RESEARCH ARTICLE

Classifying Ten Types of Major Cancers Based on Reverse Phase Protein Array Profiles

Pei-Wei Zhang[2], Lei Chen[4], Tao Huang[2]*, Ning Zhang[3]*, Xiang-Yin Kong[2]*, Yu-Dong Cai[1]*

1 College of Life Science, Shanghai University, Shanghai, P.R. China, **2** The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P.R. China, **3** Department of Biomedical Engineering, Tianjin Key Lab of BME Measurement, Tianjin University, Tianjin, P.R. China, **4** College of Information Engineering, Shanghai Maritime University, Shanghai, P.R. China

* tohuangtao@126.com (TH); zhni@tju.edu.cn (NZ); xykong@sibs.ac.cn (XYK); cai_yud@126.com (YDC)

Abstract

Gathering vast data sets of cancer genomes requires more efficient and autonomous procedures to classify cancer types and to discover a few essential genes to distinguish different cancers. Because protein expression is more stable than gene expression, we chose reverse phase protein array (RPPA) data, a powerful and robust antibody-based high-throughput approach for targeted proteomics, to perform our research. In this study, we proposed a computational framework to classify the patient samples into ten major cancer types based on the RPPA data using the SMO (Sequential minimal optimization) method. A careful feature selection procedure was employed to select 23 important proteins from the total of 187 proteins by mRMR (minimum Redundancy Maximum Relevance Feature Selection) and IFS (Incremental Feature Selection) on the training set. By using the 23 proteins, we successfully classified the ten cancer types with an MCC (Matthews Correlation Coefficient) of 0.904 on the training set, evaluated by 10-fold cross-validation, and an MCC of 0.936 on an independent test set. Further analysis of these 23 proteins was performed. Most of these proteins can present the hallmarks of cancer; Chk2, for example, plays an important role in the proliferation of cancer cells. Our analysis of these 23 proteins lends credence to the importance of these genes as indicators of cancer classification. We also believe our methods and findings may shed light on the discoveries of specific biomarkers of different types of cancers.

Citation: Zhang P-W, Chen L, Huang T, Zhang N, Kong X-Y, Cai Y-D (2015) Classifying Ten Types of Major Cancers Based on Reverse Phase Protein Array Profiles. PLoS ONE 10(3): e0123147. doi:10.1371/journal.pone.0123147

Academic Editor: Lukasz Kurgan, University of Alberta, CANADA

Received: October 20, 2014

Accepted: February 24, 2015

Published: March 30, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants from National Basic Research Program of China (2011CB510102, 2011CB510101), and National Natural Science Foundation of China (31371335, 81171342, 81201148, 61401302), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), the grant of "The First-class Discipline of Universities in Shanghai", Tianjin Research Program of Application Foundation and Advanced Technology(14JCQNJC09500), the National Research Foundation for the Doctoral

Introduction

Identifying cancer-specific genes involved in tumorigenesis and cancer progression is one of the major ways to understand the pathophysiologic mechanisms of cancers and to find therapeutic drug targets. Many efforts have been made to identify cancer biomarkers by using gene expression profiles [1]. However, the robustness of microarray-derived biomarkers is very poor [2]; this is in part because the robustness can be easily influenced in gene expression levels by small environmental changes. Without the evaluation of protein expression levels, there would

Program of Higher Education of China (20130032120070, 20120032120073) and the Independent Innovation Foundation of Tianjin University (60302064, 60302069). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

be no way to illustrate causes of tumor proliferation and differentiation. Therefore, better understanding of the translational states of these genomes will bring us a step closer to finding potential drug targets and to illustrating off-target effects in cancer medicine.

Reverse phase protein array (RPPA) is a powerful and robust antibody-based high-throughput approach for targeted proteomics that allows us to quantitatively assess target protein expression in large sample sets [3]. In this process, sample analytes are immobilized in the solid phase, and analyte-specific antibodies are used in the solution phase. Through using secondary tagging and signal amplification to detect bound antibodies, proteins may be measured. Compared with conventional protein quantify methods, such as western blotting or ELISA, the advantages of RPPA include: large-scale quantification of the protein, high sensitivity, and small sample volume requirements [4]. While mass spectrometry, usually used to quantify the numbers of phosphorylation sites or phosphopeptides, requires further protein digestion, peptide fractionation and phosphopeptide enrichment after protein extraction, RPPA can directly quantify the extracted protein [5]. The application of RPPA has been extensively validated for both cell lines and patient samples [6], and it illustrates mechanistic insights behind diseases.

