Prediction of Body Fluids where Proteins are Secreted into Based on Protein Interaction Network

Le-Le Hu[1,2], Tao Huang[3,4], Yu-Dong Cai[1,5,6]*, Kuo-Chen Chou[6]

[1] Institute of Systems Biology, Shanghai University, Shanghai, China, [2] Department of Chemistry, College of Sciences, Shanghai University, Shanghai, China, [3] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [4] Shanghai Center for Bioinformation Technology, Shanghai, China, [5] Centre for Computational Systems Biology, Fudan University, Shanghai, China, [6] Gordon Life Science Institute, San Diego, California, United States of America

Abstract

Determining the body fluids where secreted proteins can be secreted into is important for protein function annotation and disease biomarker discovery. In this study, we developed a network-based method to predict which kind of body fluids human proteins can be secreted into. For a newly constructed benchmark dataset that consists of 529 human-secreted proteins, the prediction accuracy for the most possible body fluid location predicted by our method via the jackknife test was 79.02%, significantly higher than the success rate by a random guess (29.36%). The likelihood that the predicted body fluids of the first four orders contain all the true body fluids where the proteins can be secreted into is 62.94%. Our method was further demonstrated with two independent datasets: one contains 57 proteins that can be secreted into blood; while the other contains 61 proteins that can be secreted into plasma/serum and were possible biomarkers associated with various cancers. For the 57 proteins in first dataset, 55 were correctly predicted as blood-secrete proteins. For the 61 proteins in the second dataset, 58 were predicted to be most possible in plasma/serum. These encouraging results indicate that the network-based prediction method is quite promising. It is anticipated that the method will benefit the relevant areas for both basic research and drug development.

Citation: Hu L-L, Huang T, Cai Y-D, Chou K-C (2011) Prediction of Body Fluids where Proteins are Secreted into Based on Protein Interaction Network. PLoS ONE 6(7): e22989. doi:10.1371/journal.pone.0022989

Editor: Peter Csermely, Semmelweis University, Hungary

Received April 25, 2011; **Accepted** July 8, 2011; **Published** July 29, 2011

Funding: These authors have no support or funding to report.

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

* E-mail: cai_yud@yahoo.com.cn

Introduction

Protein secretion is a universal biological process occurring in all organisms. Secreted proteins such as hormones, digestive enzymes, neurotransmitters as well as antibodies, play vital regulatory roles in various biological activities such as reproduction, digestion, nerve conduction and immunization [1]. The studies on the secreted proteins under different physiological and pathological conditions in different growth and development stages can deepen the understanding of many biological phenomena. Under the condition of the disease, some secreted proteins showed abnormal concentration level [2]. In recent years, several genes encoding secreted proteins have been found to be consistently over-expressed in various cancer specimens [3,4,5]. For example, MIC1 gene has been observed to be over-expressed in breast, colorectal and prostate cancer patients [5]. These proteins that could be detected in blood, urine or other body fluids are more suitable to serve as biomarkers for diagnosis [6]. This is because the body fluid test (e.g. blood test or urine test) is less invasive, cheaper, and easier to collect and process samples than tissue biopsy test [7,8] since the latter requires surgery to get the disease tissues. Besides, identification of body fluids where proteins can be secreted into is very helpful for protein function annotation and biomarker discovery.

However, how to realize the identification is still a big challenge even having the advanced proteomics technologies because there are a large amount of proteins with a variety of modifications in body fluids [8]. To address this problem, let us resort to computational approaches. In the past two decades, many studies have focused on predicting the subcellular locations of proteins in both prokaryotes and eukaryotes (see, e.g., [9,10,11,12,13,14,15,16,17,18,19,20,21] as well as a long list of the relevant references in a comprehensive review [22]). Unfortunately, none of the aforementioned methods was aimed at identifying the final locations where the extracellular proteins are secreted. The present study was initiated in an attempt to address this problem, with a focus on human secreted proteins and a novel approach via protein-protein interaction (PPI) network.

According to a recent comprehensive review [23], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, let us describe how to deal with these steps.

Materials and Methods

Training dataset

The human secreted proteins were retrieved from UniProt [24]. The detailed procedures for collecting the human secreted protein sequences are as follows. **(1)** Open the web-page at http://www.uniprot.org/ (Release 2011_05). **(2)** Click the button "Fields", followed by selecting "Subcellular location" for Advanced Search, "Secreted" for Term, "Homo sapiens" for Organism, and "Experimental" for Confidence. **(3)** Click Add & Search. Thus we collected a total of 1,019 experiment-validated human secreted proteins. Subsequently, these proteins were mapped to 11 different kinds of body fluids contained in the human body fluid database "Sys-BodyFluid" [25] (http://lifecenter.sgst.cn/bodyfluid/), where the body fluid proteome data was manually collected from 50 peer-review publications. Finally, a total of 682 human proteins have been obtained that can be secreted into the aforementioned body fluids.

