OPEN ACCESS Freely available online

PLoS one

# A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites

**Xuan Xiao[1,2]*, Zhi-Cheng Wu[1], Kuo-Chen Chou[2]**

**1** Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China, **2** Gordon Life Science Institute, San Diego, California, United States of America

## Abstract

Prediction of protein subcellular localization is a challenging problem, particularly when the system concerned contains both singleplex and multiplex proteins. In this paper, by introducing the ''multi-label scale'' and hybridizing the information of gene ontology with the sequential evolution information, a novel predictor called **iLoc-Gneg** is developed for predicting the subcellular localization of Gram-positive bacterial proteins with both single-location and multiple-location sites. For facilitating comparison, the same stringent benchmark dataset used to estimate the accuracy of **Gneg-mPLoc** was adopted to demonstrate the power of **iLoc-Gneg**. The dataset contains 1,392 Gram-negative bacterial proteins classified into the following eight locations: (1) cytoplasm, (2) extracellular, (3) fimbrium, (4) flagellum, (5) inner membrane, (6) nucleoid, (7) outer membrane, and (8) periplasm. Of the 1,392 proteins, 1,328 are each with only one subcellular location and the other 64 are each with two subcellular locations, but none of the proteins included has $\geq 25\%$ pairwise sequence identity to any other in a same subset (subcellular location). It was observed that the overall success rate by jackknife test on such a stringent benchmark dataset by **iLoc-Gneg** was over 91%, which is about 6% higher than that by **Gneg-mPLoc**. As a user-friendly web-server, **iLoc-Gneg** is freely accessible to the public at http://icpr.jci.edu.cn/bioinfo/iLoc-Gneg. Meanwhile, a step-by-step guide is provided on how to use the web-server to get the desired results. Furthermore, for the user's convenience, the **iLoc-Gneg** web-server also has the function to accept the batch job submission, which is not available in the existing version of **Gneg-mPLoc** web-server. It is anticipated that **iLoc-Gneg** may become a useful high throughput tool for Molecular Cell Biology, Proteomics, System Biology, and Drug Development.

**Citation:** Xiao X, Wu Z-C, Chou K-C (2011) A Multi-Label Classifier for Predicting the Subcellular Localization of Gram-Negative Bacterial Proteins with Both Single and Multiple Sites. PLoS ONE 6(6): e20592. doi:10.1371/journal.pone.0020592

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

**Received** February 26, 2011; **Accepted** May 4, 2011; **Published** June 17, 2011



**Funding:** This work was supported by grants from the National Natural Science Foundation of China (No. 60961003), the Key Project of Chinese Ministry of Education (No. 210116), the Province National Natural Science Foundation of JiangXi (2009GZS0064 and 2010GZS0122), the Department of Education of Jiang-Xi Province (No. GJJ09271), and the plan for training youth scientists (stars of Jing-Gang) of Jiangxi Province. 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: xiaoxuan0326@yahoo.com.cn

## Introduction

Bacteria can be divided into two groups: Gram-positive and Gram-negative. Gram-positive bacteria are those that are stained dark blue or violet by Gram staining; while Gram-negative bacteria cannot retain the stain, instead taking up the counter-stain and appearing red or pink.

It has special meaning for both basic research and drug design to study bacteria because (1) they are the workhorses for the fields of molecular biology, biochemistry, and genetics due to their ability to quickly grow and being relatively easier to be manipulated, and (2) they are both harmful and useful. With the explosion of protein sequences generated in the post-genomic era, we are challenged to develop computational methods for timely and accurately identifying the subcellular locations of newly discovered bacterial proteins based on their sequence information alone because this kind of knowledge will be very useful for selecting proper bacterial proteins for a special target, or screening and prioritizing candidates in drug design.

Actually, numerous predictors were developed for identifying subcellular localization of proteins in various organisms (see [1,2] as well as the long list of references cited in the two review papers). However, those that are specialized for dealing with Gram-negative proteins are only a few. They are called "**PSORT**" [1,3,4], "**PSORT-B**" [5], and **PSORTb v.2.0** [6]. All these methods have played important roles in stimulating the development of this area. To improve the prediction coverage scope and the quality of benchmark datasets, the predictor called **Gneg-PLoc** [7] was developed. Compared with the previous methods, **Gneg-PLoc** extended the coverage scope from five to eight subcellular location sites. Also, the benchmark datasets used to train and test the predictor have been significantly refined. For instance, the benchmark datasets used in **PSORT-B** [5] contain many proteins with pairwise sequence identity higher than 90%, while in the benchmark datasets of **Gneg-PLoc** [7] none of the proteins included has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; i.e., the latter is much more stringent and rigorous than the former in excluding the homology bias and redundancy. Also, **Gneg-PLoc** was able to yield higher success rates.

However, all the aforementioned predictors cannot be used to deal with multiplex proteins that may simultaneously exist at, or

move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery [8,9]. Particularly, as pointed out by Millar et al. [10], recent evidences have indicated that an increasing number of proteins have multiple locations in the cell.

To make **Gneg-PLoc** [7] be able to deal with multiplex Gram-negative proteins as well, a predictor called **Gneg-mPLoc** [11] was developed recently, where the character "m" in front of "PLoc" stands for "multiple", meaning that it can be also used to deal with Gram-negative bacterial proteins with multiple locations.

However, **Gneg-mPLoc** has the following shortcomings. **(1)** In predicting the number of subcellular location sites for a query Gram-negative protein, an optimal threshold factor $\theta^*$ (see Eq.48 of [2]) was adopted without providing its statistical implication and detailed learning process. It would be more instructive if we could find a more intuitive approach to determine this with a more natural manner. **(2)** In formulating the protein samples, only the integer numbers 0 and 1 were used to reflect the GO (gene ontology) information [12,13]. Such an over-simplified formulation might cause some useful information lost so as to limit the prediction quality. **(3)** Although a web-server for **Gneg-mPLoc** has been established at http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/, only one query protein sequence at a time is allowed when using the web-server to conduct prediction. For the convenience of users in handling many query Gram-negative protein sequences, such a rigid limit should be improved.

The present study was dedicated to develop a new and more powerful predictor, called **iLoc-Gneg**, for predicting Gram-negative bacterial protein subcellular localization by addressing the above three problems.

To establish a really useful statistical predictor for protein system, we usually need to consider the following procedures [14]: (1) select or construct a valid benchmark dataset to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly web-server [15] for the predictor that is accessible to the public. Below, let us describe how to realize these steps one by one.

## Materials and Methods

Here, we choose to use the same dataset $\mathbb{S}$ in establishing **Gneg-mPLoc** [11] as the benchmark dataset for the current study. The reasons doing so are as follows. **(1)** The dataset was constructed specialized for Gram-negative bacterial proteins and it can cover 8 subcellular location sites; compared with the other datasets such as the one in **PSORTb v.2.0** [6] that only covered 5 subcellular locations, the coverage scope of the dataset $\mathbb{S}$ from [11] is much wider. **(2)** None of proteins included in $\mathbb{S}$ has $\geq 25\%$ pairwise sequence identity to any other in a same subcellular location; compared with most of the other benchmark datasets in this area, the dataset $\mathbb{S}$ is much more rigorous in excluding homology bias and redundancy. **(3)** It contains both singleplex and multiplex proteins and hence can be used to train and test a predictor developed aimed at being able to deal with proteins with both single and multiple location sites. **(4)** Using the dataset $\mathbb{S}$ will also make it easier to compare the new predictor with the existing one because the tested results by **Gneg-mPLoc** on $\mathbb{S}$ have been well documented and reported [11].

The dataset $\mathbb{S}$ contains 1,392 Gram-negative bacterial protein sequences, of which 1,328 belong to one subcellular location, 64 to two locations, and none to three or more locations. The dataset covers 8 subcellular locations (**Fig. 1**), as can be formulated by

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \cup \mathbb{S}_6 \cup \mathbb{S}_7 \cup \mathbb{S}_8 \quad (1)$$

where $\mathbb{S}_1$ represents the subset for the subcellular location of cell inner membrane, $\mathbb{S}_2$ for cell outer membrane, $\mathbb{S}_3$ for cytoplasm, $\mathbb{S}_4$ for extracellular, and so forth (**Table 1**); while $\cup$ represents the symbol for "union" in the set theory. To avoid homology bias and redundancy, none of the proteins in $\mathbb{S}$ has $\geq 25\%$ pairwise sequence identity to any other in a same subset. For convenience, hereafter let us just use the subscripts of **Eq.1** as the codes of the 8 location sites; i.e., "1" for "cell membrane", "2" for "cell wall", "3" for "chloroplast", and so forth (**Table 2**).

For readers' convenience, the corresponding accession numbers and protein sequences in $\mathbb{S}$ are given in Supporting Information S1.

Note that because some proteins may occur in two or more locations, the 1,392 Gram-negative proteins actually correspond to 1,456 locative proteins. The concept of "locative proteins" was introduced for studying proteins with multiple subcellular location sites, as elaborated in [2].

To develop a powerful method for statistically predicting protein subcellular localization according to the sequence information, one of the most important things is to formulate the protein sequences with an effective mathematical expression that can truly reflect the intrinsic correlation with their subcellular localization [14]. However, it is by no means an easy job to realize this because this kind of correlation is usually deeply "buried" or hidden in piles of complicated sequences.

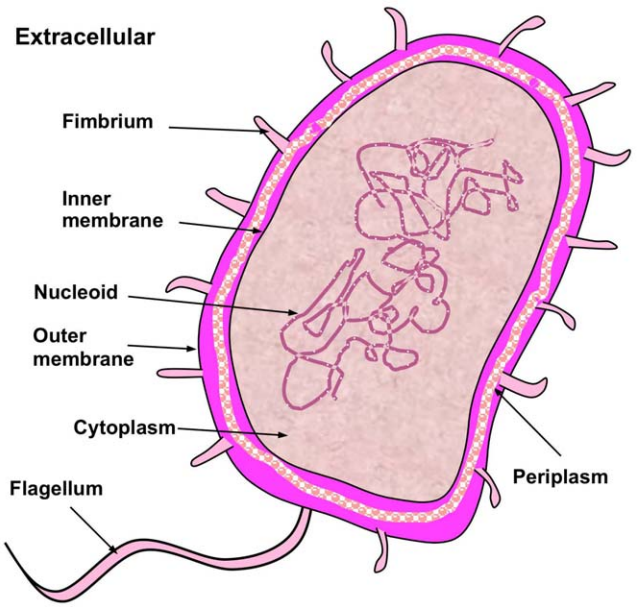


**Figure 1. Illustration to show the 8 subcellular locations of Gram-negative bacterial proteins.** The 8 locations are: (1) cytoplasm, (2) extracellular, (3) fimbrium, (4) flagellum, (5) inner membrane, (6) nucleoid, (7) outer membrane, and (8) periplasm. Note that in prokaryotic life forms, the nucleoid region is the part of the cell that contains the DNA molecule; unlike the true nucleus of eukaryotes, it is not delimited by a membrane.
doi:10.1371/journal.pone.0020592.g001

**Table 1.** Breakdown of the Gram-negative bacterial protein benchmark dataset $\mathbb{S}$ taken from [11].

| Subset | Subcellular location | Number of proteins |
|---|---|---|
| $\mathbb{S}_1$ | Cell inner membrane | 557 |
| $\mathbb{S}_2$ | Cell outer membrane | 124 |
| $\mathbb{S}_3$ | Cytoplasm | 410 |
| $\mathbb{S}_4$ | Extracellular | 133 |
| $\mathbb{S}_5$ | Fimbrium | 32 |
| $\mathbb{S}_6$ | Flagellum | 12 |
| $\mathbb{S}_7$ | Nucleoid | 8 |
| $\mathbb{S}_8$ | Periplasm | 180 |
| Total number of locative proteins $N(\mathrm{loc})$ | | 1,456[a] |
| Total number of different proteins $N(\mathrm{seq})$ | | 1,392[b] |

None of proteins included here has $\geq 25\%$ sequence identity to any other in a same subcellular location.
[a]See Eqs.36–38 of [2] for the definition about the number of locative proteins, and its relation with the number of different proteins.
[b]Of the 1,392 different proteins, 1,328 have one subcellular location, 64 have two locations, and none have three or more locations.
doi:10.1371/journal.pone.0020592.t001

**Table 2.** A comparison of the jackknife success rates by **Gnec-mPLoc** [11] and the current **iLoc-Gneg** on the benchmark dataset $\mathbb{S}$ (cf. Supporting Information S1) that covers 8 location sites of Gram-negative bacterial proteins in which none of the proteins included has $\geq 25\%$ pairwise sequence identity to any other in a same location.

| Code | Subcellular location | Success rate by jackknife test | |
|---|---|---|---|
| | | Gneg-mPLoc[a] | iLoc-Gneg[b] |
| 1 | Cell inner membrane | 525/557 = 94.3% | 539/557 = 96.8% |
| 2 | Cell outer membrane | 105/124 = 84.7% | 103/124 = 83.1% |
| 3 | Cytoplasm | 357/410 = 87.1% | 367/410 = 89.5% |
| 4 | Extracellular | 79/133 = 59.4% | 115/133 = 86.5% |
| 5 | Fimbrium | 28/32 = 87.5% | 30/32 = 93.8% |
| 6 | Flagellum | 0/12 = 0.0% | 12/12 = 100% |
| 7 | Nucleoid | 0/8 = 0.0% | 4/8 = 50% |
| 8 | Periplasm | 154/180 = 85.6% | 161/180 = 89.4% |
| Overall[c] | | 1248/1456 = **85.7%** | 1331/1456 = **91.4%** |

[a]The predictor from [11].
[b]The predictor proposed in this paper.
[c]Note that instead of 1,392 (the number of total different Gram-positive bacterial proteins), here we use 1,456 (the number of total different locative proteins) for the denominator. This is because some of the Gram-negative bacterial proteins in $\mathbb{S}$ may have more than one location site. See footnotes a and b of Table 1 for further explanation.
doi:10.1371/journal.pone.0020592.t002

The most straightforward method to formulate the sample of a query protein **P** was just using its entire amino acid sequence, as can be generally written by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{2}$$

where $\mathrm{R}_1$ represents the 1st residue of the protein **P**, $\mathrm{R}_2$ the 2nd residue, …, $\mathrm{R}_L$ the *L*-th residue, and they each belong to one of the 20 native amino acids. In order to identify its subcellular location(s), the sequence-similarity-search-based tools, such as BLAST [16,17], was utilized to search protein database for those proteins that have high sequence similarity to the query protein **P**. Subsequently, the subcellular location annotations of the proteins thus found were used to deduce the subcellular location(s) for **P**. Unfortunately, although it was quite intuitive and able to contain the entire information of a protein sequence, this kind of straightforward sequential model failed to work when the query protein **P** did not have significant sequence similarity to any location-known proteins.

Thus, various non-sequential or discrete models to formulate protein samples were proposed in hopes to establish some sort of correlation or cluster manner by which the prediction quality could be improved.

Among the discrete models for a protein sample, the simplest one is its amino acid (AA) composition or AAC [18]. According to the AAC-discrete model, the protein **P** of **Eq.2** can be formulated by [19,20]

$$\mathbf{P} = [f_1 \quad f_2 \quad \cdots \quad f_{20}]^{\mathbf{T}} \tag{3}$$

where $f_i (i = 1,2,\cdots,20)$ are the normalized occurrence frequencies of the 20 native amino acids in protein **P**, and **T** the transposing operator. Many methods for predicting protein subcellular localization were based on the AAC-discrete model (see, e.g., [19,21,22,23,24]). However, as we can see from **Eq.3**, if using the ACC model to represent the protein **P**, all its sequence-order effects would be lost, and hence the prediction quality might be limited.

To avoid completely lose the sequence-order information, the pseudo amino acid composition (PseAAC) was proposed to represent the sample of a protein, as formulated by [25]

$$\mathbf{P} = [p_1 \quad p_2 \quad \cdots \quad p_{20} \quad p_{20+1} \quad \cdots \quad p_{20+\lambda}]^{\mathbf{T}} \tag{4}$$

where the first 20 elements are associated with the 20 elements in **Eq.3** or the 20 amino acid components of the protein **P**, while the additional $\lambda$ factors are used to incorporate some sequence-order information via a series of rank-different correlation factors along a protein chain. For a brief introduction about PseAAC, please see a Wikipedia article at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition.

According to [14], the PseAAC for a protein **P** can be generally formulated as

$$\mathbf{P} = [\psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_\Omega]^{\mathbf{T}} \tag{5}$$

where the subscript $\Omega$ is an integer, and its value as well as the components $\psi_1, \psi_2, \ldots$ will depend on how to extract the desired information from the amino acid sequence of **P** (cf. **Eq.2**). As a general form, **Eq.5** can cover various different modes of PseAAC. For example, when its elements are given by

$$\psi_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & (1 \leq u \leq 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\theta_j}, & (20+1 \leq u \leq 20+\lambda = \Omega;\ \lambda < L) \end{cases} \tag{6}$$

we immediately obtain the formulation of PseAAC as originally introduced in [25], where the meanings for $w$, $\theta_j$, and $\lambda$ were clearly elaborated and hence there is no need to repeat here.

Below, let us use the general form of PseAAC (**Eq.5**) to find the formulations to reflect the core and essential features of protein samples that are closely correlated with their subcellular localization.

## 1. GO (Gene Ontology) Formulation

GO database [12] was established according to the molecular function, biological process, and cellular component. Accordingly, protein samples defined in a GO database space would be clustered in a way better reflecting their subcellular locations [2,26]. However, in order to incorporate more information, instead of only using 0 and 1 elements as done in [11], here let us use a different approach as described below.

**Step 1.** Compression and reorganization of the existing GO numbers. The GO database (version 74.0 released 30 July 2009) contains many GO numbers. However, these numbers do not increase successively and orderly. For easier handling, some reorganization and compression procedure was taken to renumber them. For example, after such a procedure, the original GO numbers GO:0000001, GO:0000002, GO:0000003, GO:0000009, GO:00000011, GO:0000012, GO:0000015, …, GO:0090204 would become GO_compress: 00001, GO_compress: 00002, GO_compress: 00003, GO_compress: 00004, GO_compress: 00005, GO_compress: 00006, GO_compress: 00007, ……, GO_compress: 11118, respectively. The GO database obtained thru such a treatment is called GO_compress database, which contains 11,118 numbers increasing successively from 1 to the last one.

**Step 2.** Using **Eq.5** with $\Omega=11{,}118$, the protein $\mathbf{P}$ can be formulated as

$$\mathbf{P}_{\mathrm{GO}}=\left[\psi_1^{\mathrm{G}} \quad \psi_2^{\mathrm{G}} \quad \cdots \quad \psi_u^{\mathrm{G}} \quad \cdots \quad \psi_{11118}^{\mathrm{G}}\right]^{\mathrm{T}} \tag{7}$$

where $\psi_u^{\mathrm{G}}$ $(u=1,2,\cdots,11118)$ are defined via the following steps.

**Step 3.** Use BLAST [27] to search the homologous proteins of the protein $\mathbf{P}$ from the Swiss-Prot database (version 55.3), with the expect value $E\le 0.001$ for the BLAST parameter.

**Step 4.** Those proteins which have $\ge 60\%$ pairwise sequence identity with the protein $\mathbf{P}$ are collected into a set, $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}$, called the "homology set" of $\mathbf{P}$. All the elements in $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}$ can be deemed as the "representative proteins" of $\mathbf{P}$, sharing some similar attributes such as structural conformations and biological functions [28,29,30]. Because they were retrieved from the Swiss-Prot database, these representative proteins must each have their own accession numbers.

**Step 5.** Search each of these accession numbers collected in Step 4 against the GO database at http://www.ebi.ac.uk/GOA/ to find the corresponding GO numbers [31].

**Step 6.** Based on the results obtained in Step 5, the elements in **Eq.7** can be written as

$$\psi_u^{\mathrm{G}}=\frac{\sum_{k=1}^{\mathbb{N}_{\mathbf{P}}^{\mathrm{homo}}}\delta(u,k)}{\mathbb{N}_{\mathbf{P}}^{\mathrm{homo}}} \quad (u=1,2,\cdots,11118) \tag{8}$$

where $\mathbb{N}_{\mathbf{P}}^{\mathrm{homo}}$ is the number of representative proteins in $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}$, and

$$\delta(u,k)=\begin{cases}1, & \text{if the } k\text{-th representative protein hits}\\ & \text{the } u\text{-th GO\_compress number}\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

As we can see from **Eq.7**, the GO formulation derived from the above steps consists of 11,118 real numbers rather than only the elements 0 and 1 as in the GO formulation adopted in [11].

Note that the GO formulation of **Eq.6** may become a naught vector or meaningless under any of the following situations: (**1**) the protein $\mathbf{P}$ does not have significant homology to any protein in the Swiss-Prot database, i.e., $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}=\varnothing$ meaning the homology set $\mathbb{S}_{\mathbf{P}}^{\mathrm{homo}}$ is an empty one; (**2**) its representative proteins do not contain any useful GO information for statistical prediction based on a given training dataset.

Under such a circumstance, let us consider using the sequential evolution formulation to represent the protein $\mathbf{P}$, as described below.

## 2. SeqEvo (Sequential Evolution) Formulation

Biology is a natural science with historic dimension. All biological species have developed continuously starting out from a very limited number of ancestral species. It is true for protein sequence as well [30]. Their evolution involves changes of single residues, insertions and deletions of several residues [32], 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 in a same subcellular location.

To incorporate the sequential evolution information into the PseAAC of **Eq.4**, here let us use the information of the PSSM (Position-Specific Scoring Matrix) [27], as described below.

**Step 1.** According to [27], the sequential evolution information of protein $\mathbf{P}$ can be expressed by a $20\times L$ matrix as given by

$$\mathbb{PSSM}=\begin{bmatrix}E_{1\to 1}^0 & E_{2\to 1}^0 & \cdots & E_{L\to 1}^0\\ E_{1\to 2}^0 & E_{2\to 2}^0 & \cdots & E_{L\to 2}^0\\ \vdots & \vdots & \ddots & \vdots\\ E_{1\to 20}^0 & E_{2\to 20}^0 & \cdots & E_{L\to 20}^0\end{bmatrix} \tag{10}$$

where $L$ is the length of $\mathbf{P}$ (counted in the total number of its constituent amino acids as shown in **Eq.1**), $E_{i\to j}^0$ represents the score of the amino acid residue in the $i$-th position of the protein sequence being changed to amino acid type $j$ during the evolutionary 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. The $20\times L$ scores in **Eq.10** were generated by using PSI-BLAST [27] 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 $\mathbf{P}$. However, according to the formulation of **Eq.10**, proteins with different lengths will correspond to column-different matrices causing difficulty for developing a predictor able to uniformly cover proteins of any length. To make the descriptor become a size-uniform matrix, let us consider the following steps.

**Step 2.** Use the elements in$\mathbb{PSSM}$ of Eq.10 to define a new matrix $\mathbf{M}$ as formulated by

$$\mathbf{M}=\begin{bmatrix}E_{1\to 1} & E_{2\to 1} & \cdots & E_{L\to 1}\\ E_{1\to 2} & E_{2\to 2} & \cdots & E_{L\to 2}\\ \vdots & \vdots & \ddots & \vdots\\ E_{1\to 20} & E_{2\to 20} & \cdots & E_{L\to 20}\end{bmatrix} \tag{11}$$

with

$$E_{i\to j}=\frac{E^0_{i\to j}-\bar{E}^0_j}{\mathrm{SD}\left(\bar{E}^0_j\right)}\quad (i=1,2,\cdots,L;j=1,2,\cdots,20) \tag{12}$$

where

$$\bar{E}^0_j=\frac{1}{L}\sum_{i=1}^{L}E^0_{i\to j}\quad (j=1,2,\cdots,20) \tag{13}$$

is the mean for $E^0_{i\to j}(i=1,2,\cdots,L)$ and

$$\mathrm{SD}\left(\bar{E}^0_j\right)=\sqrt{\sum\nolimits_{i=1}^{L}\left[E^0_{i\to j}-\bar{E}^0_j\right]^2\Big/L} \tag{14}$$

is the corresponding standard deviation.

**Step 3.** Introduce a new matrix generated by multiplying $\mathbf{M}$ with its own transpose matrix $\mathbf{M^T}$; i.e.,

$$\mathbf{MM^T}=\begin{bmatrix}\sum_{i=1}^{L}E_{i\to1}E_{i\to1} & \sum_{i=1}^{L}E_{i\to1}E_{i\to2} & \cdots & \sum_{i=1}^{L}E_{i\to1}E_{i\to20}\\ \sum_{i=1}^{L}E_{i\to2}E_{i\to1} & \sum_{i=1}^{L}E_{i\to2}E_{i\to2} & \cdots & \sum_{i=1}^{L}E_{i\to2}E_{i\to20}\\ \vdots & \vdots & \ddots & \vdots\\ \sum_{i=1}^{L}E_{i\to20}E_{i\to1} & \sum_{i=1}^{L}E_{i\to20}E_{i\to2} & \cdots & \sum_{i=1}^{L}E_{i\to20}E_{i\to20}\end{bmatrix} \tag{15}$$

which contains $20\times20=400$ elements. Since $\mathbf{MM^T}$ is a symmetric matrix, we only need the information of its 210 elements, of which 20 are the diagonal elements and $(400-20)/2=190$ are the lower triangular elements, to formulate the protein $\mathbf{P}$; i.e., the general PseAAC form of **Eq.5** can now be formulated as

$$\mathbf{P}_{\mathrm{Evo}}=\left[\begin{array}{ccccccc}\psi^{\mathrm{E}}_1 & \psi^{\mathrm{E}}_2 & \cdots & \psi^{\mathrm{E}}_u & \cdots & \psi^{\mathrm{E}}_{210}\end{array}\right]^{\mathbf{T}} \tag{16}$$

where the components $\psi^{\mathrm{E}}_u(u=1,2,\cdots,210)$ are respectively taken from the 210 diagonal and lower triangular elements of **Eq.15** by following a given order, say from left to right and from the 1st row to the last as illustrated by following equation

$$\begin{bmatrix}(1)\\(2) & (3)\\(4) & (5) & (6)\\ \vdots & \vdots & \vdots & \ddots\\(191) & (192) & (193) & \ldots & (210)\end{bmatrix} \tag{17}$$

where the numbers in parentheses indicate the order of elements taken from **Eq.15** for **Eq.16**.

## 3. The Self-consistency Formulation Principle

Regardless of using which formulation to represent protein samples, the following self-consistency principle must be observed during the course of prediction: if the query protein $\mathbf{P}$ was defined in the form of $\mathbf{P}_{\mathrm{GO}}$ (see **Eq.7**), then all the protein samples used to train the prediction engine should also be expressed in the GO formulation; if the query protein was defined in the form of $\mathbf{P}_{\mathrm{Evo}}$ (see **Eq.16**), then all the training data should be expressed in the SeqEvo formulation as well.

Below, let us consider the algorithm or operation engine for conducting the prediction.

## 4. Multi-Label KNN (K-Nearest Neighbor) Classifier

In this study, let us introduce a novel classifier, called the multi-label KNN or abbreviated as ML-KNN classifier, to predict the subcellular localization for the systems that contain both single-location and multiple-location proteins.

Suppose the $m$-th subset $\mathbb{S}_m$ of $\mathbb{S}$ (**Eq.1**) contains $N_m$ Gram-negative proteins, and $\mathbf{P}(m,j)$ is the $j$-th one in that subset. Thus, we have

$$\mathrm{P}(m,j)=\begin{cases}\mathrm{P}_{\mathrm{GO}}(m,j), & \text{in GO space}\\ \mathrm{P}_{\mathrm{Evo}}(m,j), & \text{in SeqEvo space}\end{cases} \tag{18}$$
$$(m=1,2,\cdots,8;j=1,2,\cdots,N_m)$$

where $\mathbf{P}_{\mathrm{GO}}(m,j)$ and $\mathbf{P}_{\mathrm{Evo}}(m,j)$ have the same forms as $\mathbf{P}_{\mathrm{GO}}$(**Eq.7**), and $\mathbf{P}_{\mathrm{Evo}}$(**Eq.16**), respectively; the only difference is that the corresponding constituent elements are derived from the amino acid sequence of $\mathbf{P}(m,j)$ instead of $\mathbf{P}$.

In sequence analysis, there are many different scales to define the distance between two proteins, such as Euclidean distance, Hamming distance [33], and Mahalanobis distance [18,34,35]. In [11], the distance between $\mathbf{P}(m,j)$ and $\mathbf{P}$ was defined by $1-\cos^{-1}[\mathbf{P},\mathbf{P}(m,j)]$. However, we have observed that when the GO descriptor was formulated with real numbers, better outcomes would be resulted by using the Euclidean metric; i.e., the distance between $\mathbf{P}$ and $\mathbf{P}(m,j)$ should be defined here by

$$D\{\mathbf{P},\mathbf{P}(m,j)\}=\|\mathbf{P}-\mathbf{P}(m,j)\| \tag{19}$$

where $\|\mathbf{P}-\mathbf{P}(m,j)\|$ represents the module of the vector difference between $\mathbf{P}$ and $\mathbf{P}(m,j)$ in the Euclidean space. According to **Eq.19**, when $\mathbf{P}\equiv\mathbf{P}(m,j)$ we have $D\{\mathbf{P},\mathbf{P}(m,j)\}=0$, indicating the distance between these two protein sequences is zero and hence they have perfect or 100% similarity.

Suppose $\mathbf{P}^*_1,\mathbf{P}^*_2,\cdots,\mathbf{P}^*_K$ are the $K$ nearest neighbor proteins to the protein $\mathbf{P}$ that forms a set denoted by $\mathbb{S}^{\mathbf{P}}_K$, which is a subset of $\mathbb{S}$; i.e.,$\mathbb{S}^{\mathbf{P}}_K\subseteq\mathbb{S}$. Based on the $K$ nearest neighbor proteins in $\mathbb{S}^{\mathbf{P}}_K$, let us define an accumulation-layer (AL) scale, given by

$$\mathbb{Q}(\mathbf{P},K)=\left\{\begin{array}{cccccc}\rho^K_1 & \rho^K_2 & \cdots & \rho^K_m & \cdots & \rho^K_8\end{array}\right\} \tag{20}$$

where

$$\rho_m=\frac{\sum_{i=1}^{K}\delta(\mathbf{P}^*_i,m)}{\mathbb{N}^*_K}\quad (m=1,2,\cdots,8) \tag{21}$$

where

$$\delta(\mathbf{P}^*_i,m)=\begin{cases}1, & \text{if } \mathbf{P}^*_i \text{ belongs to the } m\text{-th location}\\ 0, & \text{otherwise}\end{cases} \tag{22}$$

and

$$\mathbb{N}^*_K=\sum\nolimits_{m=1}^{8}\sum\nolimits_{i=1}^{K}\delta(\mathbf{P}^*_i,m) \tag{23}$$

Note that $\mathbb{N}^*_K\geq K$ because a protein may belong to one or more subcellular location sites in the current system.

Now, for a query protein $\mathbf{P}$, its subcellular location(s) will be predicted according to the following steps.

**Step 1.** The number of how many different subcellular locations it belongs to will be determined by its nearest neighbor

protein in $\mathbb{S}$. For example, suppose $\mathbf{P}^*$ is the nearest protein to $\mathbf{P}$ in $\mathbb{S}$. If $\mathbf{P}^*$ has only one subcellular location, then $\mathbf{P}$ will also have only one location; if $\mathbf{P}^*$ has two subcellular locations, then $\mathbf{P}$ will also have two locations; and so forth. In general, if $\mathbf{P}^*$ belongs to $\mathfrak{M}$ different location sites, then $\mathbf{P}$ will be predicted to have the same number, $\mathfrak{M}$, of subcellular locations as well, as can be formulated by

$$\mathfrak{M} = \mathrm{Num}\{\mathbf{P}^* \Rightarrow \mathbb{L}\} = \mathrm{Num}\{\mathbf{P} \Rightarrow \mathbb{L}\} \quad (24)$$

where $\mathfrak{M}$ is an integer ($\leq 8$), $\mathrm{Num}\{\mathbf{P}^* \Rightarrow \mathbb{L}\}$ represents the number of different subcellular locations to which $\mathbf{P}^*$ belongs, and $\mathrm{Num}\{\mathbf{P} \Rightarrow \mathbb{L}\}$ the number of different subcellular locations to which $\mathbf{P}$ belongs.

**Step 2.** However, the concrete location site(s) to which $\mathbf{P}$ belongs will not be determined by the location site(s) of $\mathbf{P}^*$, but by the element(s) in **Eq.20** that has (have) the highest score(s), as can be expressed by $\{\ell\}$, the subscript(s) of **Eq.1**. For example, if $\mathbf{P}$ is found belonging to only one location ($\mathfrak{M}=1$) in Step 1, and the highest score in **Eq.20** is $\rho_3^K$, then $\mathbf{P}$ will be predicted as $\{\ell\}=3$ meaning that it belongs to $\mathbb{S}_3$ or resides at "cytoplasm" (cf. **Table 1**). If $\mathbf{P}$ is found belonging to two locations ($\mathfrak{M}=2$), and the first two highest scores in **Eq.20** are $\rho_1^K$ and $\rho_8^K$, then $\mathbf{P}$ will be predicted as $\{\ell\}=(1, 8)$ meaning that it belongs to $\mathbb{S}_1$ and $\mathbb{S}_8$ or resides simultaneously at "cell inner membrane" and "periplasm". And so forth. In other words, the concrete predicted subcellular location(s) can be formulated as

$$\{\ell\} = \mathrm{Max}\triangleright_{\mathbf{Sub}}^{\mathfrak{M}}\{\,\rho_1^K \quad \rho_2^K \quad \cdots \quad \rho_m^K \quad \cdots \quad \rho_8^K\,\} \quad (\mathfrak{M} \leq 8) \quad (25)$$

where the operator "$\mathrm{Max}\triangleright_{\mathbf{Sub}}^{\mathfrak{M}}$" means identifying the $\mathfrak{M}$ highest scores for the elements in the brackets right after it, followed by taking their $\mathfrak{M}$ subscripts.

The entire classifier thus established is called **iLoc-Gneg**, which can be used to predict the subcellular localization of both singleplex and multiplex Gram-negative bacterial proteins. To provide an intuitive picture, a flowchart is provided in **Fig. 2** to illustrate the prediction process of **iLoc-Gneg**.

## 5. Protocol Guide

For user's convenience, a web-server for **iLoc-Gneg** was established. Below, let us give a step-by-step guide on how to use it to get the desired results.

**Step 1.** Open the web server at site http://icpr.jci.edu.cn/bioinfo/iLoc-Gneg and you will see the top page of the predictor on your computer screen, as shown in **Fig. 3**. Click on the Read Me button to see a brief introduction about **iLoc-Gneg** 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. 3**. 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. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box. For more information about FASTA format, visit http://en.wikipedia.org/wiki/Fasta_format. Different with **Gneg-mPLoc** [11], where only one query protein sequence at a time is allowed for each submission, now the maximum number of query proteins for each submission can be 10.

**Step 3.** Click on the Submit button to see the predicted result. For example, if you use the three query protein sequences in the Example window as the input, after clicking the Submit button, you will see **Fig. 4** shown on your screen, indicating that the predicted result for the 1st query protein is "**Cell outer membrane**", that for the 2nd one is "**Cytoplasm; Periplasm**", and that for the 3rd one is "**Cell inner membrane; Cytoplasm**". In other words, the 1st query protein (P0A3N8) is a single-location one residing at "cell outer membrane" only, the 2nd one (Q05097) can simultaneously reside in two different sites ("cytoplasm" and "periplasm"), and the 3rd one (P61380) can also simultaneously reside in two different sites ("cell inner membrane" and "cytoplasm"). All these results are exactly the same as observed by experiments as shown in the Supporting Information S1. It takes about 10 seconds for the above computation before the predicted results appear on your computer screen; the more number of query proteins and longer of each sequence, the more time it is usually needed.

**Step 4.** As shown on the lower panel of **Fig. 3**, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the "Browse" button. To see the sample of batch input file, click on the button Batch-example. The maximum number of the query proteins for each batch input file is 50. After clicking the button Batch-submit, you will see "Your batch job is under computation; once the results are available, you will be notified by e-mail." Note that if you submit a batch input file from an Apple computer, although it looks like in the FASTA format, your input might change to non-FASTA format in the server end and cause errors. Under such a circumstance, the safest way is to submit your input file with a pdf format.

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

**Step 6.** Click on the Data button to download the benchmark datasets used to train and test the **iLoc-Gneg** predictor.

**Caveat.** To obtain the predicted result with the expected success rate, the entire sequence of the query protein rather than its fragment should be used as an input. A sequence with less than 50 amino acid residues is generally deemed as a fragment. Also, if the query Gram-negative protein is known not one of the 8 locations as shown in **Fig. 1**, stop the prediction because the result thus obtained will not make any sense.

## Results and Discussion

In statistical prediction, it would be meaningless to simply report a success rate of a predictor without specifying what method and benchmark dataset were used to test its accuracy [14]. As is well known, the following three methods are often used to examine the quality of a predictor: independent dataset test, subsampling test, and jackknife test [36]. Owing to that subsampling test and jackknife test can be performed with one benchmark dataset and that independent dataset test can be treated as a special case of subsampling test, one benchmark dataset would suffice to serve all the three kinds of cross-validation. However, as demonstrated by Eq.1 of [37] and elucidated in [2], among the three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset and hence has been widely recognized and increasingly used to examine the power of various predictors (see, e.g., [38,39,40,41,42,43,44,45,46,

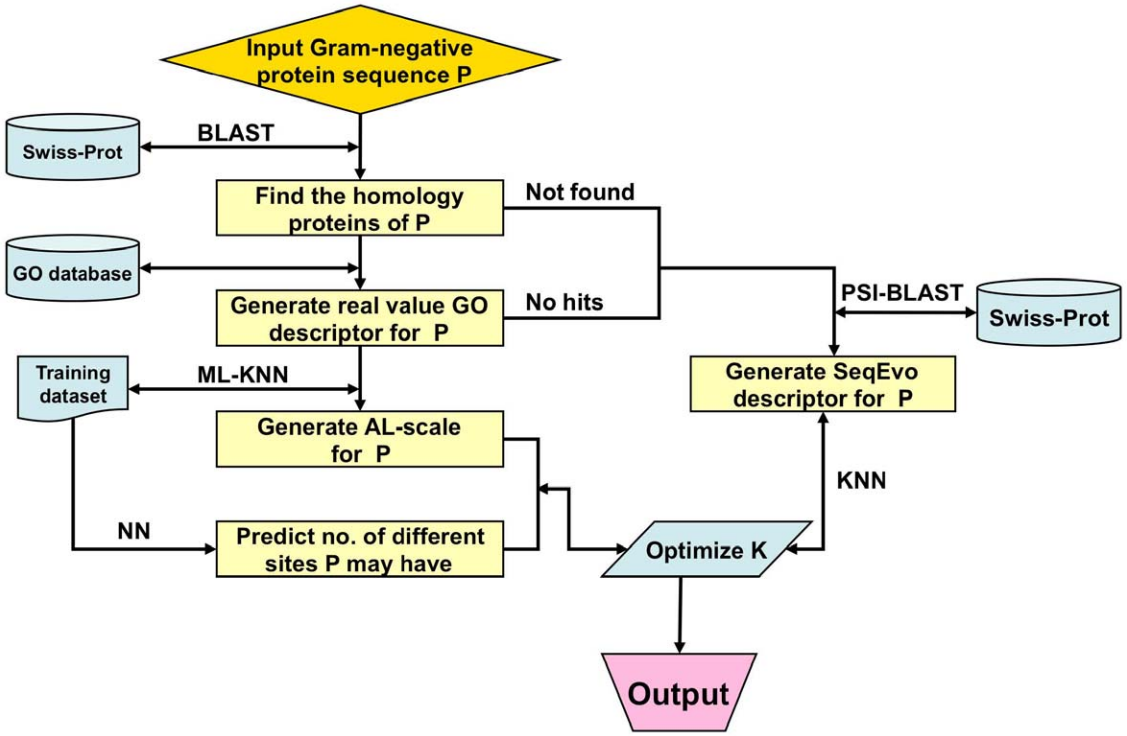


**Figure 2. A flowchart to show the prediction process of iLoc-Gneg.**
doi:10.1371/journal.pone.0020592.g002

47,48,49,50,51,52,53,54,55,56]). Accordingly, in this study, the jackknife test will be adopted to evaluate the power of **iLoc-Gneg** as well.

However, even if using the jackknife test to examine the accuracy, a same predictor may still yield obviously different success rates when tested by different benchmark datasets. This is

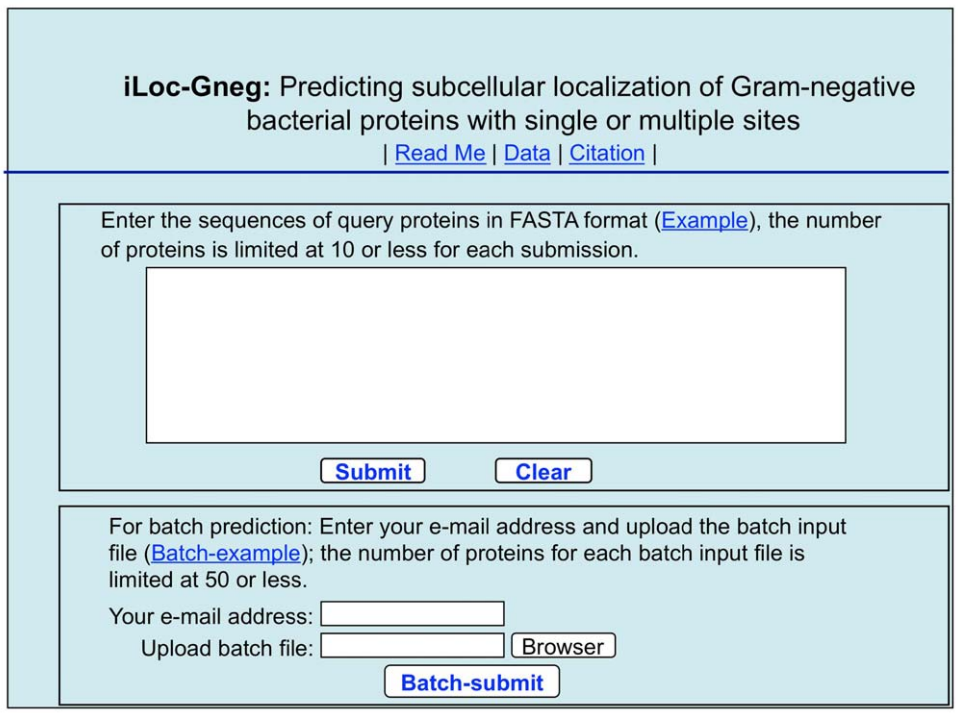


**Figure 3. A semi-screenshot to show the top page of the iLoc-Gneg web-server.** Its website address is at http://icpr.jci.edu.cn/bioinfo/iLoc-Gneg.
doi:10.1371/journal.pone.0020592.g003

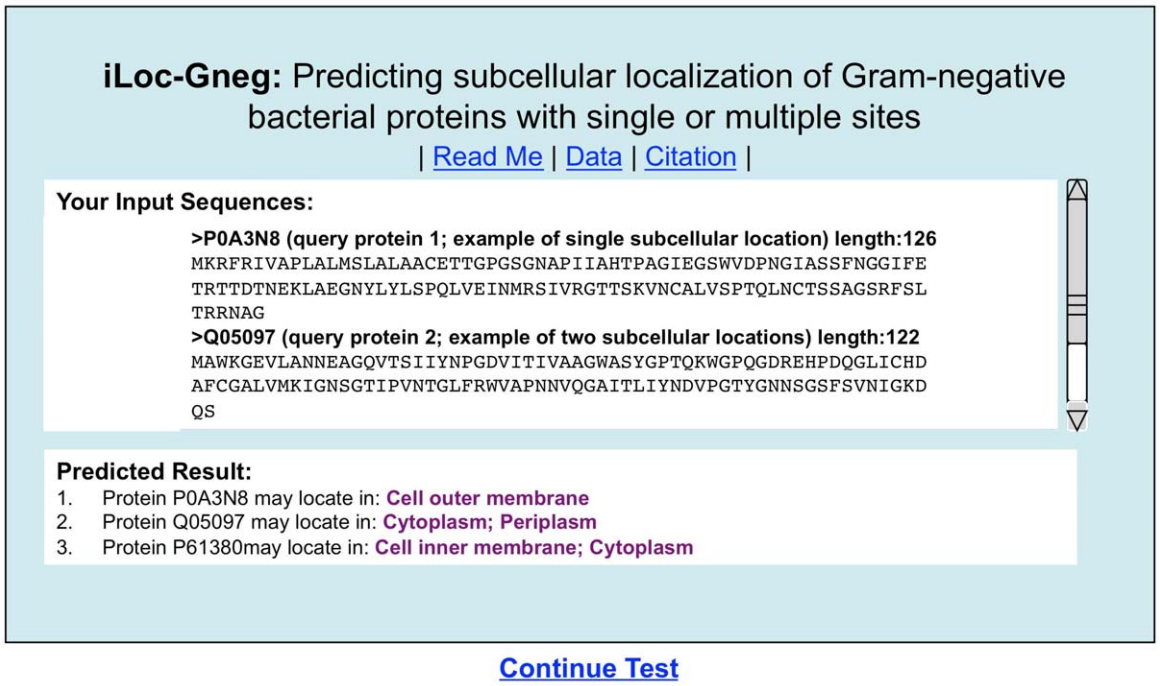


**Figure 4. A semi-screenshot to show the output of iLoc-Gneg.** The input was taken from the three protein sequences listed in the Example window of the **iLoc-Gneg** web-server (cf. Fig. 3).
doi:10.1371/journal.pone.0020592.g004

because the more stringent of a benchmark dataset in excluding homologous sequences, the more difficult for a predictor to achieve a high success rate. Also, the more number of subsets (subcellular locations) a benchmark dataset covers, the more difficult to achieve a high overall success rate, as elaborated in a recent review [14].

As mentioned in the Materials section, the benchmark dataset used in this study is $\mathbb{S}$ (cf. Supporting Information S1), which is the same benchmark dataset constructed in [11] for **Gneg-mPLoc**.

Actually, for such a dataset containing both single-location and multiple-location Gram-negative proteins distributed among 8 subcellular location sites, so far only one existing predictor, i.e., **Gneg-mPLoc** [11], had the capacity to deal with it. Therefore, to demonstrate the power of the current predictor, it would suffice to just compare **iLoc-Gneg** with **Gneg-mPLoc** [11].

Listed in **Table 2** are the results obtained with **Gneg-mPLoc** [11] and **iLoc-Gneg** on the aforementioned benchmark dataset $\mathbb{S}$ by the jackknife test. As we can see from **Table 2**, for such a stringent and complicated benchmark dataset, the overall success rate achieved by **iLoc-Gneg** is over 91.4%, which is about 6% higher than that by **Gneg-mPLoc** [11].

Note that during the course of the jackknife test by **Gneg-mPLoc** and **iLoc-Gneg**, the false positives (over-predictions) and false negatives (under-predictions) were also taken into account to reduce the scores in calculating the overall success rate. As for the detailed process of how to count the over-predictions and under-predictions for a system containing both single-location and multiple-location proteins, see Eqs.43–48 and Fig. 4 in a comprehensive review [2].

To provide a more intuitive and easier-to-understand measurement, let us introduce a new scale, the so-called "absolute true" success rate, to reflect the accuracy of a predictor, as defined by

$$\Lambda = \frac{\sum_{i=1}^{N} \Delta(i)}{N} \tag{26}$$

where $\Lambda$ represents the absolute true rate, $N$ the number of total proteins investigated, and

$$\Delta(i) = \begin{cases} 1, & \text{if all the subcellular locations of the } i\text{-th protein are} \\ & \text{correctly predicted without any overprediction} \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

According to the above definition, for a protein belonging to, say, two subcellular locations, if only one of the two is correctly predicted, or the predicted result contains a location not belonging to the two, the prediction score will be counted as 0. In other words, when and only when all the subcellular locations of a query protein are exactly predicted without any underprediction or overprediction, can the prediction be scored with 1. Therefore, the absolute true scale is much more strict and harsh than the scale used previously [2,11] in measuring the success rate. However, even if using such a stringent criterion on the same benchmark dataset by the jackknife test, the overall absolute true success rate achieved by **iLoc-Gneg** was $1252/1392 = 89.9\%$.

Why can **iLoc-Gneg** enhance the success rate so remarkably? One of the key reasons is that the GO formulation for protein samples in **iLoc-Gneg** contains more information than that in **Gneg-mPLoc** [11], as elaborated as follows. For example, for the protein with the access number "P0A8U0" as denoted by **P**(P0A8U0), according to Steps 3 and 4 in the Section of "GO (Gene Ontology) Formulation", we found 47 proteins that were homologous to it; i.e., $\mathbb{N}^{\text{homo}}_{\mathbf{P}(\text{P0A8U0})} = 47$. Each of the 47 homologous

proteins hit GO:0005886 (or GO_compress:00277) and GO:0016020 (or GO_compress:00830), and hence the two GO numbers were hit by a total of 47 times. Only one of the 47 proteins hit GO:0005737 (or GO_compress: 00269). Substituting these data into Eqs.8–9, we have

$$\psi_u^{\mathrm{G}}(\mathrm{P(P0A8U0)}) = \begin{cases} 1/47 \approx 0.0213, & \text{if } u = 269 \\ 47/47 = 1.0, & \text{if } u = 277 \\ 47/47 = 1.0, & \text{if } u = 830 \\ 0.0, & \text{otherwise} \end{cases} \quad (28)$$

$$(u = 1,2,\cdots,11118)$$

In contrast, if the same protein was represented according to the formulation in **Gneg-mPLoc** [11], it would be

$$\psi_u^{\mathrm{G}}(\mathbf{P}(\mathrm{P0A8U0})) = \begin{cases} 1, & \text{if } u = 269 \\ 1, & \text{if } u = 277 \\ 1, & \text{if } u = 830 \\ 0, & \text{otherwise} \end{cases} \quad (u = 1,2,\cdots,11118) \quad (29)$$

It can be seen by a comparison of Eq.28 with Eq.29 that although the elements in the 269$^{\mathrm{th}}$, 277$^{\mathrm{th}}$, and 830$^{\mathrm{th}}$ components are all not zero in both formulations, the differences of their weights are completely ignored in Eq.29 as formulated in **Gneg-mPLoc** [11]. That is also why, when the sequence of **P**(P0A8U0) was inputted into **iLoc-Gneg** and **Gneg-mPLoc** [11] as a query protein for prediction, the former could accurately predict its both location sites (''cell inner membrane'' and ''cytoplasm''), while the latter could predict only one site (''cell inner membrane'') but miss the site of ''cytoplasm''.

## Conclusions

Prediction of protein subcellular localization is a challenging problem, particularly when the system concerned contains both singleplex and multiplex proteins. The reasons why **iLoc-Gneg** can achieve higher success rates than **Gneg-mPLoc** are as follows. **(1)** The GO formulation used to represent protein samples in **iLoc-Gneg** is formed by the probabilities of hits (cf. Eqs.8–9) and hence contains more information than that in **Gneg-mPLoc** [11] where only the number ''0'' or ''1'' was used regardless how many hits were found to the corresponding component in the GO formulation. **(2)** The accumulation-layer scale has been introduced in **iLoc-Gneg** that is more natural and effective for dealing with proteins having both single and multiple subcellular locations.

## Supporting Information

**Supporting Information S1 This benchmark dataset $\mathbb{S}$ includes 1,456 locative protein sequences (1,392 different proteins), classified into 8 Gram-negative subcellular locations.** Among the 1,392 different proteins, 1,328 belong to one location; and 64 to two locations. Both the accession numbers and sequences are given. None of the proteins has $\geq$25% sequence identity to any other in the same subset (subcellular location). See the text of the paper for further explanation.
(PDF)

## Acknowledgments

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

## Author Contributions

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

## References


1. Nakai K (2000) Protein sorting signals and prediction of subcellular localization. Advances in Protein Chemistry 54: 277–344.
2. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370: 1–16.
3. Nakai K, Kanehisa M (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. Proteins: Structure, Function and Genetics 11: 95–110.
4. Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends in Biochemical Science 24: 34–36.
5. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, et al. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Research 31: 3613–3617.
6. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, et al. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21: 617–623.
7. Chou KC, Shen HB (2006) Large-scale predictions of Gram-negative bacterial protein subcellular locations. Journal of Proteome Research 5: 3420–3428.
8. Smith C (2008) Subcellular targeting of proteins and drugs. http://www.biocompare.com/Articles/FeaturedArticle/976/Subcellular-Targeting-Of-Proteins-And-Drugs.html.
9. Glory E, Murphy RF (2007) Automated subcellular location determination and high-throughput microscopy. Dev Cell 12: 7–16.
10. Millar AH, Carrie C, Pogson B, Whelan J (2009) Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. Plant Cell 21: 1625–1631.
11. Shen HB, Chou KC (2010) Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. Journal of Theoretical Biology 264: 326–333.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 25–29.
13. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32: D262–266.
14. 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.
15. Chou KC, Shen HB (2009) Review: recent advances in developing web-servers for predicting protein attributes. Natural Science 2: 63–92 (openly accessible at http://www.scirp.org/journal/NS/).
16. Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. In: Suhai S, ed. Theoretical and Computational Methods in Genome Research. New York: Plenum. pp 1–14.
17. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 17: 149–163.
18. Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. Journal of Biological Chemistry 269: 22014–22020.
19. Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238: 54–61.
20. Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function & Genetics 21: 319–344.
21. Cedano J, Aloy P, P'erez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266: 594–600.
22. Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Research 26: 2230–2236.
23. Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Engineering 12: 107–118.
24. Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. PROTEINS: Structure, Function, and Genetics 50: 44–48.
25. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol 44, 60) 43: 246–255.
26. 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.

27. 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.
28. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, et al. (2009) Protein function annotation by homology-based inference. Genome Biol 10: 207.
29. Gerstein M, Thornton JM (2003) Sequences and topology. Curr Opin Struct Biol 13: 341–343.
30. Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11: 2105–2134.
31. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, et al. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res 13: 662–672.
32. Chou KC (1995) The convergence-divergence duality in lectin domains of the selectin family and its implications. FEBS Letters 363: 123–126.
33. Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis: Chapter 11 Discriminant Analysis; Chapter 12 Multivariate analysis of variance; Chapter 13 cluster analysis (pp. 322–381). London: Academic Press. pp 322–381.
34. Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2: 49–55.
35. Pillai KCS (1985) Mahalanobis D2. In: Kotz S, Johnson NL, eds. Encyclopedia of Statistical Sciences. New York: John Wiley & Sons, This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics. pp 176–181.
36. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30: 275–349.
37. 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 (openly accessible at http://www.scirp.org/journal/NS/).
38. Cai YD, He J, Li X, Feng K, Lu L, et al. (2010) Predicting protein subcellular locations with feature selection and analysis. Protein Pept Lett 17: 464–472.
39. Jahandideh S, Hoseini S, Jahandideh M, Hoseini A, Disfani FM (2009) Gamma-turn types prediction in proteins using the two-stage hybrid neural discriminant model. Journal of Theoretical Biology 259: 517–522.
40. Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein & Peptide Letters 16: 27–31.
41. Kannan S, Hauth AM, Burger G (2008) Function prediction of hypothetical proteins without sequence similarity to proteins of known function. Protein & Peptide Letters 15: 1107–1116.
42. Chen C, Chen LX, Zou XY, Cai PX (2008) Predicting protein structural class based on multi-features fusion. Journal of Theoretical Biology 253: 388–392.
43. Chen K, Kurgan LA, Ruan J (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J Comput Chem 29: 1596–1604.
44. Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein & Peptide Letters 16: 351–355.
45. Du P, Cao S, Li Y (2009) SubChlo: predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm. Journal of Theoretical Biolology 261: 330–335.
46. 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.
47. Gao QB, Jin ZC, Ye XF, Wu C, He J (2009) Prediction of nuclear receptors with optimal pseudo amino acid composition. Analytical Biochemistry 387: 54–59.
48. Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128: 87–93.
49. Jahandideh S, Sarvestani AS, Abdolmaleki P, Jahandideh M, Barfeie M (2007) gamma-Turn types prediction in proteins using the support vector machines. J Theor Biol 249: 785–790.
50. 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.
51. Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. Journal of Theoretical Biology 252: 350–356.
52. Masso M, Vaisman II (2010) Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. Journal of Theoretical Biology 266: 560–568.
53. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
54. 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.
55. 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.
56. 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.