Currently, cancer types are classified by anatomical positions where they are found, such as lung cancer, breast cancer, etc. Whether these names could present their proteomic feature has not been determined until now. Although there have been some methods to find biomarker signatures for specific cancer types, there is still little research being done that considers different types of cancer as a whole in order to identify their similar or distinct proteomic expression patterns and classification features.

In this study, we proposed a computational workflow to successfully use 23 proteins to classify patient samples into ten main cancer types. First, we randomly divided the 3467 samples from ten types of cancers into a training set with 2775 samples and an independent test set with 692 samples. The proportions of each cancer type were similar in the training set and the independent test set. Then, with the training set, all features for distinguishing groups were ranked by the mRMR (minimum Redundancy Maximum Relevance Feature Selection) criteria. With 10-fold cross-validation on the training set, the SMO (Sequential minimal optimization) and the IFS (Incremental Feature Selection) [7] methods were used to choose an optimal feature set. A total of 23 proteins were selected from the training set. Their MCC (Matthews Correlation Coefficient) for the training set was 0.904 evaluated by 10-fold cross validation and their MCC on the independent test set was 0.936. Our methods could provide clinicians with knowledge of key distinct biochemical features of cancer types and could shed some new light on the discoveries of specific biomarkers of different types of cancers.

Materials and Methods

Datasets

The RPPA data were downloaded from TCPA (The Cancer Proteome Atlas) database [8] (http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/download.html under Pan-Cancer 11 RBN), which contained proteomic expression of 3467 cancer patients in 11 cancer types (**Table 1**). Because COAD (Colon adenocarcinoma) and READ (Rectum adenocarcinoma) share similar pathologies and were analyzed together in the TCGA (The Cancer Genome Atlas) colon and rectal cancer study [9], we combined the COAD and READ samples together as 'Colon adenocarcinoma and Rectum adenocarcinoma' samples. Therefore, ten cancer types were analyzed in following steps.

Because we did not have a different cohort to do multi-center validation, we randomly divided the 3467 samples into a training set with 2775 samples and an independent test set with 692 samples. The ratio of training samples over test samples was approximately 4:1 and we

Table 1. The ten types of cancers and their sample sizes.

Cancer Type	Cancer Abbreviation	Cancer Name	Sample size	Number of training samples	Number of test samples
1	BLCA	Bladder Urothelial Carcinoma	127	102	25
2	BRCA	Breast invasive carcinoma	747	598	149
3	COAD/READ	Colon adenocarcinoma and Rectum adenocarcinoma	464	371	93
4	GBM	Glioblastoma multiforme	215	172	43
5	HNSC	Head and Neck squamous cell carcinoma	212	170	42
6	KIRC	Kidney renal clear cell carcinoma	454	363	91
7	LUAD	Lung adenocarcinoma	237	190	47
8	LUSC	Lung squamous cell carcinoma	195	156	39
9	OV	Ovarian serous cystadenocarcinoma	412	330	82
10	UCEC	Uterine Corpus Endometrioid Carcinoma	404	323	81
Total			3467	2775	692

doi:10.1371/journal.pone.0123147.t001

kept the proportion of each cancer type roughly the same in the training set and the independent test set. The description of the ten cancer types and their sample sizes in are given in Table 1. The training and test data sets are provided in S1 File.

Each sample contained 187 proteins whose expression levels were measured with reverse phase protein array (RPPA). RPPA is a protein array that allows measurement of protein expression levels in a large number of samples simultaneously in a quantitative manner when high-quality antibodies are available [4]. The 187 protein expression levels were considered as 187 features to be used for the cancer type classifications in this study.

Feature selection

The expression levels of 187 proteins may not all contribute equally to the classification. The maximum relevance minimum redundancy (mRMR) method [10–13] was employed to rank the importance of the 187 features in the training set. The 187 features can be ordered by using this method according to each feature's relevance to the target and according to the redundancy among the features themselves.

Let Ω denotes the whole set of 187 features, while Ω_s denotes the already-selected feature set which includes m features and Ω_t denotes the to-be-selected feature set which includes n features. The relevance D of the feature f in Ω_t with the cancer classes c can be calculated by:

$$D = I(f, c) \tag{1}$$

And the redundancy R of the feature f in Ω_t with the already-selected features in Ω_s can be calculated by:

$$R = \frac{1}{m}\sum_{f_i \in \Omega_s} I(f, f_i) \tag{2}$$

To obtain the feature f_j in Ω_t with maximum relevance with cancer classes c and minimum redundancy with the already-selected features Ω_s, Equation (1) and Equation (2) are combined as the mRMR function:

$$\max_{f_j \in \Omega_t}\left[I(f_j, c) - \frac{1}{m}\sum_{f_i \in \Omega_s} I(f_j f_i)\right] (j = 1, 2, \ldots, n) \tag{3}$$

The feature evaluation will continue 187 rounds. After these evaluations, a ranked feature list S by mRMR method can be obtained:

$$S = \{f_1{}', f_2{}', \ldots, f_h{}', \ldots, f_N{}'\} \tag{4}$$

The feature index h indicates the importance of feature. A feature with a smaller index h indicated that it had a better trade-off between the maximum relevance and the minimum redundancy, and it may contribute more in the classification.

Based on the ranked feature list in the mRMR table, we adopted the Incremental Feature Selection (IFS) method [14, 15] to determine the optimal feature set, or one that achieves the best classification performance. To perform this method, features in the mRMR table were added one by one from higher to lower rank.

When another feature had been added, a new feature set was generated. And we get 187 feature sets, and the i-th feature set is:

$$S_i = \{f_1, f_2, \ldots, f_i\}\ (1 \leq i \leq N) \tag{5}$$

Based on each of the 187 feature sets, the classifiers were built and tested on the training set with 10-fold cross validation. With Matthews Correlation Coefficient (MCC) of 10-fold cross validation calculated on training set, we obtain an IFS table with the number of features and the performance of them. S_{optimal} is the optimal feature set that achieves the highest MCC on training set. At last, the model was build with features from S_{optimal} on training set and elevated on the test set.

Prediction methods

We randomly divided the whole data set into a training set and an independent test set. The training set was further partitioned into 10 equally sized partitions. The 10-fold cross-validation on the training set was applied to select the features and build the prediction model. The constructed prediction model was tested on the independent test set. The framework of model construction and evaluation was shown in **Fig 1**.

We tried the following four machine learning algorithms: SMO (Sequential minimal optimization), IB1 (Nearest Neighbor Algorithm), Dagging, RandomForest (Random Forest), and selected the optimal one to construct the classifier. The brief description of these algorithms was as below.

The SMO method is one of the popular algorithms for training support vector machines (SVM) [16]. It breaks the optimization problem of a SVM into a series of the smallest possible sub-problems, which are then solved analytically [16]. To tackle multi-class problems, pairwise coupling [17] is applied to build the multi-class classifier.

IB1 is a nearest neighbor classifier, in which the normalized Euclidean distance is used to measure the distance of two samples. For a query test sample, the class of a training sample with minimum distance is assigned to the test sample as the predicted result. For more information, please refer to Aha and Kibler's study [18].

Dagging is a meta classifier that combines multiple models derived from a single learning algorithm using disjoint samples from the training dataset and integrates the results of these models by majority voting [19]. Suppose there is a training dataset $\Im$ containing n samples. k subsets are constructed by randomly taking samples in $\Im$ without replacement such that each of them contain n' samples, where $kn' \leq n$. A selected basic learning algorithm is trained on these k subsets, thereby inducing k classification models M_1,M_2,. . .,M_k. For a query sample, $M_i(1 \leq i \leq k)$ provides a predict result and the final predicted result of Dagging is the class with most votes.

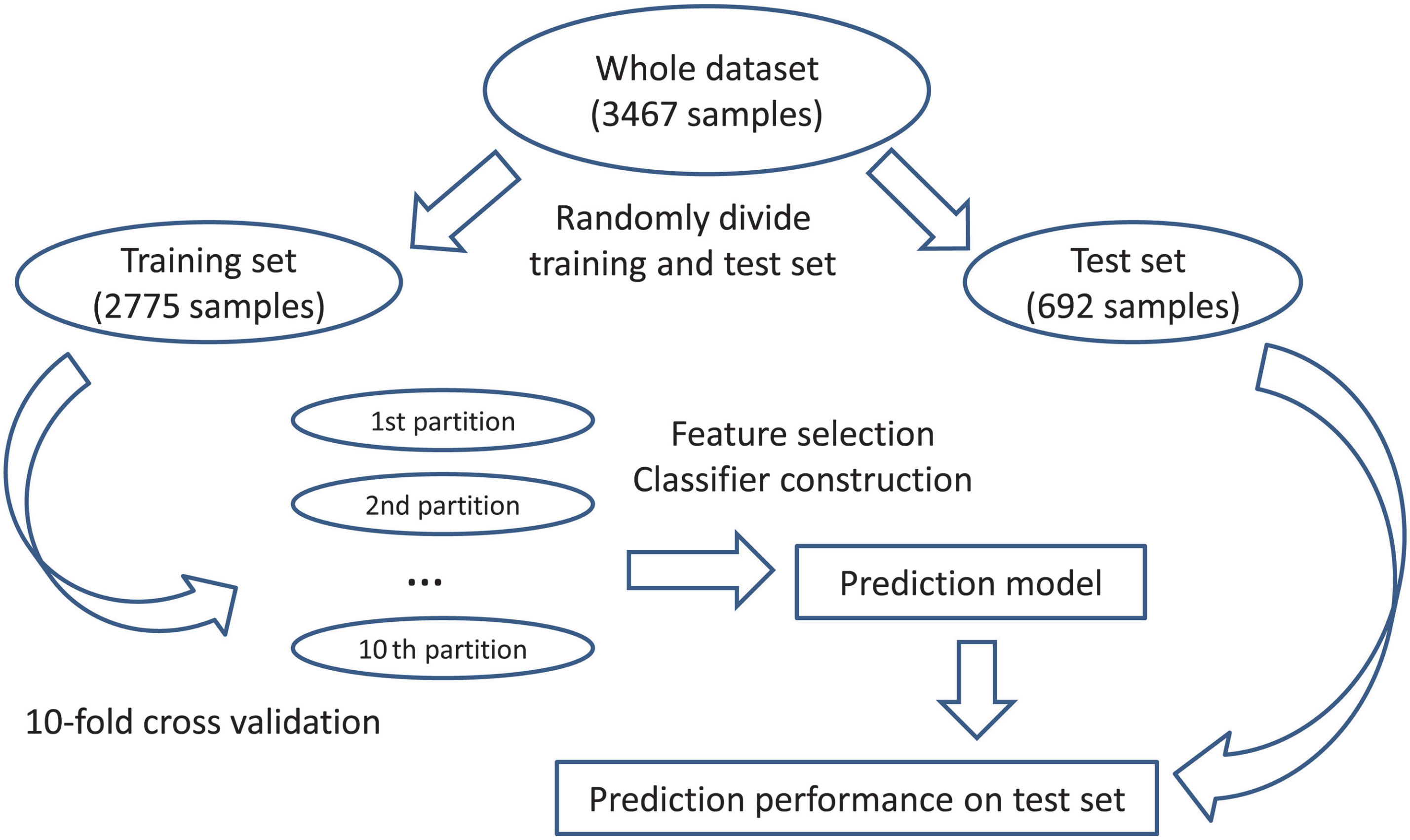

Fig 1. The workflow of model construction and evaluation. First, we randomly divided the whole data set into a training set and an independent test set. Then, the training set was further partitioned into 10 equally sized partitions to perform 10-fold cross validation. Based on the training set, the features were selected and the prediction model was built. At last, the constructed prediction model was tested on the independent test set.

doi:10.1371/journal.pone.0123147.g001

Random Forest algorithm was first proposed by Loe Breiman [20]. It is an ensemble predictor consisting of multiply decision trees. Suppose there are n samples in the training set and each sample was represented by M features. Each tree is constructed by randomly selecting N, with replacement, from the training set. At each node, randomly select m features and select the optimized split to grow the tree. After constructing multiply decision trees, the predicted result of a given sample is the class that receives the most votes from these trees.

Matthews Correlation Coefficient (MCC)

MCC [21], a balanced measure even if the classes are of very different sizes, is often used to evaluate the performance of prediction methods on a two-class classification problem. To calculate the MCC, one must count four values: true positives (TP), false positive (FP), true negative (TN) and false negative (FN) [22, 23]. Then, the MCC can be computed by

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TN + FN) \cdot (TN + FP) \cdot (TP + FN) \cdot (TP + FP)}} \tag{6}$$

However, many problems involve more than two classes, say N classes encoded by 1,2,. . .,N ($N > 2$). In this case, we can calculate the MCC for class i to partly measure the performance of prediction methods by counting TP, FP, TN and FN as following manners:

TP_i: the number of samples such that class i is their predicted class and true class;

FP$_i$: the number of samples such that class i is their predicted class and class i is not their true class;

TN$_i$: the number of samples such that class i is neither their predicted class nor their true class;

FN$_i$: the number of samples such that class i is not their predicted class and class i is their true class.

Accordingly, MCC for class i, denoted by MCC$_i$, can be computed by

$$\mathrm{MCC}_i = \frac{TP_i \cdot TN_i - FP_i \cdot FN_i}{\sqrt{(TN_i + FN_i) \cdot (TN_i + FP_i) \cdot (TP_i + FN_i) \cdot (TP_i + FP_i)}} \tag{7}$$

However, these values can't completely measure the performance of prediction methods, the overall MCC in multiclass case is still necessary. Fortunately, Gorodkin [24] has reported the MCC in multiclass case, which was used to evaluate the performance of the prediction methods mentioned in Section "Prediction methods". In parallel, The MCC for each class will also be given as references. Here, we gave the brief description of the overall MCC in multiclass case as below.

Suppose there is a classification problem on n samples, say $s_1, s_2, \ldots, s_n$, and N classes encoded by $1, 2, \ldots, N$. Define a matrix Y with n rows and N columns, where $Y_{ij} = 1$ if the i-th sample belongs to class j and $Y_{ij} = 0$ otherwise. For a classification model, its predicted results on the problem can be represented by two matrices X and C, where X has n rows and N columns,

$$X_{ij} = \begin{cases} 1 & \mathit{if} \quad \text{the } i\text{−th} \quad \text{sample} \quad \text{is} \quad \text{predicted} \quad \text{to} \quad \text{be} \quad \text{class} \quad j \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and C has N rows and N columns, C_{ij} is the number of samples in class i that have been predicted to be class j.

For Matrices X and Y, their covariance function can be calculated by

$$\mathrm{cov}(X, Y) = \frac{1}{N}\sum_{k=1}^{N} \mathrm{cov}(X_k, Y_k) = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{N}(X_{ik} - \bar{X}_k)(Y_{ik} - \bar{Y}_k) \tag{9}$$

where X_k and Y_k are the k-th column of matrices X and Y, respectively, $\bar{X}_k$ and $\bar{Y}_k$ are mean value of numbers in X_k and Y_k, respectively. Then, the MCC in multiclass case can be computed by the following formulation [25]:

$$\begin{aligned} MCC &= \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{cov}(X, X)\mathrm{cov}(Y, Y)}} \\ &= \frac{\sum_{k,l,m}^{N}(C_{kk}C_{ml} - C_{lk}C_{km})}{\sqrt{\sum_{k=1}^{N}[(\sum_{l=1}^{N} C_{lk})(\sum_{f,g=1,f\neq g}^{N} C_{gf})]}\sqrt{\sum_{k=1}^{N}[(\sum_{l=1}^{N} C_{kl})(\sum_{f,g=1,f\neq g}^{N} C_{fg})]}} \end{aligned} \tag{10}$$

Like the MCC in two-class case, the MCC in multiclass case ranges between -1 and 1, where 1 indicates the perfect classification, -1 the extreme misclassification.

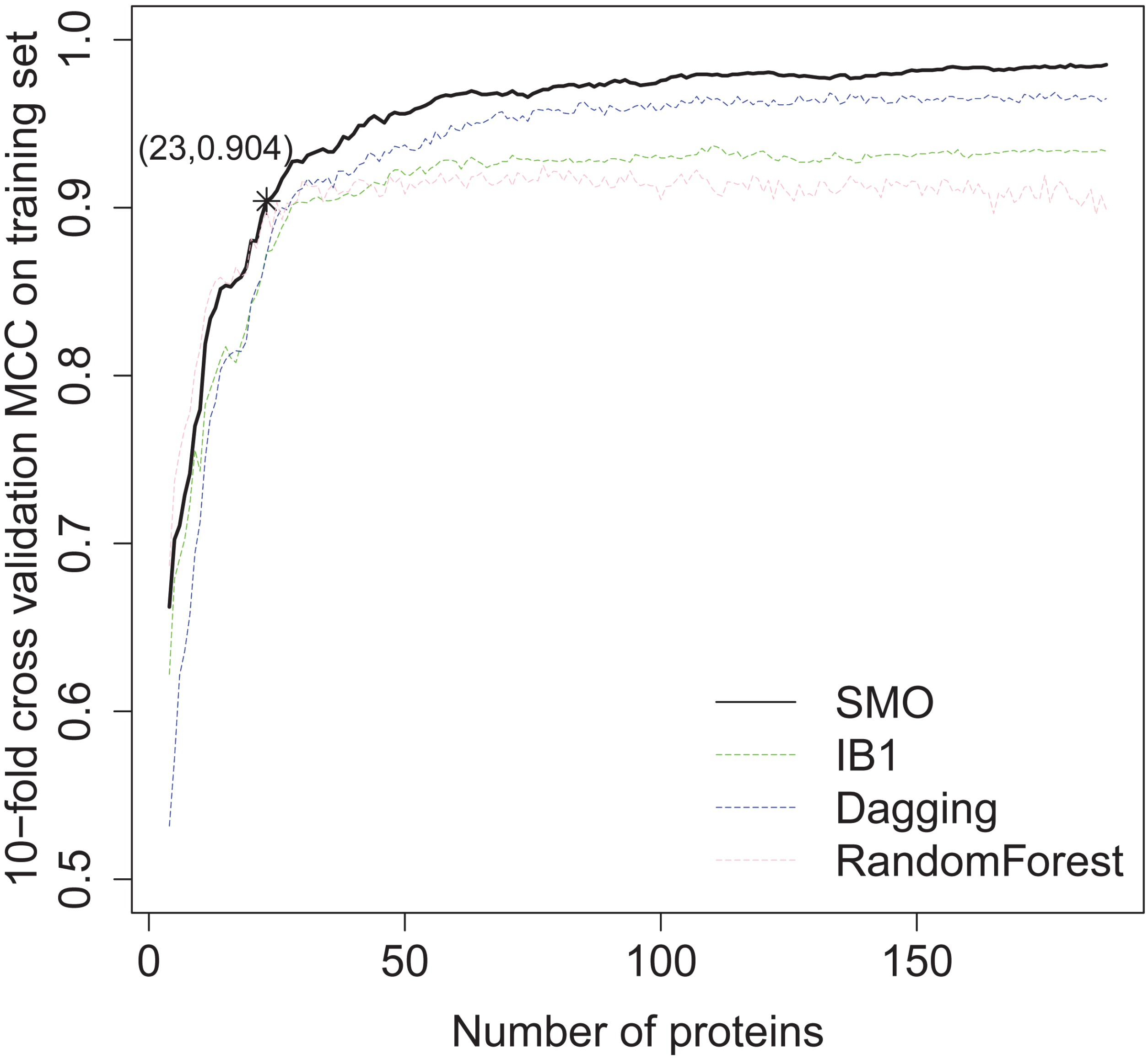

Fig 2. The IFS curves for the classifying of the 10 types of tumors. Plot to show the MCCs of the different classifiers constructed by different number of protein features selected from the mRMR table during the IFS process on training set. When the first 23 proteins were selected, the MCC reached 0.904, which was the first reach above 0.900 and with more protein features, the MCC did not increase much. We considered the 23 proteins as the most significant proteins for the classification.

doi:10.1371/journal.pone.0123147.g002

Results and Discussion

The mRMR and IFS results

By using the maximum relevance minimum redundancy (mRMR) method, the 187 features were ranked by importance in the training set. The result of the mRMR table can be found in S2 File.

During the IFS approach, each protein feature was added one by one. The classification MCCs which were obtained by four prediction methods, on the training set evaluated by 10-fold cross validation are presented in S3 File. We depicted the classification MCCs as Fig 2 from the data in S3 File. It can be observed that the highest MCCs for SMO, IB1, Dagging and

RandomForest were 0.985, 0.937, 0.969 and 0.925, indicating SMO can be used to construct an optimal classifier. By carefully checking the predicted results of SMO, it can be seen that by using the top 23 proteins, the MCC reached 0.904 which was the first reach above 0.900. With more proteins, the MCC did not increase by much. Therefore, in this study, we considered the 23 proteins as the optimal feature set and these 23 proteins were regarded as the most important proteins in classifying these ten types of cancers. We evaluated their prediction performance on the independent test set and the MCC was 0.936. The MCC for each cancer type can be found in S3 File.

The selected top 23 proteins for distinguishing cancer types

The selected top 23 proteins are summarized in Table 2. These proteins may play important roles in classifying the ten different cancer types. Most of these proteins have been reported to

Table 2. The top 23 important proteins for the classification of the 10 cancer types.

Order	Name	Gene Name	Protein function and regulatory pathways
1	FASN	FASN	Fatty acid synthase (FASN) catalyzes the synthesis of long-chain fatty acids from acetyl-CoA and malonyl-CoA. Indicated as a poor prognosis in breast and prostate cancer.
2	Claudin-7	CLDN7	Claudins make up tight junction strands.
3	PR	PGR	Progesterone receptor, Transcription Factor
4	TIGAR	C12ORF5	Regulates p53 tumor suppressor pathway and glycolysis
5	GATA3	GATA3	Transcription factor
6	NDRG1_pT346	NDRG1	A member of the NDRG family functions in growth, differentiation, and cell survival
7	AR	AR	Androgen receptor (AR). Transcription Factor
8	PREX1	REX1	Downstream of Heterotrimeric G proteins (Guanine nucleotide exchange factor)
9	PEA15_pS116	PEA15	Implicated in the regulation of multiple cellular processes including apoptosis, integrin activation, and insulin-sensitive glucose transport in insulin-responsive cells. Its activation is mediated through binding to multiple proteins, including ERK1&2, RSK2, Akt, FADD, and Caspase-8.
10	Cyclin_B1	CCNB1	Cyclin B1 regulates mitosis. Cyclin B1 levels rise during S phase and G2, and peak at mitosis.
11	ER-alpha	ESR1	Estrogen receptor, Transcription Factor
12	AMPK_alpha	PRKAA1	Involved in energy homeostasis regulation
13	Acetyl-a-Tubulin-Lys40		The cytoskeleton consists of three types of cytosolic fibers: microtubules, microfilaments (actin filaments), and intermediate filaments. Acetylation of α-tubulin at Lys40 is required for dynamic cell shape remodeling, cell motility, tubulin stability and terminal branching of cortical neurons
14	Rab-25	Rab-25	A member of Rab11 family possesses small Ras-like GTPase activity. Increased Rab25 expression is associated with aggressive growth in ovarian and breast cancer, where Rab25 may inhibit apoptosis and promote cancer cell proliferation and invasion through regulation of vesicle transport and cellular motility.
15	Chk2	CHEK2	Kinase acts downstream of ATM/ATR involving in DNA damage checkpoint control, embryonic development, and tumor suppression
16	E-Cadherin	CDH1	A member of transmembrane glycoprotein superfamily, Mediate calcium-dependent cell-cell adhesion and normal tissue development.
17	ACC1	ACACA	Key enzyme in the biosynthesis and oxidation of fatty acids. Involved in energy homeostasis regulation
18	GAPDH	GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
19	PKC-alpha_pS657	PRKCA	PKC alpha is an ubiquitously expressed PKC isozyme that has been implicated in the regulation of a broad range of cellular functions
20	TRFC	TFRC	Transferrin Receptor
21	Cyclin_E1	CCNE1	Cyclin E has been found to be associated with the transcription factor E2F in a temporally regulated manner. The cyclin E/E2F complex is detected primarily during the G1 phase of the cell cycle and decreases as cells enter S phase. E2F is known to be a critical transcription factor for expression of several S phase specific proteins.
22	CD20	CD20	A surface molecule of B-lymphocyte during the differentiation of B-cells into plasma cells
23	GAB2	GAB2	A docking protein, which mainly mediates the interaction between receptor tyrosine kinases (RTKs) and non-RTK receptors.

doi:10.1371/journal.pone.0123147.t002

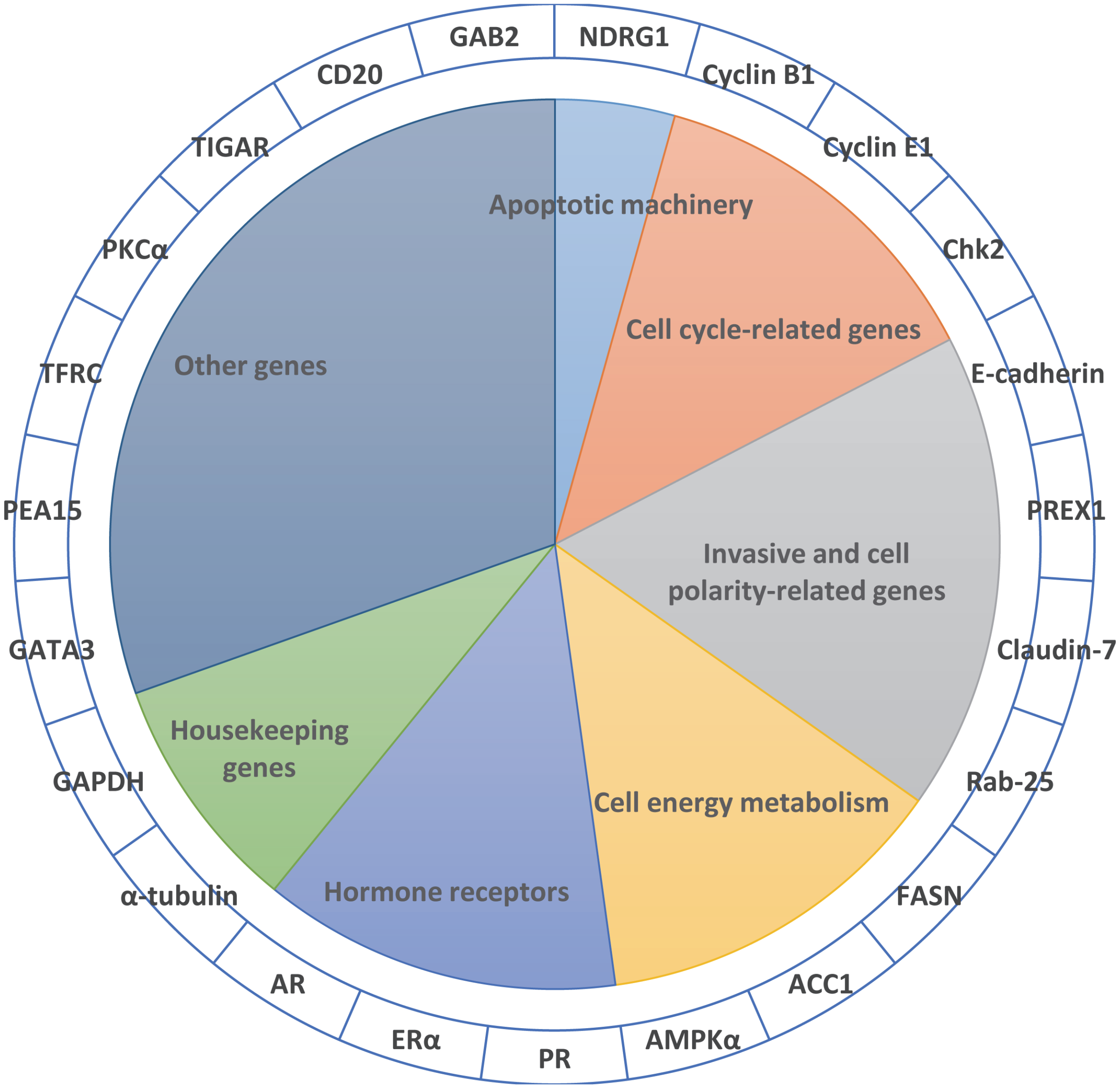

Fig 3. The 23 selected proteins. The 23 selected proteins are ascribed to seven sections mainly based on hallmarks of cancer. For those that are not associated with cancer-related pathways, we put genes with similar functions together to discuss.

doi:10.1371/journal.pone.0123147.g003

be related to certain tumors. For example, Claudin-7 has been reported to be over-expressed in breast tumors [26] and down-regulated in head and neck carcinomas [27]. TIGAR is up-regulated in colon tumors [28]. Gene amplification of ESR1 occurs frequently with breast cancer [29]. PREX1 is highly expressed in prostate cancer [30]. Thus, our findings are further corroborated by these previous results. Below, we will discuss the biological significance of the 23 proteins in detail based on gene function, cell pathways and biological functions, which may shed some light on the differences of different cancers in protein expression levels. We mainly discuss these genes in sections according to Robert A. Weinburg's [31]. For some genes that do not apply to cancer's hallmarks, we try to put these genes with similar functions together for discussion (see Fig 3).

Preventing cell death is crucial for cancer development because cancer cells are often resistant to apoptotic signaling caused by DNA damage and other factors. In our results, we found one gene that is related to apoptotic machinery and could be used to distinguish different cancers. Here, we discuss NDRG1, as well as previous findings showing its relationship to cancer. NDRG1 (N-myc downstream regulated gene 1) is a phosphorylated protein [32] that could be activated by the tumor suppressor gene p53 and required for the induction of p53-mediated apoptosis in the colon cancer cell line [33]. Because the NDRG1 protein has a crucial role in inhibiting primary tumor growth, it is well-known as a metastasis suppressor in a number of cancers including colon, prostate and breast cancers [34].

Replicative immortality is an important hallmark of cancer, which is commonly recognized as deregulated cell proliferation. Our findings on several important cell cycle-related genes in selected proteins not only illustrate their importance to the development of cancer, but are also first used as indicators of cancer classification. These cell cycle-related genes are discussed below: Cyclin B1 has a role in the regulation of cell cycle: before entering mitosis, cells flip between G2 and mitosis until there is sufficient accumulation of cyclin B to support CDK1 activity [35]. Misexpressed cyclin B1 in the nucleus has been found in a huge proportion of cells of some neoplasms, and cyclin B1 has been regarded as a potent prognostic factor in human breast carcinoma and squamous cell carcinoma [36]. Cyclin E1, encoded by CCNE1, is one of the members of the cyclin family, which controls cell cycle processes by dramatic periodicity of abundance. Recently, a genome-wide association study found that rs8102137 within the CCNE1 gene is associated with bladder cancer [37]. Meta-analysis also indicates that there is over-expression of this protein with breast cancer [38]. Chk2 (checkpoint kinase 2), as a serine/threonine protein kinase, could respond to DNA damage in order to maintain genomic integrity [39]. It has been shown that Chk2 plays an important role in the proliferation of cancer cells [40], attracting much attention to make it a possible anti-cancer drug design target [41].

It is clear that invasion is a hallmark of cancer, even if its underlying mechanisms are still an enigma. Until now, the gain and loss of cell-cell attachment proteins are the main reasons of invasion, especially the loss of E-cadherin [31]. In our results, E-cadherin and some polarity-related proteins are found that could be used to distinguish different cancer types. These proteins are discussed below: E-Cadherin, as the type-1 classical cadherin, mediates cell interactions. