frontiers in Genetics

METHODS
published: 23 October 2020
doi: 10.3389/fgene.2020.600454

DPPN-SVM: Computational Identification of Mis-Localized Proteins in Cancers by Integrating Differential Gene Expressions With Dynamic Protein-Protein Interaction Networks

Guang-Ping Li, Pu-Feng Du, Zi-Ang Shen, Hang-Yu Liu and Tao Luo**

College of Intelligence and Computing, Tianjin University, Tianjin, China

OPEN ACCESS

Edited by:
Wen Zhang,
Huazhong Agricultural University, China

Reviewed by:
Ting Wang,
Fred Hutchinson Cancer Research Center, United States
Shiwei Sun,
Chinese Academy of Sciences (CAS), China
Lei Xu,
Shenzhen Polytechnic, China

***Correspondence:**
Pu-Feng Du
pdu@tju.edu.cn
Tao Luo
luo_tao@tju.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 August 2020*
Accepted: *07 October 2020*
Published: *23 October 2020*

Citation:
Li G-P, Du P-F, Shen Z-A, Liu H-Y and Luo T (2020) DPPN-SVM: Computational Identification of Mis-Localized Proteins in Cancers by Integrating Differential Gene Expressions With Dynamic Protein-Protein Interaction Networks. Front. Genet. 11:600454. doi: 10.3389/fgene.2020.600454

Eukaryotic cells contain numerous components, which are known as subcellular compartments or subcellular organelles. Proteins must be sorted to proper subcellular compartments to carry out their molecular functions. Mis-localized proteins are related to various cancers. Identifying mis-localized proteins is important in understanding the pathology of cancers and in developing therapies. However, experimental methods, which are used to determine protein subcellular locations, are always costly and time-consuming. We tried to identify cancer-related mis-localized proteins in three different cancers using computational approaches. By integrating gene expression profiles and dynamic protein-protein interaction networks, we established DPPN-SVM (Dynamic Protein-Protein Network with Support Vector Machine), a predictive model using the SVM classifier with diffusion kernels. With this predictive model, we identified a number of mis-localized proteins. Since we introduced the dynamic protein-protein network, which has never been considered in existing works, our model is capable of identifying more mis-localized proteins than existing studies. As far as we know, this is the first study to incorporate dynamic protein-protein interaction network in identifying mis-localized proteins in cancers.

Keywords: protein subcellular localization, differentially gene expression, protein-protein interactions, mis-localized proteins, diffusion kernel

INTRODUCTION

Eukaryotic cell is the most basic structural and functional unit of eukaryotic living creatures. Every cell contains numerous more basic components named subcellular compartments or subcellular organelles (Reece, 2015). According to the presence or absence of membranes, these subcellular organelles can be divided into two categories, the membrane bounded subcellular compartments and the non-membrane bounded subcellular structures (Perez-Ordonez et al., 2006). The membrane bounded subcellular compartments are those compartments surrounded by a single or double lipid layer membrane, such as mitochondria, nucleus and chloroplasts

(in photosynthetic organisms). The non-membrane bounded subcellular structures, for example, the ribosomes, the cytoskeletons and the centrioles, are those structures without a membrane.

Proteins, which are translated in cytosol or rough ER (Endoplasmic Reticulum), must be transported to proper compartments during or after the translations to perform their biological functions (Mitra et al., 2006; Nyathi et al., 2013; Johnson et al., 2013). This process is known as the protein sorting process (Alberts et al., 2002). The subcellular organelles, where a protein performs its biological functions, are called the subcellular localization of the protein. A protein may have one or more than one subcellular localizations (Cheng et al., 2017). In complex disease conditions, some proteins may be sorted to incorrect subcellular locations, which results in abnormal intracellular behavior (Lee et al., 2008). For example, Zellweger syndrome is a rare congenital disorder characterized by the reduction or absence of functional peroxisomes in the cells of an individual (Brul et al., 1988). A study showed that many diseases such as Swyer syndrome, speech-language disorder, Alzheimer's disease, kidney stones and Diamond-Blackfan anemia were all associated with mis-localized proteins (Hung and Link, 2011). Therefore, tracking alternative subcellular locations in different cellular conditions is important in understanding the pathology of complex diseases, like cancers.

With the help of automatic image processing and understanding technology, the first comprehensive human protein localization map was finally established (Uhlen et al., 2010; Thul et al., 2017). However, the experimental methods used to establish this kind of comprehensive localization map is still costly and time consuming (Horwitz and Johnson, 2017), which makes it difficult to establish this kind of localization map in different cellular conditions, such as disease conditions, drug perturbations and environmental stress conditions. Therefore, computational prediction approaches are still demanded in analyzing altered protein subcellular locations in different conditions.

During the last twenty years, hundreds of works have been done in predicting protein subcellular locations using various types of information at various levels of cellular structure in various species (Chou and Shen, 2006; Briesemeister et al., 2010; Mooney et al., 2011; Zhou et al., 2017; Cheng et al., 2017). For example, many works have been done in predicting protein subcellular locations using protein sequences and sequence related information (Chou and Shen, 2007; Du et al., 2011; Du and Xu, 2013). Most of these works rely on machine learning algorithms (Chou, 2011). Unfortunately, almost all existing studies, which focus on predicting protein subcellular locations, only predict subcellular locations for a given protein in only one condition (Liu and Hu, 2016).

This is because almost all existing studies of this kind utilize only the static information as the input data. For example, most of the existing methods tried to extract informative features from the primary sequence of proteins, while the mutations and the SNPs were not taken into considerations. For another example, some of the existing methods make use of the gene ontology annotations, as well as the functional domain composition of proteins (Zhou et al., 2017). There is still no distinguishable information that can be extracted from the gene ontology annotations or the functional domain compositions for different cellular conditions.

Several existing methods are designed to find the alternative protein subcellular locations in different cellular conditions. PROLocalizer makes use of sequence mutations to detect mis-localized protein in diseases (Laurila and Vihinen, 2009, 2011). Lee et al. integrated protein sequences, PPI (Protein-Protein Interaction) networks, and gene expression profiles to predict mis-localized proteins in glioma (Lee et al., 2008). Liu and Hu improved the Lee's method to predict mis-localized protein in several types of cancers (Liu and Hu, 2016).

In these existing works, the information to distinguish different cellular conditions comes from two sources, one is the mutations and SNPs, while the other is the differential gene expressions. Although the gene mutation and SNP information is useful, it is not easy to utilize them in sequence based features. On the contrary, many gene expression datasets have been deposited in the NCBI GEO (Gene Expression Omnibus) database (Barrett et al., 2013), which have been proved to be useful if they are combined with the protein-protein interaction networks (Ideker and Krogan, 2012). Therefore, combining the gene expression profiles and the PPI network is a feasible way to explore mis-localized proteins in cancers, as well as other kinds of complex diseases.

Although state-of-the-arts methods, which applied gene expression profiles and PPI networks to predict mis-localized proteins in cancers, have achieved success in several specific types of cancers, it should be noted that these methods have two common issues.

First, all state-of-the-arts methods used identical PPI network structures in both the disease and non-disease conditions. This is the result of lacking PPI network data in specific disease conditions. However, if a protein is mis-localized in the disease condition, its interacting proteins must be changed, as the physical distances between the mis-localized protein and the other proteins are changed. Therefore, the topological structure of the PPI network in the disease condition must not be identical to the non-disease condition.

Second, as the topological structure of the PPI network should be changed in the disease condition, the difference of the topological structure of the PPI network should be utilized to predict mis-localized proteins.

In this work, we tried to solve the above two issues by building a model named DPPN-SVM (Dynamic Protein-Protein Network with Support Vector Machine). We made changes to the PPI network in the non-disease condition according to the changes of co-expression scores in disease condition to establish an adjusted PPI network in the disease condition. We applied the ECC (edge clustering coefficient), which has already been applied in predicting essential proteins and protein subcellular locations (Wang et al., 2012; Du and Wang, 2014), to extract the PPI network structure information. By training SVM classifiers with diffusion kernels (Kondor and Lafferty, 2002) on the PPI network, we can predict protein subcellular locations in different cellular conditions. We developed a mis-localization score, which

describes how likely a protein will move to or leave from a specific subcellular location in a specific cellular condition. We hope this work may provide a better way in predicting mis-localized protein in various types of cancers.

MATERIALS AND METHODS

PPI Network Construction

We downloaded our PPI data from the BioGRID database version 3.5.179 (Oughtred et al., 2019). To construct a high quality working dataset, we screened the raw PPI data strictly using the following criteria. (1) Only interactions between two human proteins were kept. (2) The interactions between two identical proteins were discarded, as this kind of interactions does not provide useful information for protein subcellular localizations. (3) Duplicate interaction records were reduced to unique interactions. (4) Only physical interactions were kept. All other types of interactions were removed. This is because the physical interactions implied that the two interactors have a very short physical distance, which contributes to protein subcellular location predictions. To achieve this, we kept only those interaction records with interaction type MI:0915 (physical association) or MI:0407 (direct interaction). After all above filtering procedures, we obtained 341088 interactions involving 23810 proteins.

Subcellular Localization Annotations

We obtained reviewed human protein records from the UniProt database (UniProt Consortium, 2019), which include 20432 proteins. We employed the online ID mapping function of the UniProt database to convert the BioGRID protein IDs of every node in the PPI network to the UniProt database IDs. There are 16319 proteins in our PPI network, which can be mapped uniquely between the UniProt database and the BioGRID database. Although this covers just about 68% nodes in the PPI network, the number of interactions between these mapped proteins is 301366, which covers over 88% of all interactions.

After the mapping procedure, we transferred the GO (Gene Ontology) annotations in cellular component ontology category from the UniProt records to the BioGRID proteins. We chose the following 12 subcellular locations, including Cell cortex(GO:0005938), Cytosol(GO:0005829), Actin cytoskeleton(GO:0015629), Golgi apparatus(GO:0005794), Endoplasmic reticulum(GO:0005783), Nucleolus(GO:0005730), Peroxisome(GO:0005777), Mitochondrion(GO:0005739), Lysosome(GO:0005764), Centrosome(GO:0005813), Nucleus(GO:0005634), and Plasma membrane(GO:0005886). When the GO annotations were transferred from the Uniprot records to the BioGRID proteins, we choose to transfer only those GO terms with experimental evidences. This is achieved by choosing only those terms with evidence code IDA (Inferred from Direct Assay) or HDA (Inferred from High Throughput Direct Assay). We have 6461 BioGRID proteins that were experimentally annotated with at least one of the above 12 subcellular locations.

Among the 6461 annotated BioGRID proteins, there were 4112 proteins with only one subcellular location, 1731 proteins with two locations, 503 proteins with three locations, 98 proteins with four locations, 15 proteins with five locations and 2 proteins with six locations. The average multiplicity degree of the dataset was 1.48. The breakdown of the dataset for different location multiplicity is illustrated in **Figure 1A**.

Virtual Locative Proteins

Since one protein may have more than one subcellular locations, it is necessary to introduce the virtual locative protein concept (Chou and Shen, 2006). In the view of machine learning, computational prediction of multiple subcellular locations for a single protein is a multi-label classification problem. Therefore, it should be converted to a single-label classification problem before it can be dealt with traditional machine-learning algorithms.

Every protein with κ ($\kappa > 1$) subcellular locations was split into κ virtual locative proteins. Each of the κ virtual locative proteins has one and only one of the κ subcellular locations. For example, if a protein p_i has two subcellular locations l_1 and l_2, we split the protein p_i into two different virtual proteins, located at l_1 and l_2, respectively.

The virtual locative proteins inherited the properties of the original real proteins, including all PPI connections and gene expression profiles. Since the virtual locative proteins have different subcellular locations, we assumed that there is no PPI between the virtual locative proteins that are generated from the same real protein.

The original 6461 proteins with experimentally annotated subcellular locations are split into 9562 virtual locative proteins, resulting in a multiplicity degree of 1.48. Therefore, the number of proteins that are mapped between UniProt and BioGRID increased to 19420, which is about 120% of the original. The number of PPI in the network increased to 601693, which is about 200% of the original. **Figure 1B** gives the breakdown of the dataset in the term of virtual locative proteins in different subcellular locations.

Edge Clustering Coefficients

Edge clustering coefficient was originally developed in analyzing social networks (Radicchi et al., 2004). It has been introduced in identifying essential proteins (Wang et al., 2012), as well as in predicting protein subcellular locations (Du and Wang, 2014). Particularly, ECC has been proved to be an indicator of whether two interacting proteins tend to have common subcellular locations (Du and Wang, 2014). For a pair of interacting proteins, which can be noted as the u-th and the v-th proteins, the ECC can be defined as follows:

$$\eta_{u,v} = \frac{z_{u,v}}{\min\left(d_u - 1, d_u - 1\right)}, \quad (1)$$

where $\eta_{u,v}$ is the ECC between the u-th and the v-th proteins, $z_{u,v}$ the number of triangles that involve the edge between the u-th and the v-th proteins, and d_u and d_v the degree of the u-th and the v-th proteins, respectively.

The denominator in Eq (1) represents the possible most number of triangles that may involve the u-th and the v-th

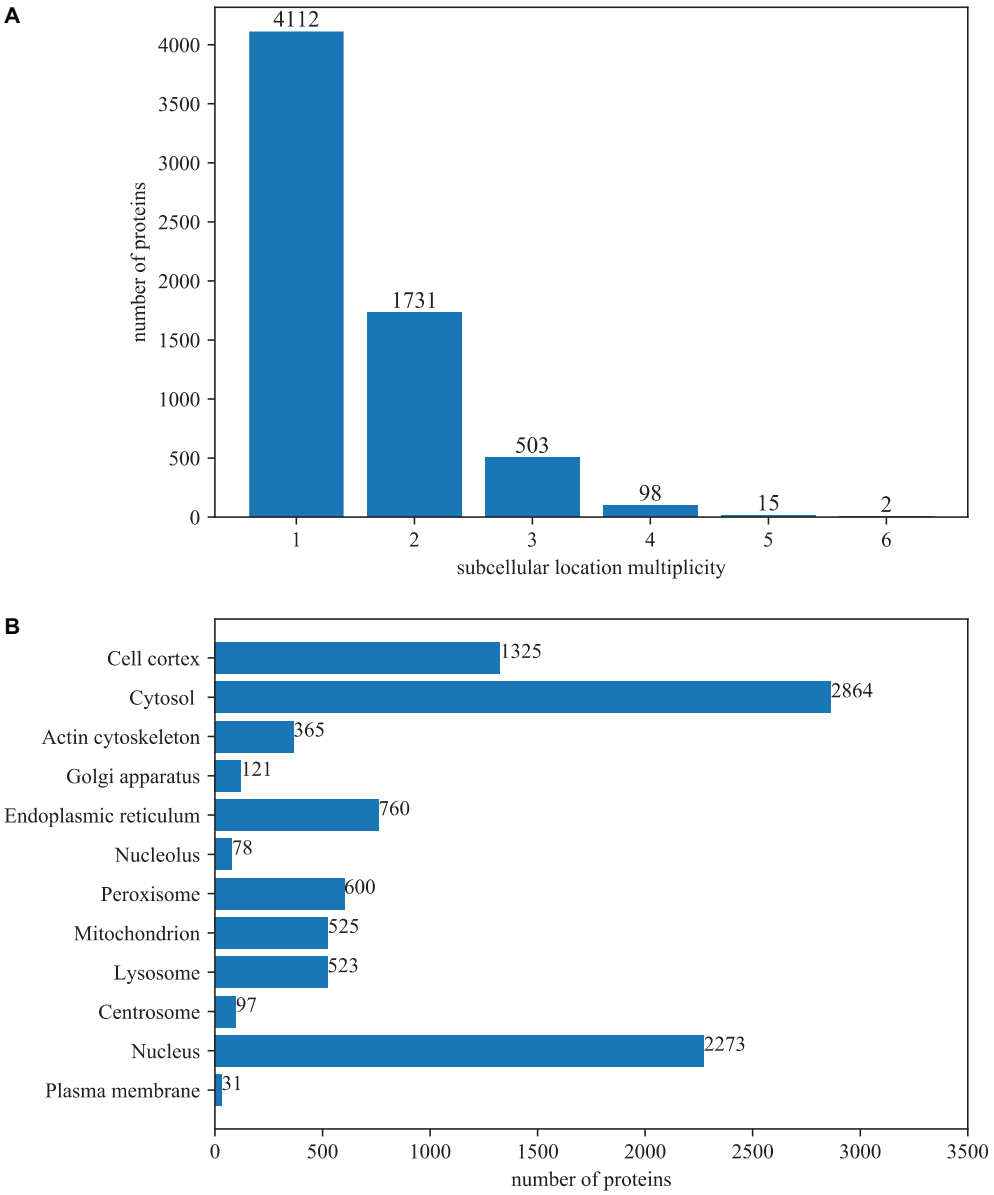

FIGURE 1 | The summary of dataset. **(A)** The number of proteins with different number of subcellular locations. Among the 6461 annotated BioGRID proteins, there were 4112 proteins with only one subcellular location, 1731 proteins with two locations, 503 proteins with three locations, 98 proteins with four locations, 15 proteins with five locations and 2 proteins with six locations. The average multiplicity degree of the dataset was 1.48. **(B)** The number of locative proteins in different subcellular locations. There are 6461 proteins with experimentally annotated subcellular locations in the dataset. Because one protein may have more than one subcellular location, the number of locative proteins is 9562.

proteins. We set $\eta_{u,v} = 0$ in the case that the denominator is degraded to zero.

Diffusion Kernel Matrix

In order to apply machine learning techniques to graph-like structures, diffusion kernel was proposed to capture the long-range relationships between vertices induced by the local structure of a graph (Kondor and Lafferty, 2002). The diffusion kernels provide means to incorporate all neighbors of proteins in the network (Lee et al., 2006).

Let G be a simple graph. Its Laplacian matrix can be defined as:

$$L = D - A, \tag{2}$$

where **A** is the adjacency matrix of the graph, and **D** the degree matrix. The matrix **D** can be defined as:

$$D = \{d_{i,j}\} = \begin{cases} d_i & i = j \\ 0 & Otherwise \end{cases}, \tag{3}$$

where d_i is the degree of the i-th vertex in the graph. The diffusion kernel matrix $\mathbf{K}(\tau)$ is given by:

$$K(\tau) = \exp(-\tau L), \quad (4)$$

where τ is a constant parameter, exp() the matrix exponential function. It can be easily shown that the $\mathbf{K}(\tau)$ is a valid kernel function.

Co-expression Network Construction

Three cancer-related gene expression profile datasets were obtained from the NCBI GEO database. These datasets are from studies on acute myeloid leukemia, breast cancer and hepatitis carcinoma, respectively. The datasets include GSE9476 (myeloid leukemia, 25 cases and 38 controls), GSE27567 (breast cancer, 51 cases and 31 controls) and GSE121248 (hepatitis carcinoma, 70 cases and 37 controls). All gene expression datasets were retrieved using the Affymetrix platforms (Dalma-Weiszhausz et al., 2006). We used the "simpleaffy" package in the Bioconductor to perform quality controls (Wilson and Miller, 2005). For each dataset, the following filtering steps were carried out. (1) The samples with scale factors larger than 3 were removed. (2) The samples with 3′ to 5′ ratios for β-actin less than 3 were kept. (3) The samples with 3′ to 5′ ratios for GAPDH (Glyceraldehyde 3-phosphate dehydrogenase) less than 1.25 were kept. We also checked the RLE (relative log expression) and NUSE (normalized unscaled standard errors) of samples. Samples with significant different RLE or NUSE values to other samples were removed. The case and control samples in each dataset were grouped, respectively. The MAS5 algorithm (Pepper et al., 2007) were applied to generate expression values for every sample. We applied the affymetrix templates and annotation packages in Bioconductor to map the gene expression values to UniProt proteins. In case of a many-to-one mapping, we used the mean value as the final expression value for proteins.

Let $x_{i,u}$ be the u-th protein expression values of the i-th sample, n the number of samples in a group. We define the sample-wise centered expression vector $\mathbf{X}_u$ as follows:

$$X_u = \left[x_{1,u} - a_u \; x_{2,u} - a_u \cdots x_{n,u} - a_u \right]^T, \quad (5)$$

where T is the transpose operator for matrix, and

$$a_u = \frac{1}{n} \sum_{i-1}^{n} x_{i,u}. \quad (6)$$

We now defined the pair-wise PCC (Pearson Correlation Coefficient) between the u-th protein and the v-th protein as the follows:

$$\rho_{u,v} = \frac{X_u^T X_v}{\sqrt{X_u^T X_u}\sqrt{X_v^T X_v}}, \quad (7)$$

where $\rho_{u,v}$ is the PCC between the u-th and the v-th proteins.

The PCC was used to quantify the coherent extent of two proteins in terms of gene expressions. Regardless to whether two proteins have physical interactions, their PCC was calculated as above.

Disease-Related Mis-Localized Protein Identification

Given a specific disease status θ, we term the case sample set as θ_1, while the control sample set as θ_0.

We can compute the PCC for all pairs of proteins as Eq(7) using only the samples in θ_0. The PCC between the u-th and the v-th proteins in non-disease states can be noted as $\rho_{u,v}(\theta_0)$. Similarly, we can compute the ECC for each interaction as Eq(1). The ECC between the u-th and the v-th proteins in non-disease states can be noted as $\eta_{u,v}(\theta_0)$.

Let $\mathbf{A}(\theta_0)$ be the adjacency matrix of the PPI network in non-disease states, which can be defined as follows:

$$A(\theta_0) = \{a_{u,v}(\theta_0)\} = \begin{cases} \rho_{u,v}(\theta_0) \; \eta_{u,v}(\theta_0) & \textit{The u-th and v-th protein are interacting} \\ 0 & \textit{otherwise} \end{cases}. \quad (8)$$

The Laplacian matrix in non-disease state can be defined as:

$$L(\theta_0) = D(\theta_0) - A(\theta_0), \quad (9)$$

where $\mathbf{D}(\theta_0)$ is the degree matrix that is computed using Eq(3).

With $\mathbf{L}(\theta_0)$, we can create the diffusion kernel matrix $\mathbf{K}(\tau,\theta_0)$ using Eq(4). This kernel matrix is used in an SVM model to predict protein subcellular locations in the non-disease state. Since we took the multi-label scenario into the consideration, we employed the libSVM package (Chang and Lin, 2011) to derive the probability that each locative protein localized to each subcellular locations.

Let $p_{u,k}(\theta_0)$ be the probability score that the u-th protein localize to the k-th subcellular location. The libSVM package ensures that

$$\sum_{k=1}^{m} p_{u,k}(\theta_0) = 1, \quad (10)$$

where m is the number of all possible subcellular locations.

Due to the imbalanced dataset, the ranges of $p_{u,k}(\theta_0)$ of different subcellular locations varies a lot. Therefore, we defined the following adjusted probability score, $q_{u,k}(\theta_0)$, which is for the u-th protein and the k-th subcellular location:

$$q_{u,k}(\theta_0) = \frac{\hat{p}_{u,k}(\theta_0)}{\sum_{k=1}^{m} \hat{p}_{u,k}(\theta_0)}, \quad (11)$$

where

$$\hat{p}_{u,k}(\theta_0) = \frac{p_{u,k}(\theta_0) - \min_u p_{u,k}(\theta_0)}{\max_u p_{u,k}(\theta_0) - \min_u p_{u,k}(\theta_0)}. \quad (12)$$

With all above definitions, the u-th protein localize to the k-th subcellular location if the following condition is satisfied:

$$q_{u,k}(\theta_0) \geq \max_k q_{u,k}(\theta_0) - \alpha \left(\max_k q_{u,k}(\theta_0) - \min_k q_{u,k}(\theta_0) \right), \quad (13)$$

where α is a real number parameter between 0 and 1. The subcellular locations, which are predicted for the u-th protein using Eq(13), can be denoted as a set $\mathbf{S}_u(\theta_0)$.

For the disease state, all above computation can be performed on θ_1. However, to amplify the differences between disease and non-disease status, we altered the topology of the PPI network before all computations in disease status. This is different to all existing works in predicting mis-localized proteins in diseases.

For the u-th protein and the v-th protein, we first compute the PCC in θ_1, which can be noted as $\rho_{u,v}(\theta_1)$. We define the disease status difference of PCC as follows:

$$h_{u,v} = \rho_{u,v}(\theta_1) - \rho_{u,v}(\theta_0). \tag{14}$$

We define two threshold parameters as follows:

$$t_+ = h + 3\sigma, \text{ and} \tag{15}$$

$$t_- = h - 3\sigma, \tag{16}$$

where h is the average value of all $h_{u,v}$, and σ the standard deviation of all $h_{u,v}$.

If the u-th protein and the v-th protein are two interacting proteins in non-disease status, the interaction would be removed, if $h_{u,v} < t_-$ is satisfied. Similarly, if the u-th protein and the v-th protein are two non-interacting proteins in non-disease status, the interaction between them should be established, if $h_{u,v} > t_+$ is satisfied.

After altering the topology of the PPI network as above, we compute the $\mathbf{S}_u(\theta_1)$ according to the Eq(8) to Eq(13) using the updated PPI network and gene expression samples in θ_1. It should be noted that the $\eta_{u,v}(\theta_1)$ may be different to $\eta_{u,v}(\theta_0)$, as the topology of the PPI network is altered in the disease state.

By comparing the $\mathbf{S}_u(\theta_1)$ and $\mathbf{S}_u(\theta_0)$, we can identify whether the subcellular locations of the u-th protein were altered in the disease state. However, this method cannot quantify how likely a protein would be mis-localized in the disease state. Therefore, we developed the following method to quantify the mis-localized proteins, which we termed as the mis-localization scores.

For each disease, we compute the differences of adjusted probability scores between the disease and non-disease states. The mis-localization score of the u-th protein in the k-th subcellular location of disease θ can be defined as follows:

$$\varphi_{u,k}(\theta) = \frac{q_{u,k}(\theta_1) - q_{u,k}(\theta_0)}{q_{u,k}(\theta_0)}. \tag{17}$$

The $\varphi_{u,k}(\theta)$ indicates the extent that the u-th protein would localize to or move from the k-th subcellular location. For each protein, we define the following two boundaries:

$$\sup[\varphi_u(\theta)] = \max_k \varphi_{u,k}(\theta), \text{ and} \tag{18}$$

$$\inf[\varphi_u(\theta)] = \min_k \varphi_{u,k}(\theta) \tag{19}$$

We sorted the proteins according to the $\sup[\varphi_u(\theta)]$ and $\inf[\varphi_u(\theta)]$ in descending and ascending orders, respectively. The top-ranked proteins within a fixed proportion of the entire list are considered as mis-localized proteins. The proportion is fixed as 0.1% in this work.

Performance Evaluation Methods

In this study, we used 10-fold cross-validation to evaluate the prediction performance of our method in the non-disease state. Four statistics, including aiming (AIM), coverage (CVR), multi-label accuracy (mlACC), absolute-true rate (ATR) were applied to measure the prediction performances (Jiao and Du, 2016). These statistics are defined as follows:

$$AIM = \frac{1}{b}\sum_{u=1}^{b}\left|\frac{S_u(\theta_0) \bigcap S_u}{|S_u(\theta_0)|}\right|, \tag{20}$$

$$CVR = \frac{1}{b}\sum_{u=1}^{b}\left|\frac{S_u(\theta_0) \bigcap S_u}{|S_u|}\right|, \tag{21}$$

$$\text{mlACC} = \frac{1}{\text{b}}\sum_{\text{u}=1}^{\text{b}}\left|\frac{\text{S}_\text{u}(\theta_0) \bigcap \text{S}_\text{u}}{\text{S}_\text{u}(\theta_0) \bigcup \text{S}_\text{u}}\right|, \text{and} \tag{22}$$

$$ATR = \frac{1}{b}\sum_{u=1}^{b}\delta[S_u(\theta_0), S_u], \tag{23}$$

where $\mathbf{S}_u(\theta_0)$ is the set of predicted protein subcellular locations of the u-th protein in the non-disease state, $\mathbf{S}_u$ the set of experimental protein subcellular locations, b the number of proteins, $|.|$ the cardinal operator in set theory, and

$$\delta[S_u(\theta_0), S_u] = \begin{cases} 1 & S_u(\theta_0) = S_u \\ 0 & otherwise \end{cases}. \tag{24}$$

Since we have introduced the virtual locative proteins in our work, we also applied single-label performance measures. Five statistics, including sensitivity (Sen), specificity (Spe), virtual-locative accuracy (vlAcc), positive-predictive value (PPV) and Matthew's Correlation Coefficients (MCC) are applied in our work. These statistics can be defined as follows:

$$Sen = \frac{TP}{TP + FN}, \tag{25}$$

$$Spe = \frac{TN}{TN + FP}, \tag{26}$$

$$PPV = \frac{TP}{TP + FP}, \tag{27}$$

$$\text{vlAcc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \text{and} \tag{28}$$

$$MCC = \frac{TPTN - FPFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{29}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives, and false negatives in the cross-validation, respectively.

Parameter Calibrations

We used a grid search strategy to find the parameter combination of τ and α that optimize the 10-fold cross validation performances in the non-disease state. The parameter τ in computing the diffusion kernel was searched from 0.1 to 2.0 with step 0.1. The parameter α in Eq(13) was searched from 0.1 to 0.3 with a step of 0.1. **Supplementary Figure 1** showed the global MCC score under different parameters. We chose the parameter values $\tau = 1.1$ and $\alpha = 0.3$ in our works.

RESULTS AND DISCUSSION

Prediction Performance Analysis in the Non-disease State

We used 10-fold cross-validation to evaluate the prediction performances in non-disease state. It should be noted that our method is designed to find out the alteration of protein subcellular locations, rather than the exact subcellular locations in non-disease state. Therefore, we choose to compare our method to Liu and Hu's method (Liu and Hu, 2016). Since we applied virtual locative protein concept in our work, while Liu and Hu employed the top-k accuracy performance measure, it is difficult to perform an exact apple-to-apple orange-to-orange comparison. However, we managed to compare the global sensitivity of our work to the top-1 accuracy of Liu and Hu's work. As our performance value was obtained by using 10-fold cross-validation, this gives some advantage to Liu and Hu's work. Our global sensitivity is 0.556, while the top-1 accuracy of Liu and Hu's work is 0.364. Although both values are not high enough in the general protein subcellular location predictions, we still achieved a comparable or little higher performance. Other global performance measure in terms of virtual locative proteins are a specificity of 0.899, a PPV of 0.437, an accuracy of 0.857 and an MCC of 0.412.

To make further performance assessment, we choose to compare the multi-label performance of our method to the Hum-mPLoc 3.0, which was developed by using gene ontology information. Since our method does not rely on the gene ontology annotations, which has been proved to have superior performances in predicting protein subcellular locations, it should be noted that the Hum-mPLoc 3.0 (Zhou et al., 2017) has intrinsic performance advantages.

Since Hum-mPLoc 3.0 does not use identical subcellular locations annotations as our method, we choose to compare the overlapped locations. To achieve a fair enough comparison, we compose a testing dataset of 3842 proteins. All these proteins are with at least one overlapped subcellular location. This testing dataset was fed into the Hum-mPLoc 3.0 and our method in non-disease state. The overall multi-label performances were compared in **Table 1**. It can be seen that our method has better performance in terms of aiming, coverage, accuracy and absolute true rate. This is an expectable result, as our method incorporates PPI information and gene expression profiles.

TABLE 1 | Performance comparison in non-disease state.

Measures[a]	Our method	Hum-mPLoc 3.0
AIM	72.00%	68.10%
CVR	69.50%	65.10%
mIACC	68.60%	65.00%
ATR	64.30%	61.80%

[a] *All performance measures are defined in Eq (20), (21), (22), and (23).*

Discovery of Potentially Mis-Localized Proteins in Cancers

We applied our method on three different type of cancers, including leukemia, breast cancer and hepatitis carcinoma. **Table 2** gives a list of representative mis-localized proteins in these cancer cells. For each disease, we listed the top six (0.1% of the entire list) proteins, which are most likely to mis-localize to an abnormal location, and the top six proteins, which are most likely to mis-localize from their normal locations. The corresponding location, the mis-localization score and the score rank can also be found in **Table 2**. In addition, we listed some highly ranked proteins that has been reported to be related to cancers by other literatures.

A comprehensive list of all proteins with the mis-localization scores can be found in supplementary data. In supplementary data, **Supplementary Tables 1–3** contain the comprehensive lists of the localization scores under different state and mis-localization scores of three diseases with all locations, one table per disease. **Supplementary Tables 4–15** are comprehensive lists of sorted mis-localization scores for hepatitis in different locations, one table per location. **Supplementary Tables 16–27** are comprehensive lists of sorted mis-localization scores for leukemia in different locations. **Supplementary Tables 28–39** are comprehensive lists of sorted mis-localization scores for breast cancer in different locations. **Supplementary Tables 40–42** are comprehensive lists of sorted maximum mis-localization scores, one table per disease. **Supplementary Tables 43–45** are comprehensive lists of sorted minimum mis-localization scores, one table per disease.

Leukemia

In acute myeloid leukemia, we used 25 cases and 38 controls. Our prediction showed that protein SETBP1 mis-localized to ER in cancer cells, as its localization score in ER increased from 0.083 to 0.633 with a mis-localization score +658.04%, while its localization score in nucleus dropped from 0.226 to 0.036 with the mis-localization score −83.94%. A recent study have suggested a direct involvement of SETBP1 in leukemia development (Oakley et al., 2012). We predicted that EI24 mis-localized from ER in cancer cells, as its localization score drops from 0.94 to 0.468 with a mis-localization score −50.22%, while Zhao et al. (2005) found that EI24/PIG8 was an ER-localized Bcl2-binding protein, which was highly mutated in aggressive breast cancers.

Breast Cancer

For breast cancer, we used 51 cases and 31 controls. We made a prediction that the protein B7H1 mis-localized from plasma

TABLE 2 | Representative prediction of mis-localized proteins.

Disorder	Uniprot ID	Mis-localizations[a]	Rank[b]
Leukemia	MAGA3_HUMAN	+Cell cortex (+Inf)	1
	F217B_HUMAN	+Peroxisome (+Inf)	2
	EI24_HUMAN [41]	+Mitochondrion (+3349.02%)	3
	ROP1A_HUMAN	+Peroxisome(+3086.82%)	4
	THYN1_HUMAN	+Nucleus (+2461.70%)	5
	CLGN_HUMAN	+Nucleus (+2425.50%)	6
	SETBP_HUMAN [40]	+Endoplasmic reticulum (+658.04%)	30
	TF2L1_HUMAN	−Nucleus (−99.57%)	1
	UPP1_HUMAN	−Cell cortex (−97.49%)	2
	AL1A1_HUMAN	−Peroxisome (−96.63%)	3
	ABCA1_HUMAN	−Lysosome (−95.71%)	4
	PARP4_HUMAN	−Lysosome (−94.87%)	5
	AL7A1_HUMAN	−Peroxisome (−93.09%)	6
	SETBP_HUMAN [40]	−Nucleus (−83.94%)	28
	EI24_HUMAN [41]	−Endoplasmic reticulum(−50.22%)	348
Breast cancer	TM258_HUMAN	+Peroxisome (+Inf)	1
	KCNKI_HUMAN	+Mitochondrion (+Inf)	2
	MARC2_HUMAN	+Cell cortex (+Inf)	3
	HEBP2_HUMAN	+Lysosome (+Inf)	4
	SIT1_HUMAN	+Mitochondrion (+13310.16%)	5
	PIM3_HUMAN	+Lysosome (+9723.01%)	6
	PD1L1_HUMAN [42]	+Nucleolus (+290.65%)	242
	INGR2_HUMAN [43]	+Mitochondrion (+184.50%)	437
	VGFR3_HUMAN [42]	+Nucleolus (+125.16%)	755
	ANO4_HUMAN	−Nucleus (−98.91%)	1
	ABCA1_HUMAN	−Lysosome (−98.78%)	2
	NDUB7_HUMAN	−Mitochondrion (−98.35%)	3
	TM127_HUMAN	−Plasma membrane (−98.25%)	4
	RUBIC_HUMAN	−Endoplasmic reticulum (−96.45%)	5
	TRIM4_HUMAN	−Nucleus (−96.45%)	6
	INGR2_HUMAN [43]	−Plasma membrane (−63.74%)	595
Hepatitis carcinoma	TBCA_HUMAN	+Cell cortex (+Inf)	1
	F217B_HUMAN	+Peroxisome (+Inf)	2
	HKDC1_HUMAN	+Nucleus (+65006.48%)	3
	SYAC_HUMAN	+Peroxisome (+10652.77%)	4
	RFWD3_HUMAN	+Lysosome (+10599.05%)	5
	ABCA1_HUMAN [10]	+Lysosome (+8115.45%)	6
	S10AB_HUMAN [44]	+Peroxisome (+6868.17%)	12
	FOXP1_HUMAN [10]	+Peroxisome (+612.39%)	478
	RM14_HUMAN	−Cell cortex (−99.99%)	1
	RM47_HUMAN	−Cell cortex (−99.99%)	2
	RT30_HUMAN	−Cell cortex (−99.99%)	3
	DUS11_HUMAN	−Cell cortex (−99.99%)	4
	CLGN_HUMAN	−Cell cortex (−99.99%)	5
	RM01_HUMAN	−Cell cortex (−99.99%)	6
	ABCA1_HUMAN [10]	−Cell cortex (−99.28%)	249

[a] The mis-localization score is marked after the altered location. The "+" prefix indicates this is a new subcellular location in disease state. The "−" prefix indicates this non-disease subcellular location is lost in the disease state. The "Inf" indicates a positive infinity value, which is produced by the zero original localization probability. [b] The score ranks are sorted using the boundary values in Eq(18) and Eq(19). The mis-localization scores with value of −100% does not participate in the ranking, as it does not necessarily indicate a completely loss of a subcellular location, but just a bias of available data.

membrane and to nucleus, as its localization score in plasma membrane dropped from 0.243 to 0.105 with a mis-localization score −56.76%, while its localization score in nucleolus increased from 0.023 to 0.092 with the mis-localization score +290.65%. This consists with the record in literature (Wang and Li, 2014). Our method also reported that the protein VEGFR3 mis-localized from plasma membrane and to cell nucleus, as its localization score in cell nucleus increased from 0.047 to 0.106 (with the

mis-localization score +125.16%). This also consists with the record in literature (Wang and Li, 2014). The protein IFNgR2 was annotated with location ER and Golgi apparatus in the Uniprot database. We predicted that IFNgR2 mis-localized from plasma membrane to mitochondria in cancer cells, as its localization score in plasma membrane drop from 0.214 to 0.078, with a mis-localization score −63.74%, while the localization score in mitochondria increased from 0.136 to 0.388 (with a mis-localization score +184.5%). It was reported that the IFNgR2 molecules can be mainly detected in mitochondria in cancer cells (Ngo et al., 2012).

Hepatitis Carcinoma

For hepatitis carcinoma, we used 70 cases and 37 controls. The protein S100A11 was reported to have very weak nuclear expression in adenocarcinomas (Rehman et al., 2004), while our method reported that it mis-localized to peroxisome, as its localization score in peroxisome increased from 0.014 to 1.0 (with a mis-localization score +6868.17%). We predicted that FOXP mis-localized to peroxisome, as its localization score in peroxisome increased from 0.003 to 0.019 with mis-localization score 612.39%. It has been reported that FOXP would lose its nuclear localization in cancers (Hung and Link, 2011). ABCA1 was reported to mis-localize from plasma membrane to lysosome in cancers (Hung and Link, 2011). Our method reported the same result, as the localization score in plasma membrane dropped from 0.162 to 0.004 with mis-localization score −99.28%), and lysosome from 0.012 to 0.972 (with a mis-localization score +8115.45%).

Potential Results Validation

Using our method, we identified some proteins that may mis-localize from or to a specific location. Some of them have been verified by existing studies. But most of the predicted proteins have not been verified. Due to our limited resources, we cannot perform experimental validations. This may be considered as a future work. It should also be noted that, there is still no database for mis-localized proteins. The information regarding the mis-localized proteins is still scattered in many literatures. Establishing such kind of database is a valuable yet impacting work, which is also in our consideration as a future work in this research topic. Since mis-localized proteins are of great significance on revealing the mechanism of diseases, we believe that it is valuable to establish a database to summarize and store relevant discoveries in future.

CONCLUSION

Computational prediction of proteins subcellular locations has been studied for over twenty years. However, computationally detecting disease-related mis-localized proteins was rarely discussed. By integrating gene expression profiles and protein-protein interaction networks, we developed a computational approach, DPPN-SVM, to detect mis-localized proteins in various cancers. The results indicated that our method can successfully identify cancer-related or mis-localized proteins that has been reported in various literatures. Comparing to existing studies, our method not only provide a comparable or better prediction performance in non-disease state, but also further amplify the differentially expressed gene information by introducing the dynamic PPI network and the SVM classifiers with diffusion kernels. The prediction results of our method provide candidate proteins as spatial cancer markers, while the method of our work gives a new way to explore the spatial distribution of proteins within a cell.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/brown-2/mis_localization.

AUTHOR CONTRIBUTIONS

G-PL collected data, process the data, implement the algorithm, performed most of the experiments, and analyzed the results. P-FD designed and directed the study, proposed the algorithm, analyzed the results, and wrote the manuscript. Z-AS and H-YL performed part of the experiments. TL analyzed the results, and participated in writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China [NSFC 61872268]; National Key R&D Program of China [2018YFC0910405]; and Open Project Funding of CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences [CASNDST201705].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.600454/full#supplementary-material

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Intracellular Compartments and Protein Sorting*. Available online at: http://www.ncbi.nlm.nih.gov/books/NBK21053/ (accessed October 03, 2013).

Barrett, T., Stephen, E. W., Pierre, L., Carlos, E., Irene, F. K., and Maxim, T. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). YLoc—an interpretable web server for predicting subcellular localization.

Nucleic Acids Res. 38(Suppl._2), W497–W502. doi: 10.1093/nar/gkq477

Brul, S., Westerveld, A., Strijland, A., Wanders, R. J., Schram, A. W., Heymans, H. S., et al. (1988). Genetic heterogeneity in the cerebrohepatorenal (Zellweger) syndrome and other inherited disorders with a generalized impairment of peroxisomal functions. A study using complementation analysis. *J. Clin. Invest.* 81, 1710–1715.

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Cheng, X., Zhao, S. G., Lin, W. Z., Xiao, X., and Chou, K. C. (2017). pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics* 33, 3524–3531. doi: 10.1093/bioinformatics/btx476

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024

Chou, K. C., and Shen, H. B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.* 5, 1888–1897. doi: 10.1021/pr060167c

Chou, K. C., and Shen, H. B. (2007). Recent progress in protein subcellular location prediction. *Anal. Biochem.* 370, 1–16. doi: 10.1016/j.ab.2007.07.006

Dalma-Weiszhausz, D., Warrington, J., Tanimoto, E. Y., and Miyada, C. G. (2006). The affymetrix GeneChip platform: an overview. *Meth. Enzymol.* 410, 3–28. doi: 10.1016/S0076-6879(06)10001-4

Du, P., Li, T., and Wang, X. (2011). Recent progress in predicting protein sub-subcellular locations. *Expert Rev. Proteom.* 8, 391–404. doi: 10.1586/EPR.11.20

Du, P., and Wang, L. (2014). Predicting human protein subcellular locations by the ensemble of multiple predictors via protein-protein interaction network with edge clustering coefficients. *PLoS One* 9:e86879. doi: 10.1371/journal.pone.0086879

Du, P., and Xu, C. (2013). Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev. Proteomics* 10, 227–237. doi: 10.1586/epr.13.16

Horwitz, R., and Johnson, G. T. (2017). Whole cell maps chart a course for 21st-century cell biology. *Science* 356, 806–807.

Hung, M. C., and Link, W. (2011). Protein localization in disease and therapy. *J. Cell Sci.* 124, 3381–3392. doi: 10.1242/jcs.089110

Ideker, T., and Krogan, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* 8:565. doi: 10.1038/msb.2011.99

Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 4, 320–330. doi: 10.1007/s40484-016-0081-2

Johnson, N., Powis, K., and High, S. (2013). Post-translational translocation into the endoplasmic reticulum. *Biochim. Biophys. Acta Mol. Cell Res.* 1833, 2403–2409. doi: 10.1016/j.bbamcr.2012.12.008

Kondor, R. I., and Lafferty, J. D. (2002). "Diffusion kernels on graphs and other discrete input spaces," in *Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, CA, 315–322.

Laurila, K., and Vihinen, M. (2009). Prediction of disease-related mutations affecting protein localization. *BMC Genomics* 10:122. doi: 10.1186/1471-2164-10-122

Laurila, K., and Vihinen, M. (2011). PROlocalizer: integrated web service for protein subcellular localization prediction. *Amino Acids* 40, 975–980. doi: 10.1007/s00726-010-0724-y

Lee, H., Tu, Z., Deng, M., Sun, F., and Chen, T. (2006). Diffusion kernel-based logistic regression models for protein function prediction. *OMICS J. Integr. Biol.* 10, 40–55. doi: 10.1089/omi.2006.10.40

Lee, K., Han-Yu, C., Andreas, B., Min-Kyung, S., Won-Ki, H., Bonghee, L., et al. (2008). Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res.* 36:e136. doi: 10.1093/nar/gkn619

Liu, Z., and Hu, J. (2016). Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction. *Methods* 93, 119–127. doi: 10.1016/j.ymeth.2015.09.022

Mitra, K., Frank, J., and Driessen, A. (2006). Co- and post-translational translocation through the protein-conducting channel: analogous mechanisms at work? *Nat. Struct. Mol. Biol.* 13, 957–964. doi: 10.1038/nsmb1166

Mooney, C., Wang, Y. H., and Pollastri, G. (2011). SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics* 27, 2812–2819. doi: 10.1093/bioinformatics/btr494

Ngo, J., Joseph, C., Mengzhi, W., Alexander, A., Hajime, M., and Kuen-Young, J. (2012). Interferon gamma receptor 2 (IFNgR2) has a ligand (IFNg)-independent activity as a Bax inhibitor in cancer cells. *Cancer Res.* 72:2013. doi: 10.1158/1538-7445.AM2012-2013

Nyathi, Y., Wilkinson, B. M., and Pool, M. R. (2013). Co-translational targeting and translocation of proteins to the endoplasmic reticulum. *Biochim. Biophys. Acta* 1833, 2392–2402. doi: 10.1016/j.bbamcr.2013.02.021

Oakley, K., Yufen, H., Bandana, A. V., Su, C., Ravi, B., and Kristbjorn, O. G. (2012). Setbp1 promotes the self-renewal of murine myeloid progenitors via activation of Hoxa9 and Hoxa10. *Blood J. Am. Soc. Hematol.* 119, 6099–6108.

Oughtred, R., Chris, S., Bobby-Joe, B., Jennifer, R., Lorrie, B., Christie, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47, D529–D541. doi: 10.1093/nar/gky1079

Pepper, S. D., Saunders, E. K., Edwards, L. E., Wilson, C. L., and Miller, C. J. (2007). The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics* 8:273. doi: 10.1186/1471-2105-8-273

Perez-Ordonez, B., Beauchemin, M., and Jordan, R. C. K. (2006). Molecular biology of squamous cell carcinoma of the head and neck. *J. Clin. Pathol.* 59, 445–453.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2658–2663.

Reece, J. B. (2015). *Campbell Biology*. London: Pearson.

Rehman, I., Abdel-Rahmene, A., Simon, S. C., Jean, C. D., James, W. F. C., and Natasha, W. (2004). Dysregulated expression of S100A11 (calgizzarin) in prostate cancer and precursor lesions. *Hum. Pathol.* 35, 1385–1391. doi: 10.1016/j.humpath.2004.07.015

Thul, P. J., Lovisa, A., Mikaela, W., Diana, M., Aikaterini, G., and Hammou, A. B. (2017). A subcellular map of the human proteome. *Science* 356:eaal3321. doi: 10.1126/science.aal3321

Uhlen, M., Oksvold, P., Linn, F., Emma, L., Kalle, J., and Mattias, F. (2010). Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* 28, 1248–1250.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Wang, J., Li, M., Wang, H., and Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE ACM Trans. Comput. Biol. Bioinform.* 9, 1070–1080.

Wang, X., and Li, S. (2014). Protein mislocalization: mechanisms, functions and clinical applications in cancer. *Biochim. Biophys. Acta Rev. Cancer* 1846, 13–25. doi: 10.1016/j.bbcan.2014.03.006

Wilson, C. L., and Miller, C. J. (2005). Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. *Bioinformatics* 21, 3683–3685. doi: 10.1093/bioinformatics/bti605

Zhao, X., Robert, E. A., Shannon, L. D., Sarah, J. A., Brian, F., and Cyrus, T. (2005). Apoptosis factor EI24/PIG8 is a novel endoplasmic reticulum–localized Bcl-2–binding protein which is associated with suppression of breast cancer invasiveness. *Cancer Res.* 65, 2125–2129. doi: 10.1158/0008-5472.CAN-04-3377

Zhou, H., Yang, Y., and Shen, H. B. (2017). Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics* 33, 843–853.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.