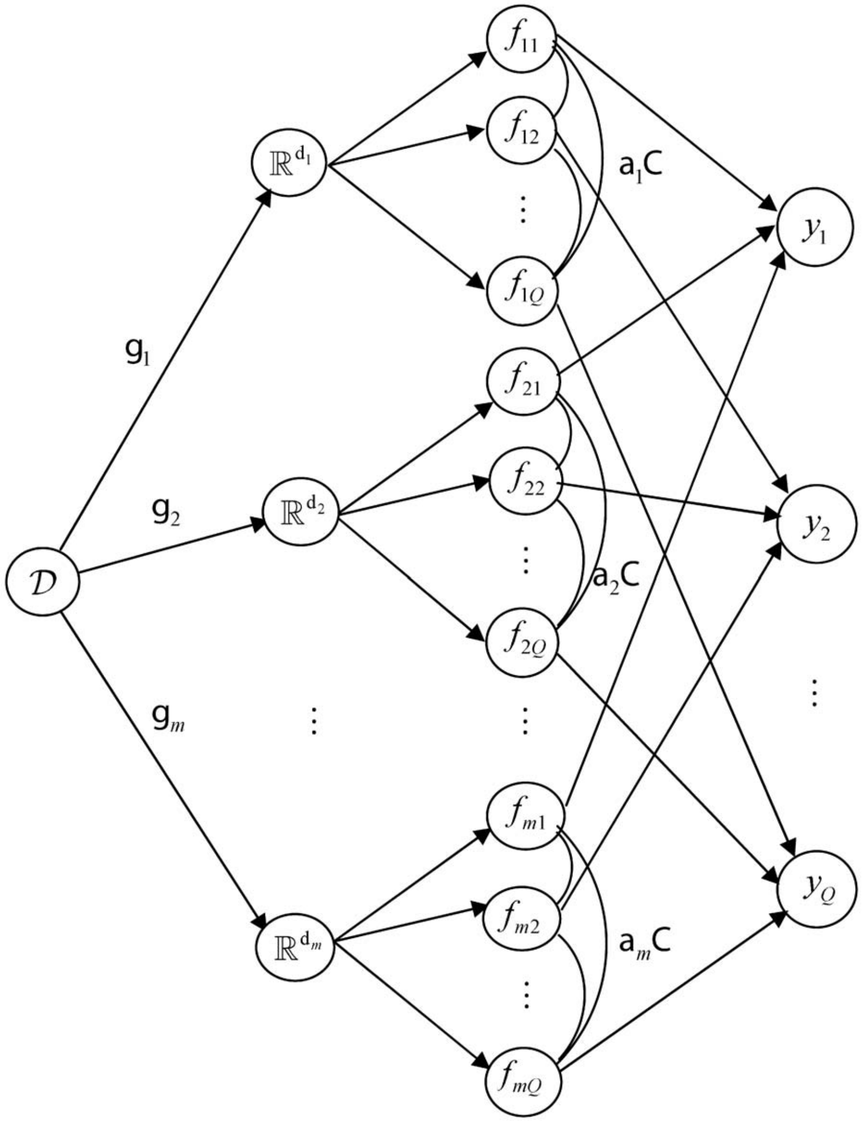

Figure 1. Graphical model for IMMMLGP.
doi:10.1371/journal.pone.0037155.g001

sequence similarity. Recently, four new approaches called iLoc-Euk [53], iLoc-Gneg [54], iLoc-Plant [55] and iLoc-Virus [56] were proposed based on a multi-label classifier to predict the subcellular locations of eukaryotic, Gram-negative bacterial, plant, and virus proteins, respectively. In [57], Wu et al. presented a multi-layer classifier to predict the subcellular locations of Gram-positive bacterial proteins. In [58], a new predictor, called iLoc-Hum, was developed based on the accumulation-label scale for predicting the subcellular locations of human proteins.

In order to deal with the protein with multiple sites, the common idea of the existing approaches is to train one or more single-label classifiers by transforming the original multi-label data into single-label ones and classify the query protein to the locations whose score outputted by the single-label classifiers satisfying some conditions. Three strategies were mainly used to transform the original multi-label data into single-label data. The first category such as Chou and Shen's work [46,49] is to take the protein with multiple sites as multiple proteins with single site; the second category [50] is to transform the original data set into multiple binary data sets, one for each location, and each binary data set includes all protein samples of the original data set, which are labeled positively if in the original data set they belong to the location corresponding to this binary data set and negatively otherwise; the third category [51] is to regard every possible combination of locations as a new class. However, the third strategy is infeasible in most cases because the number of classes will increase exponentially and the data in the new classes usually are sparse; the first and second one have limitations as well because they neglect the correlations among the locations caused by the protein with multiple sites. In fact, the correlations among the locations are the important information for improving the prediction accuracy. Taking the data set of eukaryotic proteins [46] as an example, it can be seen that almost all the proteins of cyanelle and hydrogenosome only have one site and about 30% proteins of cytoplasm also belong to nucleus. If a classifier can obtain these correlations from the training data set, it will think over the correctness of prediction result "a certain protein belongs to cyanelle and other locations simultaneously", and will have to reconsider whether the location 'nucleus' is missed when a protein was located to cytoplasm only. Thus, the first research content of this paper is to improve the performance of the classifier by considering the correlations among the locations caused by the protein with multiple sites.

In addition, to improve the whole performance of protein subcellular localization prediction approaches, another important factor is to represent the proteins with an effective feature extraction technology. Although the proteins may contain all the information such that they can be transported to the due subcellular compartments exactly, to establish a quality feature extraction technology that can mine this information is still a challenging problem. However, with the efforts of researchers, various types of feature extraction technologies based on the different local information of proteins such as N-terminus, sequence motifs, amino acid composition, and gene ontology terms have been proposed. Thus, we can try to improve the prediction performance by incorporating multiple local feature information of proteins. In fact, researchers have already done some work in this aspect. However, in many cases, different types of feature information were included in one predictor based on the subjective understanding of researchers, and it is hardly to realize the optimal combination of them. Thus, the second research context of our work is to optimally combine multiple feature extraction technologies in the predictor.

Furthermore, the subcellular distribution of proteins is usually extremely imbalanced. For example, in the data set of eukaryotic proteins [46], the number of proteins in 'cytoplasm' is 2186, while the number of proteins in 'Hydrogenosome' is only 10. In this case, the common classifier will tend to be overwhelmed by the majority classes and ignore the minority ones. Thus, the third research context of our work is to address the imbalanced data problem.

In order to consider aforementioned three problems simultaneously, a new classifier is proposed in this paper by using Gaussian process model. The basic idea of the proposed algorithm is to define multiple latent functions on the feature spaces, then the correlations among the locations can be identified by the

Table 1. The experimental results (mean) on human protein data sets for investigating the usefulness of the correlations among the locations.

Evaluation metric		The proposed algorithm					
		The original data set			The new data set (40%)		
		Normal	Variation	The gap	Normal	Variation	The gap
The whole test set	Average precision ↑	0.661	0.655	0.006	0.653	0.636	0.017
	Recall ↑	0.595	0.587	0.008	0.562	0.543	0.019
	F1-score ↑	0.530	0.522	0.008	0.516	0.504	0.012
	Absolute true success rate ↑	0.274	0.261	0.013	0.204	0.189	0.015
	Coverage ↓	2.003	2.047	−0.044	2.630	2.711	−0.081
	Ranking loss ↓	0.129	0.132	−0.003	0.143	0.148	−0.005
Samples with multiple sites	Average precision ↑	0.688	0.673	0.015	0.700	0.678	0.022
	Recall ↑	0.478	0.459	0.019	0.535	0.498	0.037
	F1-score ↑	0.535	0.518	0.017	0.572	0.545	0.027
	Absolute true success rate ↑	0.179	0.148	0.031	0.231	0.181	0.050
	Coverage ↓	3.889	4.030	−0.141	3.825	3.954	−0.129
	Ranking loss ↓	0.152	0.158	−0.006	0.148	0.155	−0.007

doi:10.1371/journal.pone.0037155.t001

covariance matrix between these latent functions, the optimal linear combination of different feature extraction technologies can be realized by defining a likelihood function and the imbalance of data can be coped with by the weighting coefficient of the likelihood on each sample. Since it can deal with the problems possessing the following properties: (1) the distribution of data on different classes may be imbalanced, (2) the data are represented in multiple feature spaces, and (3) each datum may associate with multiple labels simultaneously, the machine learning framework described in this paper is named imbalanced multi-modal multi-label learning (IMMML). we also call the proposed algorithm imbalanced multi-modal multi-label Gaussian process (IMMMLGP).

According to a recent comprehensive review [59], to establish a really useful predictor for determining the subcellular locations of proteins based on their sequence information, we need to consider the following procedures: (i) construct or select a valid benchmark data set to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect the intrinsic correlation with their subcellular locations; (iii) introduce or develop a powerful algorithm (or engine, classifier) 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.

Methods

Let D and $\mathcal{Y}=\{y_1,y_2,\cdots,y_Q\}$ respectively denote the sets of proteins and the subcellular locations for a certain subcellular localization prediction problem, where Q is the number of subcellular locations. Let $\{g_1,g_2,\cdots,g_m\}$ be the set of m feature extraction technologies used to extract the feature information of the proteins. Thus, each protein $X\in\mathcal{D}$ can be represented by $\{x_1,x_2,\cdots,x_m\}$, where, $x_j\in\mathbb{R}^{\mathrm{d_j}}$ is the feature vector of X associated with g_j, $\mathbb{R}^{\mathrm{d_j}}$ is the feature space corresponding to g_j, $j=1,2,\cdots,m$. Suppose $S=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)\}$ be a data set including n proteins with known sites, where, $X_i\in\mathcal{D}$ denotes the ith protein, $\{x_{i1},x_{i2},\cdots,x_{im}\}$ are the feature vectors of X_i and $Y_i\subset\mathcal{Y}$ is the set of subcellular locations associated with X_i, $i=1,2,\cdots,n$. For notation's convenience, Y_i can be represented by a vector $[y_{i1},y_{i2},\cdots,y_{iQ}]^T$, in which $y_{ik}=1$ denotes that protein X_i belongs to y_k, otherwise $y_{ik}=-1$. The goal of subcellular localization prediction of proteins with both single and multiple sites is to learn a function $h:\mathcal{D}\rightarrow 2^{\mathcal{Y}}$ from S which can correctly predict the subcellular locations of a new protein $X_*\in\mathcal{D}$. Being different with the traditional predictor for the proteins with single site, the output of h is a set of the locations.

Due to the desirable properties such as the natural Bayesian interpretation, explicit probabilistic formulation, and the ability to infer model parameters, Gaussian process model (GP) has received extensive attentions in recent years and become an important tool for many machine learning technologies. We will omit an introduction to it and refer the readers to the excellent books on this topic [60]. The main reason for using Gaussian process model but not other methods in this paper is that it can infer the correlations among the subcellular locations and the optimal combination coefficients of feature extraction technologies in a more convenient way.

To represent our uncertainty over subcellular locations for a protein, a better method is to output a probability for each subcellular location. As shown in Fig. 1, the main idea of IMMMLGP is to assume an unobservable latent function f_{jk} for every subcellular location y_k on the feature space $\mathbb{R}^{\mathrm{d_j}}$, $j=1,2,\cdots,m,k=1,2,\cdots,Q$, and then the probability that a protein X belongs to subcellular location y_k can be obtained by the combination of latent functions $\{f_{1k},f_{2k},\cdots,f_{mk}\}$ that assumed for y_k. In IMMMLGP, the correlations among the subcellular locations can be identified by the covariance matrix of the latent functions; the optimal linear combination of different feature extraction technologies can be realized by defining a likelihood function and the combination coefficient of the jth feature extraction technology is just a parameter of the kernel function over feature space $\mathbb{R}^{\mathrm{d_j}}$; the imbalance of data can be coped with by giving a weighting coefficient to each sample in the joint likelihood. The details of IMMMLGP algorithm are shown as follows.

Gaussian Process Prior

The basic idea behind Gaussian process model is to place a Gaussian process prior over the latent functions. In this paper, we place the Gaussian process priors with zero mean and the following covariance function over the latent functions $\{f_{jk}|j=1,2,\cdots,m,k=1,2,\cdots,Q\}$,

$$\begin{cases}\langle f_{jl}(x),f_{js}(x')\rangle=C_{ls}\cdot a_j\cdot k^j(x,x'),j=1,2,\cdots,m;\,l,s=1,2,\cdots,Q;\\ \langle f_{j_1l}(x),f_{j_2s}(x')\rangle=0,j_1\neq j_2;j_1,j_2=1,2,\cdots,m;\,l,s=1,2,\cdots,Q\end{cases}\qquad(1)$$

where, $C=(C_{ls})_{Q\times Q}$ is a positive semi-definite matrix that specifies the correlations among the subcellular locations, so that the observation of one location can affect the prediction on another one. As will be seen from the Section "Joint Likelihood", the main role of $a_j(a_j>0)$ is the weighting coefficient of the jth feature extraction technology. k^j is a covariance function over feature space $\mathbb{R}^{\mathrm{d_j}}$. In this paper, the Gaussian kernel was used as the covariance function k^j, i.e., $k^j(x,x')=e^{-\|x-x'\|^2/\beta_j}$. Since f_{j_1l} and f_{j_2s} are the functions defined on different input spaces when $j_1\neq j_2$, we can regard them as mutually independent functions.

We assume that all the parameters can be given except C and $\{a_j\}$. For notation's convenience, let $Y=[Y_1^T,Y_2^T,\cdots,Y_n^T]^T$, $D^j=\{x_{1j},x_{2j},\cdots,x_{nj}\}$, $D=\{X_i|i=1,2,\cdots,n\}$, $a=[a_1,a_2,\cdots,a_m]$, $f_{jki}=f_{jk}(x_{ij})$, $F_j=[f_{j11},f_{j21},\cdots,f_{jQ1},f_{j12},\cdots,f_{jQ2},\cdots,f_{j1n},\cdots,f_{jQn}]^T$, and $F=[F_1^T,F_2^T,\cdots,F_m^T]^T$, $f_{jk*}=f_{jk}(x_{*j})$, $F_{*j}=[f_{j1*},f_{j2*},\cdots,f_{jQ*}]^T$, $F_*=[F_{*1}^T,F_{*2}^T,\cdots,F_{*m}^T]^T$, $i=1,2,\cdots,n$, $j=1,2,\cdots,m$, $k=1,2,\cdots,Q$, $\{x_{*1},x_{*2},\cdots,x_{*m}\}$ are the feature vectors of X_*.

According to (1), the joint distribution $p(F|D,C,a)$ and $p(F,F_*|D,X_*,C,a)$ can be written as

$$p(F|D,C,a)=\prod_{j=1}^{m}\mathcal{N}\left(F_j|0,(a_jK^j)\otimes C\right)\qquad(2)$$

and

$$p(F_*,F|D,X_*,C,a)=\prod_{j=1}^{m}\mathcal{N}\left(\begin{bmatrix}F_{*j}\\ F_j\\ \end{bmatrix}|0,\begin{bmatrix}a_jK_{**}^j & a_j(K_*^j)^T\\ a_jK_*^j & a_jK^j\end{bmatrix}\otimes C\right)\qquad(3)$$

respectively, where $\otimes$ denotes the Kronecker product, the element of Kj is $k^j(x,x'),x,x'\in D^j$, $K_{**}^j=k^j(x_{*j},x_{*j})$, and K_*^j is a column vector and its ith element is $k^j(x_{ij},x_{*j}),i=1,2,\cdots,n$. Thus, the conditional prior $p(F_*|F,D,X_*,C,a)$ can be deduced analytically,

$$p(F_*|F,D,X_*,C,a)=\prod_{j=1}^{m}\mathcal{N}(F_{*j}|((K_*^j)^T(K^j)^{-1}\otimes E) \\ F_j,a_j(K_{**}^j-(K_*^j)^T(K^j)^{-1}K_*^j)\otimes C) \quad (4)$$

where, E is an identity matrix.

Joint Likelihood

Let $p(Y|F)$ denote the joint likelihood, i.e., the joint probability of observing the class labels Y given the latent functions. Generally, the class labels can be regarded as independent variables given the latent functions. Thus, $p(Y|F)$ may be evaluated as a product of the likelihoods on individual observation, that is

$$p(Y|F)=\prod_{i=1}^{n}\prod_{k=1}^{Q}p(y_{ik}|f_{1ki},f_{2ki},\cdots,f_{mki}) \quad (5)$$

Since the imbalance of data should be considered, we can set a weighting coefficient to the likelihood of each observation such that it can enhance the influence of minority classes on joint likelihood and reduce the influence of the majority classes, i.e.,

$$p(Y|F)=\prod_{i=1}^{n}\prod_{k=1}^{Q}(p(y_{ik}|f_{1ki},f_{2ki},\cdots,f_{mki}))^{r_{ik}} \quad (6)$$

A detailed explanation of why the likelihood (6) can deal with the imbalance of data and the details of determining $\{r_{ik}|i=1,2,\cdots,n;k=1,2,\cdots,Q\}$ will be given in Appendix S1.

In this paper, we also would like to realize the optimal linear combination of various feature extraction technologies. It can be seen from (1) that the scale of f_{jki} can be determined by the covariance function, this suggests that the linear combination $\sum_{j=1}^{m}b_jf_{jki}(b_j>0)$ of $\{f_{jki}\}$ with covariance function $\{Ck^j\}$ is equivalent to the sum $\sum_{j=1}^{m}f_{jki}$ of $\{f_{jki}\}$ with covariance function $\{b_j^2Ck^j\}$. Thus, we can define likelihood $p(y_{ik}=1|f_{1ki},f_{2ki},\cdots,f_{mki})$ as

$$p(y_{ik}=1|f_{1ki},f_{2ki},\cdots,f_{mki})=\sigma(\sum_{j=1}^{m}f_{jki}) \quad (7)$$

then the optimal linear combination of various feature extraction technologies may be realized indirectly by choosing the weighting coefficients $\{a_j\}$ in (1). Here, $\sigma(t)=1/(1+e^{-t})$ is the logistic function. As the values of $p(y_{ik}=1|f_{1ki},f_{2ki},\cdots,f_{mki})$ and $p(y_{ik}=-1|f_{1ki},f_{2ki},\cdots,f_{mki})$ must sum to 1, thus likelihood $p(y_{ik}|f_{1ki},f_{2ki},\cdots,f_{mki})$ can be written as

$$p(y_{ik}|f_{1ki},f_{2ki},\cdots,f_{mki})=\sigma(y_{ik}\sum_{j=1}^{m}f_{jki}) \quad (8)$$

Posterior Distribution

By using Bayes's rule, the posterior distribution over F for given C and a becomes

$$p(F|D,Y,C,a)=\frac{p(Y|F)p(F|D,C,a)}{p(Y|D,C,a)} \quad (9)$$

where,

$$p(Y|D,C,a)=\int p(Y|F)p(F|D,C,a)\mathrm{d}F \quad (10)$$

is the marginal likelihood of the parameters C and a. It can be seen that the posterior distribution $p(F|D,Y,C,a)$ is a non-Gaussian distribution which can not be computed analytically. The same as the traditional GP classification models, Laplace's method can be utilized to obtain a Gaussian approximation of $p(F|D,Y,C,a)$, that is

$$p(F|D,Y,C,a)\approx q(F|D,Y,C,a)=\mathcal{N}(F|\hat{F},A^{-1}) \quad (11)$$

where $\hat{F}=\arg\max_F p(F|D,Y,C,a)$ and $A=-\nabla\nabla\log p(F|D,Y,C,a)|_{F=\hat{F}}$ is the Hessian matrix of the negative log posterior at $\hat{F}$. The details of solving $\hat{F}$ and A can be found in Appendix S1.

Prediction

By using the approximation $q(F|D,Y,C,a)$ of posterior (9) and the conditional prior $p(F_*|F,D,X_*,C,a)$ (4), the distribution of F_* can be deduced analytically

$$p(F_*|D,Y,X_*,C,a)=\int p(F_*|F,D,X_*,C,a)q(F|D,Y,C,a)\mathrm{d}F \\ =\mathcal{N}(F_*|\tilde{K}\hat{F},H+\tilde{K}A^{-1}\tilde{K}^T) \quad (12)$$

where,
$\tilde{K}=\mathrm{diag}\{(K_*^1)^T(K^1)^{-1},(K_*^2)^T(K^2)^{-1},\cdots,(K_*^m)^T(K^m)^{-1}\}\otimes E$,
$H=\mathrm{diag}\{H_1,H_2,\cdots,H_m\}\otimes C$, $\quad H_j=a_j(K_{**}^j-(K_*^j)^T(K^j)^{-1}K_*^j)$,
$\mathrm{diag}\{\cdot\}$ denotes block diagonal matrix.

Thus, the probability $p(y_{*k}=1|D,Y,X_*,C,a)$ that protein X_* belongs to subcellular location $k(k=1,2,...,Q)$ may be predicted by averaging out F_*, i.e.,

$$p(y_{*k}=1|D,Y,X_*,C,a) \\ =\int\sigma(\sum_{j=1}^{m}f_{jk*})p(F_*|D,Y,X_*,C,a)\mathrm{d}F_* \quad (13) \\ =\int\cdots\int\sigma(\sum_{j=1}^{m}f_{jk*})p(f_{1k*},f_{2k*},\cdots, \\ f_{mk*}|D,Y,X_*,C,a)\mathrm{d}f_{1k*}\mathrm{d}f_{2k*}\cdots\mathrm{d}f_{mk*}$$

Notice that the predictive probability (??) also can not be computed analytically. In this paper, we resort to Monte Carlo sampling method to compute it.

Until now, we have presented the whole IMMMLGP algorithm under the assumption that C, a and $\{r_{ik}|i=1,2,\cdots,n;k=1,2,\cdots,Q\}$ have been obtained. The details of computing C, a and $\{r_{ik}|i=1,2,\cdots,n;k=1,2,\cdots,Q\}$ can be found in **Appendix S1**.

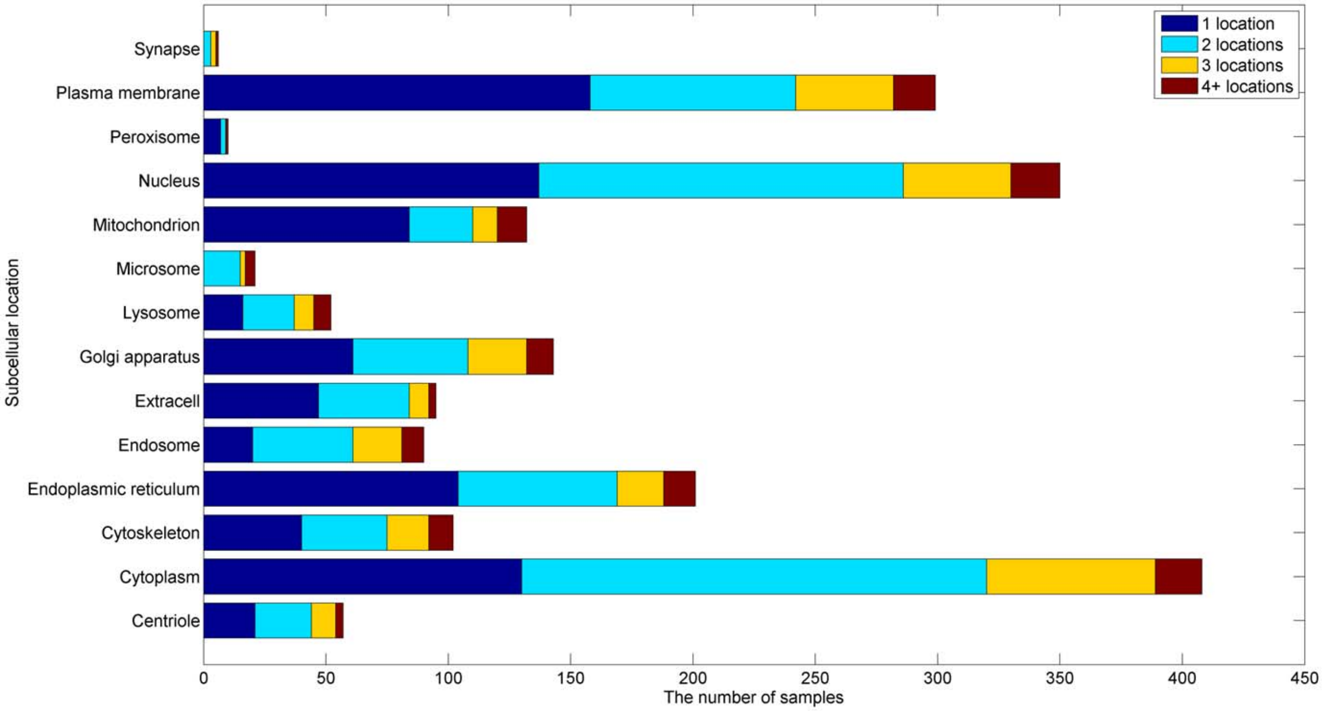

Figure 2. Subcellular distribution of the test samples.
doi:10.1371/journal.pone.0037155.g002

Results and Discussion

In this section, we test the proposed algorithm on a human protein data set collected from the Swiss-Prot database by Shen and Chou [47]. This data set includes 3106 different protein sequences covering 14 subcellular locations, where 2580 proteins belong to one subcellular location, 480 to two locations, 43 to three locations, and 3 to four locations. None of proteins included here has >25% pairwise sequence identity to any other in a same subcellular location. Five feature extraction technologies including GO process, GO function, GO component, composition of amino acids, and pseudo amino acid composition with $\lambda=11$, which measure the similarity of proteins from different aspects, are chosen in the experiments. The details of these feature extraction technologies can be found in [61] or [62]. In each experiment, the approach proposed in [63] is used to determine the parameter β_j of covariance function $k^j(x,x')=e^{-\|x-x'\|^2/\beta_j}$.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent data set test, subsampling test, and jackknife test [64]. Of the three test methods, the jackknife test is deemed the most objective [65]. The reasons are as follows. (i) For the independent data set test, although all the proteins used to test the predictor are outside the training data set used to train it so as to exclude the "memory" effect or bias, the

Table 2. The performance comparison between the proposed algorithm and Hum-mPLoc 2.0.

Evaluation metric		The proposed algorithm	Hum-mPLoc 2.0
The whole test set	Average precision ↑	**0.581**	0.579
	Recall ↑	**0.643**	0.519
	F1-score ↑	0.506	**0.541**
	Absolute true success rate ↑	0.202	**0.294**
	Coverage ↓	**4.303**	5.317
	Ranking loss ↓	**0.419**	0.496
Samples with multiple sites	Average precision ↑	**0.596**	0.568
	Recall ↑	**0.579**	0.443
	F1-score ↑	**0.576**	0.548
	Absolute true success rate ↑	**0.153**	0.114
	Coverage ↓	**6.800**	8.453
	Ranking loss ↓	**0.463**	0.568

doi:10.1371/journal.pone.0037155.t002

way of how to select the independent proteins to test the predictor could be arbitrary unless the number of independent proteins is sufficiently large. This kind of arbitrariness might result in completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent test data set might fail to keep so when tested by another independent test data set [64]. (ii) For the subsampling test, the concrete procedure usually used in literatures is the 2-fold, 5-fold, 7-fold or 10-fold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a benchmark data set is an astronomical figure even for a very simple data set, as elucidated in [65] and demonstrated by Equations (28)-(30) in [59]. Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark data set and a same predictor, the subsampling test cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as a good one. (iii) In the jackknife test, all the proteins in the benchmark data set will be singled out one-by-one and tested by the predictor trained by the remaining protein samples. During the process of jackknifing, both the training data set and test data set are actually open, and each protein sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem as mentioned above for the independent data set test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark data set. Accordingly, the jackknife test has been increasingly and widely used by those investigators to examine the quality of various predictors (see, e.g., [4–10,66–72]). However, to reduce the computational time, we will adopt the independent data set and subsampling test methods to examine the proposed predictor as done by many predictors with SVM or Bayesian network as the classifier [42,50]. And we will try to prevent the influence of the arbitrariness problem mentioned above on the experimental results through constructing an independent data set as large as possible or repeating subsampling test many times.

Since the performance evaluation of multi-label problems is much more complicated than the traditional single-label ones, the following popular multi-label evaluation metrics are used to comprehensively evaluate the performance of the proposed approach. Here, $S=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_p,Y_p)\}$ denotes a test set, $h(X_i)$ returns a set of proper labels of X_i; $h(X_i,y)$ returns a probability indicating the confidence for y to be a proper label of X_i; $Rank_h(X_i,y)$ is the rank of y derived from $h(X_i,y)$.

- Average precision: $\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{|\{y'|Rank_h(X_i,y')\leq Rank_h(X_i,y),y'\in Y_i\}|}{Rank_h(X_i,y)}$. It can compute the average fraction of labels ranked above a particular label $y\in Y_i$.
- Coverage: $\frac{1}{p}\sum_{i=1}^{p}(\max_{y\in Y_i} Rank_h(X_i,y)-1)$. It can evaluate how far one needs to go in the list of labels in order to cover all the proper labels of a sample.
- Ranking loss: $\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i||\overline{Y_i}|}|\{(y,y')|h(X_i,y)\leq h(X_i,y'),(y,y')\in Y_i\times\overline{Y_i}\}|$, where $\overline{Y_i}$ is the complementary set of Yi. It can evaluate the average fraction of label pairs that are not correctly ordered for a sample.
- Recall: $\frac{1}{p}\sum_{i=1}^{p}\frac{|h(X_i)\cap Y_i|}{|Y_i|}$.
- F1-score: $2\frac{Rec\cdot Pre}{Rec+Pre}$, where $Pre=\frac{1}{p}\sum_{i=1}^{p}\frac{|h(X_i)\cap Y_i|}{|h(X_i)|}$ and $Rec=\frac{1}{p}\sum_{i=1}^{p}\frac{|h(X_i)\cap Y_i|}{|Y_i|}$.
- Absolute true success rate: $\frac{1}{p}\sum_{i=1}^{p}\Delta(i)$, where $\Delta(i)=\begin{cases}1 & h(X_i)\equiv Y_i\\ 0 & \text{otherwise}\end{cases}$. According to the definition, the prediction score of a test protein can be counted as 1 when and only when all its subcellular locations are exactly predicted without any underprediction or overprediction. Therefore, the absolute true success rate is much more strict and harsh than other metrics.

Table 3. Some examples of the experimental results outputted by the two algorithms.

Accession number	Locations annotated in Swiss-Prot database	The predicted results of Hum-mPLoc 2.0	The predicted results of the proposed algorithm
P60852	Plasma membrane; Extracell	Extracell	Plasma membrane; Extracell
O75396	Endoplasmic reticulum; Golgi apparatus	Endoplasmic reticulum	Endoplasmic reticulum; Golgi apparatus
Q2VWA4	Cytoplasm; Nucleus	Nucleus	Cytoplasm; Nucleus
Q6NT55	Endoplasmic reticulum; Microsome	Endoplasmic reticulum; Microsome; Extracell	Endoplasmic reticulum; Microsome
P42261	Plasma membrane; Endoplasmic reticulum; Synapse	Plasma membrane; Synapse; Extracell	Plasma membrane; Endoplasmic reticulum; Synapse
Q9Y3A5	Cytoplasm; Nucleus; Cytoskeleton	Mitochondrion	Cytoplasm; Nucleus
P49419	Cytoplasm; Nucleus; Mitochondrion	Mitochondrion	Cytoplasm; Mitochondrion
Q86WV6	Endoplasmic reticulum; Cytoplasm; Mitochondrion; Plasma membrane	Cytoplasm	Cytoplasm; Endoplasmic reticulum
Q99527	Plasma membrane; Golgi apparatus; Endoplasmic reticulum	Plasma membrane	Plasma membrane; Endoplasmic reticulum
O75410	Cytoplasm; Nucleus; Centriole	Nucleus	Cytoplasm; Nucleus; Centriole; Mitochondrion

doi:10.1371/journal.pone.0037155.t003

The more detailed definitions of the first five metrics can be found in [73] and [74], and the definition of absolute true success rate can be found in [58] or [53].

As shown in the Section "Methods", a main contribution of the proposed approach is that the correlations among the locations are exploited by using a covariance matrix *C*. In order to justify the fact that the superior performance of the proposed algorithm benefits by considering the correlations among labels, we firstly investigate the performance difference between the proposed approach and its variation in which the covariance matrix *C* is assumed to be an identity matrix (i.e., the locations are considered as mutually independent ones). Table 1 shows the experimental results on the human protein data set. For each evaluation metric, '↓' indicates 'the smaller the better' while '↑' indicates 'the bigger the better'. In our experiments, the data were randomly partitioned in half to form a training set and a test set. We repeated each experiment for 5 random splits, and reported the average of the results obtained over 5 different test sets. In order to study the influence of the percentage of the proteins with multiple sites on the proposed approach, we construct a new human protein data set which contains around 40% proteins with multiple sites by randomly removing some proteins with single site from the original data set. Table 1 also presents the experimental results on this new data set. It can be seen from Table 1 that the proposed approach can achieve superior performance than its variation no matter on the whole test set or the test samples with multiple sites only. Moreover, the performance gap tends to increase when the percentage of the proteins with multiple sites increases. Thus, as what we expected, the correlations among the locations are the useful information for improving the prediction accuracy of the predictor and the covariance matrix could exploit this information effectively.

In order to evaluate the relative performance of the proposed algorithm, it is compared with an existing algorithm named Hum-mPLoc 2.0 [47], which is a popular web-server predictor for the subcellular localization prediction of human proteins with multiple sites. Since the whole human data set has been taken as the training set of Hum-mPLoc 2.0, to make a fair and comprehensive comparison, we have to take it as the training set of the proposed algorithm also and construct a test set according to the following criteria: (1) they must belong to human proteins, as clearly annotated in Swiss-Prot database; (2) None of proteins included here has >25% sequence identity to the ones of the training set in a same subcellular location. By following the above procedures, we obtained a test set containing 1315 proteins, of which 825 located to one site, 369 to two sites, 91 to three sites, and 30 to more than three sites. The details about the distribution of these samples can be seen in Fig. 2. Table 2 presents the experimental results of the proposed algorithm and Hum-mPLoc 2.0, where the best result on each metric is shown in bold face. It can be seen from Table 2 that the proposed algorithm achieves the best performance on four of the six evaluation metrics as far as the whole test set is concerned. Since these evaluation metrics measure the performance of algorithms from different aspects, one algorithm usually is difficult to outperform another on all the metrics. Thus, overall, the proposed algorithm can achieve superior performance than Hum-mPLoc 2.0 on this test set. In addition, Table 2 also presents the experimental results of each algorithm on the test samples with multiple sites only. It can be seen that the proposed algorithm consistently outperforms Hum-mPLoc 2.0 on the samples with multiple sites in terms of all evaluation metrics. This suggests that the proposed algorithm has the obvious advantage than Hum-mPLoc 2.0 for predicting the subcellular locations of proteins with multiple sites.

In order to understand why the proposed algorithm can achieve superior performance than Hum-mPLoc 2.0 on the proteins with multiple sites, we analysis the difference of the results outputted by the two algorithm. Table 3 shows some examples of the experimental results outputted by them. For the first 5 proteins, all their sites are correctly identified by the proposed algorithm but only partial sites can be correctly predicted by Hum-mPLoc 2.0. For the others, all of the two algorithms only can correctly predict their partial sites or incorrectly predict all their sites. It can be seen from Table 3 that the proposed algorithm can output as much as possible corrected locations than Hum-mPLoc 2.0 in most cases. For example, according to the experimental annotation in Swiss-Prot, the protein with accession number P60852 belongs to two locations: Plasma membrane and Extracell. If using Hum-mPLoc 2.0 to predict its sites, the output is 'Extracell', and 'Plasma membrane' is missed; however, the proposed algorithm can correctly output all of them. This may be the main reason why the proposed algorithm achieves superior performance than Hum-mPLoc 2.0.

Finally, it should be pointed out that although the proposed algorithm can achieve superior performance than the existing ones, it mainly benefits by the novel classifier but not the feature information. In the future, we will try to improve the algorithm by using more feature information such as FunD (functional domain) representation and SeqEvo (sequential evolution) representation. Moreover, since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors [75], we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Supporting Information

Appendix S1
(PDF)

Acknowledgments

The authors wish to thank the reviewers for the valuable suggestions and comments, which are very helpful for strengthening the presentation of this paper.

Author Contributions

Conceived and designed the experiments: HG. Performed the experiments: JH WL. Analyzed the data: HG JH. Contributed reagents/materials/analysis tools: JH WL. Wrote the paper: JH.

References

1. Ma J, Gu H (2010) A novel method for predicting protein subcellular localization based on pseudo amino acid composition. BMB Reports 43: 670–676.
2. Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. Analytical Biochemistry 370: 1–16.
3. Imai K, Nakai K (2010) Prediction of subcellular locations of proteins: where to proceed? Proteomics 10: 3970–3983.
4. He Z, Zhang J, Shi XH, Hu LL, Kong X, et al. (2010) Predicting drugtarget interaction networks based on functional groups and biological features. PLoS ONE 5: e9603.
5. Huang T, Niu S, Xu Z, Huang Y, Kong X, et al. (2011) Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. PLoS ONE 6: e22940.

6. Hu LL, Huang T, Cai YD, Chou KC (2011) Prediction of body fluids where proteins are secreted into based on protein interaction network. PLoS ONE 6: e22989.
7. Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 5: e10972.
8. Wang P, Hu L, Liu G, Jiang N, Chen X, et al. (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS ONE 6: e18476.
9. Lin WZ, Fang JA, Xiao X, Chou KC (2011) iDNA-Prot: Identification of DNA binding proteins using random forest with grey model. PLoS ONE 6: e24756.
10. Huang T, Chen L, Cai YD, Chou KC (2011) Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS ONE 6: e25297.
11. Chou KC, Shen HB (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochemical and Biophysical Research Communications 357: 633–640.
12. Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry 268: 16938–16948.
13. Chou KC (1996) Review: Prediction of HIV protease cleavage sites in proteins. Analytical Biochemistry 233: 1–14.
14. Chen L, Feng KY, Cai YD, Chou KC, Li HP (2010) Predicting the network of substrate-enzymeproduct triads by combining compound similarity and functional domain composition. BMC Bioinformatics 11: 293.
15. Chen L, Huang T, Shi XH, Cai YD, Chou KC (2010) Analysis of protein pathway networks using hybrid properties. Molecules 15: 8177–8192.
16. Chou KC, Shen HB (2008) ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochemical and Biophysical Research Communications 376: 321–325.
17. Chou KC, Shen HB (2007) MemType-2L: AWeb server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications 360: 339–345.
18. Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11: 2105–2134.
19. Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. Journal of Molecular Biology 238: 54–61.
20. Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Engineering 12: 107–118.
21. Emanuelsson O, Nielsen H, Brunak S, Heijne GV (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology 300: 1005–1016.
22. Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728.
23. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biological Chemistry 277: 45765–45769.
24. Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins: Structure, Function, and Genetics 50: 44–48.
25. Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. Journal of Theoretical Biology 245: 775–783.
26. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure, Function, and Genetics 43: 246–255.
27. Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein & Peptide Letters 15: 612–616.
28. Jiang X, Wei R, Zhang TL, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein & Peptide Letters 15: 392–396.
29. Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. Journal of Theoretical Biology 248: 377–381.
30. Liao B, Jiang JB, Zeng QG, Zhu W (2011) Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. Protein & Peptide Letters 18: 1086–1092.
31. Liu T, Zheng X, Wang C, Wang J (2010) Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: An approach from auto covariance transformation. Protein & Peptide Letters 17: 1263–1269.
32. Shi JY, Zhang SW, Pan Q, Zhou GP (2008) Using pseudo amino acid composition to predict protein subcellular location: Approached with amino acid composition distribution. Amino Acids 35: 321–327.
33. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, et al. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. Journal of Theoretical Biology 259: 366–372.
34. Mei S (2011) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. Journal of Theoretical Biology 293: 121–130.
35. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, et al. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20: 547–556.
36. Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22: 1158–1165.
37. Chou KC, Shen HB (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3: 153–162.
38. Ma J, Liu W, Gu H (2010) Using elman networks ensemble for protein subnuclear location prediction. International Journal of Innovative Computing, Information & Control 6: 5093–5103.
39. Tian J, Gu H, Liu W, Gao C (2011) Robust prediction of protein subcellular localization combining PCA and WSVMs. Computers in Biology and Medicine 41: 648–652.
40. Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, et al. (2006) A mammalian organelle map by protein correlation profiling. Cell 125: 187–199.
41. Zhang S, Xia X, Shen J, Zhou Y, Sun Z (2008) DBMLoc: a Database of proteins with multiple subcellular localizations. BMC Bioinformatics 9: 127.
42. Scott MS, Thomas DY, Hallett MT (2004) Predicting subcellular localization via protein motif co-occurrence. Genome Research 14: 1957–1966.
43. Chou KC, Cai YD (2005) Predicting protein localization in budding Yeast. Bioinformatics 21: 944–950.
44. Chou KC, Shen HB (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research 6: 1728–1734.
45. Shen HB, Chou KC (2007) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochemical and Biophysical Research Communications 355: 1006–1011.
46. Chou KC, Shen HB (2010) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PloS ONE 5: e9931.
47. Shen HB, Chou KC (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Analytical Biochemistry 394: 269–274.
48. Chou KC, Shen HB (2010) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PloS ONE 5: e11335.
49. Shen HB, Chou KC (2010) Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. Journal of Biomolecular Structure & Dynamics 28: 175–186.
50. Lee KY, Kim DW, Na DK, Lee KH, Lee D (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. Nucleic Acids Research 34: 4655–4666.
51. Briesemeister S, Rahnenfuhrer J, Kohlbacher O (2010) Going from where to why–interpretable prediction of protein subcellular localization. Bioinformatics 26: 1232–1238.
52. Lin HN, Chen CT, Sung TY, Ho SY, Hsu WL (2009) Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. BMC Bioinformatics 10: S8.
53. 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.
54. Xiao X, Wu ZC, Chou KC (2011) A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites. PloS ONE 6: e20592.
55. 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.
56. 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.
57. Wu ZC, Xiao X, Chou KC (2011) iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein & Peptide Letters, DOI: BSP/PPL/E pub/0380 [pii].
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. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273: 236–247.
60. Rasmussen CE, Williams KI (2006) Gaussian process for machine learning. The MIT press.
61. Mei S, Fei W, Zhou S (2011) Gene ontology based transfer learning for protein subcellular localization. BMC Bioinformatics 12: 44.
62. 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.
63. Yin J, Li T, Shen HB (2011) Gaussian kernel optimization: Complex problem and a simple solution. Neurocomputing 74: 3816–3822.
64. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30: 275–349.
65. Chou KC, Shen HB (2010) Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Science 2: 1090–1103.

66. 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.
67. 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.
68. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
69. 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.
70. Hu L, Huang T, Shi X, Lu WC, Cai YD, et al. (2011) Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS ONE 6: e14556.
71. Wang P, Xiao X, Chou KC (2011) NR-2L: A two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS ONE 6: e23505.
72. Xiao X, Wang P, Chou KC (2011) GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Molecular BioSystems 7: 911–919.
73. Schapire RE, Singer Y (2000) BoosTexter: A boosting-based system for text categorization. Machine Learning 39: 135–168.
74. Tsoumakas G, Katakis I (2007) Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3: 1–13.
75. Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. Natural Science 2: 63–92.