The human protein-protein interaction (PPI) networks were retrieved from STRING [26,27] (http://string.embl.de/), which is a database dedicated to both physical and functional interactions. Information derived from 3 kinds of sources (high-throughput experiments, mining of databases and literature, and prediction from genomic context analysis) was integrated into several PPI networks. As done by previous investigators in using the intuitive graphic representation to deal with complicated biological systems, such as enzyme-catalyzed system [28,29,30], protein-folding system [31], and drug metabolism system [32], here the PPI network can also be intuitively expressed via a graph, in which each of the proteins is represented by a node, and the interaction is represented by the edge between two nodes. The edge is weighted by the interaction confidence, i.e., the likelihood that the interaction exists between two nodes. The interaction confidence score of two proteins is obtained as follows: first, the interactions from each source were scored by benchmarking them against a common reference set; then these scores were combined in the naive Bayesian fashion [26].

Of the 682 human secreted proteins, we have found that 153 proteins have no PPI information nor interact with any of the other secreted proteins, while 529 proteins interact with at least one of the other proteins in the human PPI network from STRING. Thus, we obtained a working PPI network that consists of 529 nodes (proteins) and 27,176 interaction units. Such 529 human secreted proteins in the newly constructed PPI network were used as the training dataset for developing the current network-based method.

The distribution of the 529 human secreted proteins classified according to the 11 different types of body fluids is shown in **Table 1**, from which we can see that the sum of numbers in column 3 is 1708 that is much more than 529, the number of secreted proteins. This is because many proteins can be secreted into more than one body fluid [25], as illustrated in **Figure 1**. As we can see from the figure, of the 529 human secreted proteins, 179 can be secreted into one body fluid, and 350 proteins can be secreted into two or more different types of body fluids. Therefore, we are to deal with a multi-label classification problem.

Testing datasets

Two testing datasets were used in this study. The first one contains 57 blood-secreted proteins, which was obtained as follows. First of all, 305 blood-secreted proteins were retrieved from the positive dataset in [33], where the proteins met the criteria that they were not only secreted but also serum/plasma detected. Of the 305 proteins thus obtained, 172 were excluded because they occurred the training dataset, and 76 proteins were also excluded because they had no interaction with the proteins in the training dataset and hence could not be processed by the current method (see the Network-based Method section). The remaining 57 blood-secreted proteins were used to test our method (**Table S1**).

Table 1. A breakdown of the 529 human secreted proteins in the training dataset according to the 11 different types of body fluids into which they can be secreted.

Type	Body fluid	Number of proteins in dataset
1	Amniotic fluid	192
2	Bronchoalveolar lavage fluid	65
3	Cerebrospinal fluid	204
4	Milk	71
5	Nipple aspiration fluid	37
6	Plasma/Serum	418
7	Saliva	175
8	Seminal fluid	155
9	Synovial fluid	63
10	Tear	84
11	Urine	244
Sum		1,708

doi:10.1371/journal.pone.0022989.t001

The second testing dataset contains 61 proteins as obtained as follows. From [33], we first collected 122 abnormally expressed proteins involved with various cancers as indicated by many published proteomics studies. From these proteins, we obtained 77 plasma/serum secreted proteins. After removing those that had been contained in the training dataset and those that had no interaction with the proteins in the training dataset, we finally obtained the remaining 61 possible marker proteins (**Table S2**) for the second testing dataset.

Network-based method

Many interacting proteins must co-occur in the same location to participate in the biological processes [34]. Accordingly, we can presume that the interacting secreted proteins are likely to be secreted into the same body fluids. In other words, the following assumptions would be valid.

Given a query protein, the higher interaction confidence score between it and its interacting counterpart, the more likely they are to be secreted into the same body fluid. Also, the more its interacting proteins in a certain body fluid, the more likely it is to be secreted into such body fluid [35]. With these points in mind, the body fluids that secreted proteins can be secreted into can be predicted as follows.

First, let us denote the n proteins in the PPI network as $\{\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_n\}$ and the 11 body fluids as $F = [F_1, F_2, ..., F_{11}]$, where F_1 stands for the "Amniotic fluid", F_2 the "Bronchoalveolar lavage fluid", F_3 the "Cerebrospinal fluid", and so forth (cf. **Table 1**). Thus, the body fluids that the proteins in the PPI network is secreted into can be described as

$$\mathbb{F} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,j} & \cdots & f_{1,11} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,j} & \cdots & f_{2,11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{i,1} & f_{i,2} & \cdots & f_{i,j} & \cdots & f_{i,11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,j} & \cdots & f_{n,11} \end{bmatrix} \quad \begin{pmatrix} i = 1,2,\cdots,n \\ j = 1,2,\cdots,11 \end{pmatrix} \qquad (1)$$

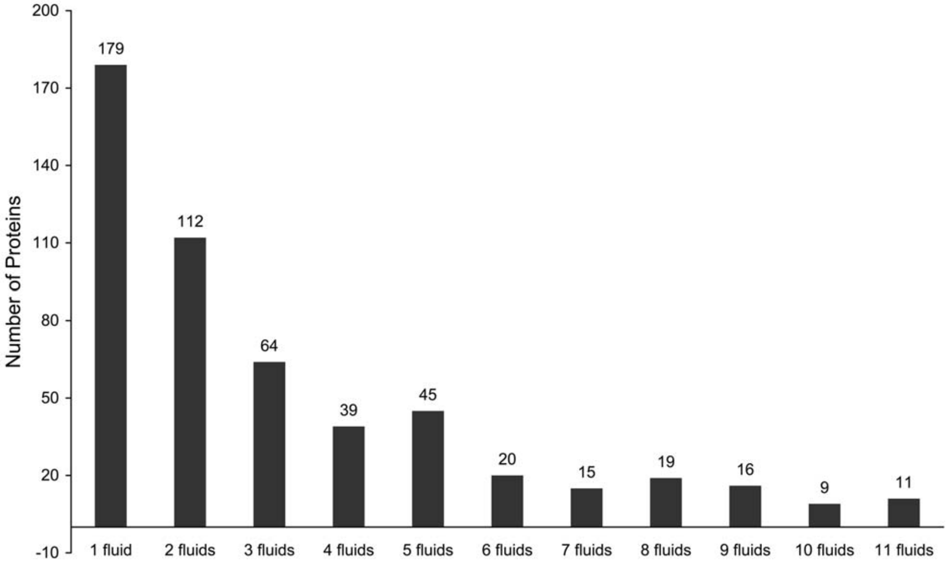

Figure 1. The numbers of proteins that are secreted in different types of body fluids. See Table 1 for the definition of the numerical codes used here for the body fluid types.
doi:10.1371/journal.pone.0022989.g001

where

$$f_{i,j}=\begin{cases}1, & \text{if protein } \mathbf{P}_i \text{ can be secreted into the } j\text{-th} \\ & \text{body fluid } F_j \\ 0, & \text{otherwise}\end{cases} \tag{2}$$

For several query proteins $\{\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_m\}$, their interactions with the n proteins in the PPI network can be described as

$$\mathbb{W}=\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,i} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,i} & \cdots & w_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{k,1} & w_{k,2} & \cdots & w_{k,i} & \cdots & w_{k,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,i} & \cdots & w_{m,n} \end{bmatrix}\left(\begin{array}{l} i=1,2,\cdots,n \\ k=1,2,\cdots,m \end{array}\right) \tag{3}$$

where $w_{k,i}$ denotes the interaction confidence score [27] between $\mathbf{P}_k$ and $\mathbf{P}_i$. If there is no interaction between $\mathbf{P}_k$ and $\mathbf{P}_i$, we have $w_{k,i}=0$. Since no self-interaction exists in the PPI network, $w_{k,i}=0$ if $k=i$. Now, let us use $S(\mathbf{P}_k\Rightarrow j)$to denote the likelihood that the query protein $\mathbf{P}_k$ is secreted into the j-th body fluid F_j. Thus, the likelihood that the m query proteins are secreted into the 11 body fluids can be formulated as

$$\mathbb{M}=\mathbb{W}\cdot\mathbb{F}$$

$$=\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,i} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,i} & \cdots & w_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{k,1} & w_{k,2} & \cdots & w_{k,i} & \cdots & w_{k,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,i} & \cdots & w_{m,n} \end{bmatrix}\begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,j} & \cdots & f_{1,11} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,j} & \cdots & f_{2,11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{i,1} & f_{i,2} & \cdots & f_{i,j} & \cdots & f_{i,11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,j} & \cdots & f_{n,11} \end{bmatrix}$$

$$=\begin{bmatrix} S(\mathbf{P}_1\Rightarrow 1) & S(\mathbf{P}_1\Rightarrow 2) & \cdots & S(\mathbf{P}_1\Rightarrow j) & \cdots & S(\mathbf{P}_1\Rightarrow 11) \\ S(\mathbf{P}_2\Rightarrow 1) & S(\mathbf{P}_2\Rightarrow 2) & \cdots & S(\mathbf{P}_2\Rightarrow j) & \cdots & S(\mathbf{P}_2\Rightarrow 11) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S(\mathbf{P}_k\Rightarrow 1) & S(\mathbf{P}_k\Rightarrow 2) & \cdots & S(\mathbf{P}_k\Rightarrow j) & \cdots & S(\mathbf{P}_k\Rightarrow 11) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S(\mathbf{P}_m\Rightarrow 1) & S(\mathbf{P}_m\Rightarrow 2) & \cdots & S(\mathbf{P}_m\Rightarrow j) & \cdots & S(\mathbf{P}_m\Rightarrow 11) \end{bmatrix} \tag{4}$$

where

$$S(\mathbf{P}_k\Rightarrow j)=\sum\nolimits_{i=1}^{n} w_{k,i}f_{i,j} \tag{5}$$

The 11 elements of each row in Eq.4 represent the likelihoods that protein $\mathbf{P}_k$ is secreted into the 11 body fluids, respectively. It can be seen from Eq.5 that the likelihood $S(\mathbf{P}_k\Rightarrow j)$ can be formulated as the sum of the interaction confidence scores of the protein $\mathbf{P}_k$ with its interacting proteins that can be secreted into the j-th body fluid F_j. Such scoring approach takes both the interaction confidence score and the number of the interacting proteins into consideration, just like the weighted vote. Obviously, the higher the score, the more likely $\mathbf{P}_k$ is to be secreted into the j-th body fluid F_j. In Eq.4, the 11 scores in the k-th row for the query protein $\mathbf{P}_k$ are used to reflect the likelihoods that it is secreted into the 11 body fluids, respectively. Accordingly, the most likely body fluid F_μ where $\mathbf{P}_k$ is secreted should be the one with the maximum score, as can be formulated below

$$\mu=\textbf{arg max}_j\{S(\mathbf{P}_k\Rightarrow j)|j=1,2,\cdots,11\} \tag{6}$$

where μ is the j that maximizes the value of $S(\mathbf{P}_k\Rightarrow j)$.

Since many secreted proteins can be secreted into more than one body fluid, our method is dedicated to provide flexible information by predicting possible body fluids for secreted proteins, rather than the most likely body fluid. To realize this, let us sort the 11 elements of each row in Eq.4 according to descending order. By doing so, we obtain a $m \times 11$ matrix as formulated by

$$\mathbf{D}^{\downarrow}\mathbb{M} = \begin{bmatrix} Q^1(\mathbf{P}_1) & Q^2(\mathbf{P}_1) & \cdots & Q^u(\mathbf{P}_1) & \cdots & Q^{11}(\mathbf{P}_1) \\ Q^1(\mathbf{P}_2) & Q^2(\mathbf{P}_2) & \cdots & Q^u(\mathbf{P}_2) & \cdots & Q^{11}(\mathbf{P}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Q^1(\mathbf{P}_k) & Q^2(\mathbf{P}_k) & \cdots & Q^u(\mathbf{P}_k) & \cdots & Q^{11}(\mathbf{P}_k) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Q^1(\mathbf{P}_m) & Q^2(\mathbf{P}_m) & \cdots & Q^u(\mathbf{P}_m) & \cdots & Q^{11}(\mathbf{P}_m) \end{bmatrix} \quad (7)$$

$$\begin{pmatrix} k=1,2,\cdots,m \\ u=1,2,\cdots,11 \end{pmatrix}$$

where $\mathbf{D}^{\downarrow}$ is a descending operator that arranges the 11 $S(\mathbf{P}_k \Rightarrow j)$ of each row in Eq.4 in descending order: $Q^1(\mathbf{P}_k) \geq Q^2(\mathbf{P}_k) \geq \cdots \geq Q^u(\mathbf{P}_k) \geq \cdots \geq Q^{11}(\mathbf{P}_k)$. If two or more elements of the row in Eq.4 are equal to one another, they will be sorted in random order. Accordingly, the predicted results for the secreted protein $\mathbf{P}_k$ can be obtained from the descending order. For instance, if $Q^1(\mathbf{P}_k) = S(\mathbf{P}_k \Rightarrow 3)$, $Q^2(\mathbf{P}_k) = S(\mathbf{P}_k \Rightarrow 6)$, and $Q^3(\mathbf{P}_k) = S(\mathbf{P}_k \Rightarrow 11)$, then that the query protein $\mathbf{P}$ is secreted into the 3rd body fluid (Cerebrospinal fluid) will have the maximum likelihood (cf. **Table 1**), that $\mathbf{P}$ is secreted into the 6th body fluid (Plasma/Serum) will have the second maximum likelihood, and that $\mathbf{P}$ is secreted into the 11th body fluid (Urine) will have the third maximum likelihood. And so forth. The predicted results thus obtained are called the 1st order predicted body fluid, the 2nd order predicted body fluid, the 3rd order predicted body fluid, and so forth.

Validation and Demonstration

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [36]. However, of the three test methods, the jackknife test is deemed the most objective [37,38]. The reasons are as follows. **(1)** For the independent dataset test, although all the proteins used to test the predictor are outside the training dataset used to train it so as to exclude the "memory" effect or bias, the way of how to select the independent proteins to test the predictor could be quite 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 testing dataset might fail to keep so when tested by another independent testing dataset [36]. **(2)** For the subsampling test, the concrete procedure usually used in literatures is the 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 dataset is an astronomical figure even for a very simple dataset, as demonstrated by Eqs.28–30 in [23]. 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 dataset 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. **(3)** In the jackknife test, all the proteins in the benchmark dataset 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 dataset and testing dataset 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 dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. Accordingly, the jackknife test has been increasingly and widely used by those investigators with strong math background to examine the quality of various predictors (see, e.g., [39,40,41,42,43,44,45,46,47,48]). In view of this, here the jackknife cross-validation was also used to examine the prediction quality of the network-based method. Meanwhile, just for a demonstration to show biologists how to use the predictor for practical application, we also performed the computation for some independent datasets.

For the j-th order prediction, the accuracy Φ_j obtained by the jackknife test can be formulated as

$$\Phi_j = \frac{M_j}{N} \qquad (j=1,2,...,11) \qquad (8)$$

where M_j represents the number of the secreted proteins whose j-th order predicted body fluid is one of the true body fluids where the protein is secreted, and N represents the total number of proteins in the PPI network. These 11-order jackknife cross-validation accuracies were used as an evaluation for the network-based method. According to Eq.8, high Φ_j with small j and low Φ_j with large j will indicate a good prediction based on the current prediction method.

In the PPI network, the average number of body fluids that each secreted protein is secreted into can be calculated by

$$\langle \Phi \rangle = \frac{\sum_{i=1}^{n} E_i}{N} \qquad (9)$$

where E_i represents the number of body fluids that the secreted protein $\mathbf{P}_i$ is secreted into. Hence, a new evaluation for the network-based method was proposed to calculate the likelihood that the first k order predicted body fluids contain all the true body fluids that the proteins can be secreted; it can be formulated as

$$L_k = \frac{\sum_{u=1}^{k} \Phi_u}{\sum_{j=1}^{11} \Phi_j} \qquad (10)$$

where k represents the smallest integer equal or greater than $\langle \Phi \rangle$ of Eq.9. Also, a large L_k indicates a good prediction of the network-based method.

Results and Discussion

Performance of network-based method

In this study, the network-based method was applied to the 529 human secreted proteins to predict the body fluids where they were secreted. All the 11 order jackknife cross-validation

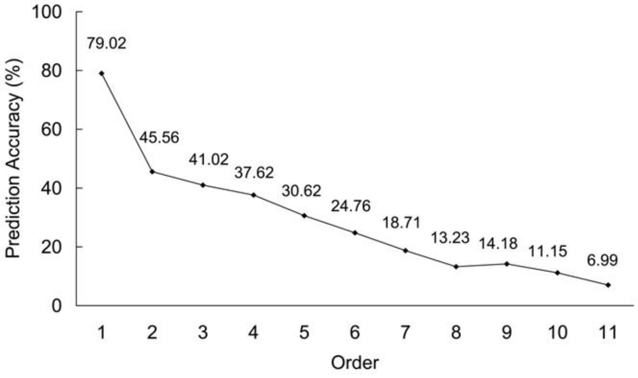

Figure 2. All the 11 order jackknife cross-validation accuracies by the network-based method for the 529 human secreted proteins.
doi:10.1371/journal.pone.0022989.g002

accuracies are shown in **Figure 2**. From the downward-slope curve, we can see that except the 8th-order prediction accuracy, all the other higher-order prediction accuracies are higher than the lower-order ones, indicting that the body fluids were well prioritized for the proteins by the method. The 1st-order (most likely) prediction accuracy is 79.02%, indicating that the 1st-order predicted body fluid for the secreted proteins is believable. The 11th-order (least likely) prediction accuracy is 6.99%, indicting that the likelihood that the query protein is secreted into the 11th-order predicted body fluid is very low and such predicted body fluid can be ignored.

The average number of body fluids that each secreted protein in the PPI network can be secreted into is 3.23 according to Eq.9. Hence, a random guess of body fluid for the secreted proteins will have a 29.36% (3.23/11) success rate, lower than the first 5 order prediction accuracies. The parameter *k* (cf. Eq.10) was set to be 4 ([3.23]+1), i.e., we consider the first 4 order predicted body fluids from the 11 order prediction. The likelihood that the first 4 order predicted body fluids contain all the true body fluids that the proteins can be secreted into is 0.6294 according to Eq.10, indicating that the first 4 order predicted body fluids should be paid more attention to than others in the 11 predicted body fluids.

The availability of using the PPI information to predict the body fluids that secreted proteins can be secreted into

Many important biological activities are mediated by proteins interactions. The interacting proteins should co-occur spatially and temporally to intact with each other [34]. Similarly, the interacting secreted proteins often are secreted into the same body fluids to perform their functions. For example, peptidoglycan recognition protein 1 (O75594, UniProt Protein) can be secreted into plasma/serum [49], saliva [50,51], and urine [52,53,54]. Its interactions with the other proteins are shown in **Table 2**. Except 3 proteins (P07492, Q13410, and P05814), the other 20 neighbor

Table 2. Interactions of peptidoglycan recognition protein 1 (O75594, UniProt Protein) with its neighbor proteins in the PPI network.

Protein A	Body fluid type number[a]	Protein B	Body fluid type number[a]	Interaction confidence
O75594	6, 7, 11	P61626	1, 2, 3, 4, 6, 7, 8, 10, 11	0.532
O75594	6, 7, 11	O15263	7	0.501
O75594	6, 7, 11	P05231	6	0.300
O75594	6, 7, 11	P13500	6	0.291
O75594	6, 7, 11	P60022	6, 11, 7	0.291
O75594	6, 7, 11	P01350	6	0.286
O75594	6, 7, 11	P78380	11	0.279
O75594	6, 7, 11	P07492	8	0.257
O75594	6, 7, 11	P02743	3, 6, 7, 8, 9, 10, 11	0.249
O75594	6, 7, 11	P05120	6, 7, 10	0.243
O75594	6, 7, 11	P35858	1, 3, 6, 9, 11	0.235
O75594	6, 7, 11	P49913	1, 6, 7, 8, 11	0.232
O75594	6, 7, 11	P01375	6	0.227
O75594	6, 7, 11	Q13410	4, 5	0.221
O75594	6, 7, 11	P48023	6	0.218
O75594	6, 7, 11	P19883	6	0.207
O75594	6, 7, 11	P05814	3, 4, 5	0.196
O75594	6, 7, 11	P11226	6	0.191
O75594	6, 7, 11	Q14116	6	0.162
O75594	6, 7, 11	P13236	6	0.156
O75594	6, 7, 11	P02788	1, 3, 5, 6, 7, 8, 10, 11	0.154
O75594	6, 7, 11	P13501	6	0.154
O75594	6, 7, 11	P13591	3, 6, 11	0.154

[a]See **Table 1** for the definition of the body fluid type number.
doi:10.1371/journal.pone.0022989.t002

Table 3. The prediction accuracies with 11 different orders for the 57 blood-secreted proteins by the network-based method, with order 1 corresponding to the most likely prediction and order 11 the least likely prediction.

Order	Accuracy (%)
1	96.49
2	3.51
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0

doi:10.1371/journal.pone.0022989.t003

proteins can be secreted into the plasma/serum or saliva or urine just like peptidoglycan recognition protein 1. According to the prediction criteria, when peptidoglycan recognition protein 1 was considered as a query protein, the first three order predicted body fluids that peptidoglycan recognition protein 1 can be secreted into are plasma/serum, saliva, and urine, which are consistent with the real locations.

Further demonstration

Now, let us demonstrate the prediction method on an independent testing dataset that contains 57 blood-secreted proteins (**Table S1**). The 11 order prediction accuracies for the 57 blood-secreted proteins by the network-based method are listed in **Table 3**. The 1st prediction accuracy is 96.49%, i.e., 55 of 57 proteins were predicted to be secreted into plasma/serum in the 1st prediction. And the 2nd prediction accuracy is 3.51%, and all the other accuracies are 0. In other words, the first 2 predictions cover the secreted locations of all the 57 blood-secreted proteins. Apparently, the results indicate a good performance of the network-based method for secreted proteins in blood. Except the proteins in the training dataset and the 57 blood-secreted proteins, few secreted proteins in other body fluids have been found in the present researches. Therefore our method was evaluated on the blood-secreted proteins.

Table 4. The prediction accuracies with 11 different orders for the 61 marker proteins by the network-based method, with order 1 corresponding to the most likely prediction and order 11 the least likely prediction.

Order	Accuracy (%)
1	95.08
2	3.28
3	1.64
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0

doi:10.1371/journal.pone.0022989.t004

Disease biomarker discovery

The 61 possible marker proteins listed in **Table S2** were also used to demonstrate our method. The 11 order prediction accuracies for the 61 marker proteins are listed in **Table 4**. The 1st prediction accuracy is 95.08%, indicating 58 of 61 proteins were predicted to be secreted into plasma/serum in the 1st prediction. The remaining 3 proteins were arranged into the plasma/serum in the 2nd and 3rd prediction. The collected 61 biomarkers were well arranged into the correct body fluid (plasma/serum).

Based on the quite promising results obtained through this study, we can now propos a way to discover disease biomarker in body fluids. After screening the proteins showing abnormal expression levels in various diseases and identifying their subcellular locations [11,12,13,14,15,18,19], they can be arranged into body fluids using our method. Therefore, suitable biomarkers, such as proteins in plasma/serum or urine can be discovered.

Application and improvement

As is discussed above, the predicted body fluids of the first 4 orders can be regarded as the candidate locations of the secreted proteins. Biologists can focus on these body fluid candidates, which can save a lot of time and labor so as to accelerate the research progress. The predicted body fluids with the last 5 or 6 orders might be excluded for consideration owing to their low accuracies.

Considering the effectiveness of the network-based method for human secreted protein, it is possible to apply the current method to predict the locations of secreted proteins in other species. The PPI network can be collected from numerous sources including STRING [27] (Version 8.0 covered 630 organisms), worm PPI database [15], fly database [55], human PPI database [56,57,58], BIND [59], BioGRID [60], CYGD [61], DIP [62], HPRD [63], MINT [64], IntAct [65], and so forth. Based on the approach proposed in this paper, we can predict the body fluids for proteins of other organisms as well.

The performance of the network-based method can be further improved via the following two avenues. The first one is to collect the PPI data of high quality to exclude the false positive interaction, which was expected to improve the prediction accuracies. The second way is to collect as much PPI data as possible for constructing the PPI network, which was expected to make the method cover as many secreted proteins as possible.

Conclusion

In this study, a multi-target model was developed for assigning the human secreted proteins to the body fluid categories based on the PPI network. Since it is the first computational method to annotate the body fluids where human protein can be secreted into, it is anticipated that the method will benefit the relevant experimental researches and stimulate a series of follow-up investigations into this emerging and challenging area.

Supporting Information

Table S1 The 57 blood-secreted proteins used to test the network-based method.
(DOC)

Table S2 The 61 abnormally expressed proteins (possible biomarkers) used to test the network-based method that were involved with various cancers.
(DOC)

Acknowledgments

We thank UniProt, Sys-BodyFluid and STRING databases for providing data to support not-for-profit research efforts.

Author Contributions

Conceived and designed the experiments: L-LH TH Y-DC. Performed the experiments: L-LH. Analyzed the data: L-LH Y-DC K-CC. Contributed reagents/materials/analysis tools: L-LH TH. Wrote the paper: L-LH TH Y-DC K-CC.

References

1. Chen Y, Zhang Y, Yin Y, Gao G, Li S, et al. (2005) SPD–a web-based secreted protein database. Nucleic acids research 33: D169–173.
2. Huang LJ, Chen SX, Huang Y, Luo WJ, Jiang HH, et al. (2006) Proteomics-based identification of secreted protein dihydrodiol dehydrogenase as a novel serum markers of non-small cell lung cancer. Lung cancer 54: 87–94.
3. Buckhaults P, Rago C, St Croix B, Romans KE, Saha S, et al. (2001) Secreted and cell surface genes expressed in benign and malignant colorectal tumors. Cancer research 61: 6996–7001.
4. Welsh JB, Zarrinkar PP, Sapinoso LM, Kern SG, Behling CA, et al. (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America 98: 1176–1181.
5. Welsh JB, Sapinoso LM, Kern SG, Brown DA, Liu T, et al. (2003) Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. Proceedings of the National Academy of Sciences of the United States of America 100: 3410–3415.
6. Basil CF, Zhao YD, Zavaglia K, Jin P, Panelli MC, et al. (2006) Common cancer biomarkers. Cancer research 66: 2953–2961.
7. Veenstra TD, Conrads TP, Hood BL, Avellino AM, Ellenbogen RG, et al. (2005) Biomarkers: mining the biofluid proteome. Molecular & cellular proteomics. MCP 4: 409–418.
8. Hu S, Loo JA, Wong DT (2006) Human body fluid proteome analysis. Proteomics 6: 6326–6353.
9. Nakai K, Kanehisa M (1991) Expert system for predicting protein localization sites in gram-negative bacteria. Proteins-Structure Function and Genetics 11: 95–110.
10. Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14: 897–911.
11. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins-Structure Function and Genetics 43: 246–255.
12. Cai YD, Chou KC (2004) Predicting 22 protein localizations in budding yeast. Biochem Biophys Res Commun 323: 425–428.
13. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, et al. (2007) WoLF PSORT: protein localization predictor. Nucleic acids research 35: W585–587.
14. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. Journal of molecular biology 340: 783–795.
15. Hiller K, Grote A, Scheer M, Munch R, Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic acids research 32: W375–379.
16. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nature protocols 2: 953–971.
17. Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 348: 85–100.
18. Cai Y, He J, Li X, Feng K, Lu L, et al. (2010) Predicting Protein Subcellular Locations Based on Feature Selection Method and Analysis. Protein Pept Lett 17: 464–472.
19. 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.
20. 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.
21. 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. J Theor Biol;doi: 10.1016/j.jtbi.2011.1006.1005.
22. Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. Anal Biochem 370: 1–16.
23. 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.
24. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31: 365–370.
25. Li SJ, Peng M, Li H, Liu BS, Wang C, et al. (2009) Sys-BodyFluid: a systematical database for human body fluid proteome research. Nucleic Acids Res 37: D907–912.
26. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic acids research 33: D433–437.
27. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–416.
28. Chou KC (1989) Graphic rules in steady and non-steady enzyme kinetics. Journal of Biological Chemistry 264: 12074–12079.
29. Zhou GP, Deng MH (1984) An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochemical Journal 222: 169–176.
30. Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Canadian Journal of Chemistry 86: 342–357.
31. Chou KC (1990) Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry 35: 1–24.
32. Chou KC (2010) Graphic rule for drug metabolism systems. Current Drug Metabolism 11: 369–378.
33. Cui J, Liu Q, Puett D, Xu Y (2008) Computational prediction of human proteins that can be secreted into the bloodstream. Bioinformatics 24: 2370–2375.
34. Shin CJ, Wong S, Davis MJ, Ragan MA (2009) Protein-protein interaction as a predictor of subcellular location. BMC systems biology 3: 28.
35. Wang YC, Lan CY, Hsieh WP, Murillo LA, Agabian N, et al. (2010) Global screening of potential Candida albicans biofilm-related transcription factors via network comparison. BMC Bioinformatics 11: 53.
36. Chou KC, Zhang CT (1995) Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology 30: 275–349.
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. 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/).
39. 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.
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 and Peptide Letters 16: 27–31.
41. 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.
42. Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E (2009) A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. Journal of Theoretical Biology 261: 449–458.
43. Gu Q, Ding YS, Zhang TL (2010) Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. Protein & Peptide Letters 17: 559–567.
44. 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.
45. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
46. Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, et al. (2011) AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology 270: 56–62.
47. Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. Journal of Theoretical Biology 271: 10–17.
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. Sennels L, Salek M, Lomas L, Boschetti E, Righetti PG, et al. (2007) Proteomic analysis of human blood serum using peptide library beads. Journal of proteome research 6: 4055–4062.

50. Denny P, Hagen FK, Hardt M, Liao L, Yan W, et al. (2008) The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions. Journal of proteome research 7: 1994–2006.
51. Fang X, Yang L, Wang W, Song T, Lee CS, et al. (2007) Comparison of electrokinetics-based multidimensional separations coupled with electrospray ionization-tandem mass spectrometry for characterization of human salivary proteins. Analytical chemistry 79: 5785–5792.
52. Adachi J, Kumar C, Zhang Y, Olsen JV, Mann M (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. Genome biology 7: R80.
53. Sun W, Li F, Wu S, Wang X, Zheng D, et al. (2005) Human urine proteome analysis by three separation approaches. Proteomics 5: 4994–5001.
54. Castagna A, Cecconi D, Sennels L, Rappsilber J, Guerrier L, et al. (2005) Exploring the hidden human urinary proteome via ligand library beads. Journal of proteome research 4: 1917–1930.
55. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727–1736.
56. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173–1178.
57. Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome biology 5: R63.
58. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968.
59. Gilbert D (2005) Biomolecular interaction network database. Briefings in bioinformatics 6: 194–198.
60. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic acids research 39: D698–704.
61. Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, et al. (2005) CYGD: the Comprehensive Yeast Genome Database. Nucleic acids research 33: D364–368.
62. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic acids research 32: D449–451.
63. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database–2009 update. Nucleic acids research 37: D767–772.
64. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, et al. (2010) MINT, the molecular interaction database: 2009 update. Nucleic acids research 38: D532–539.
65. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, et al. (2010) The IntAct molecular interaction database in 2010. Nucleic acids research 38: D525–531.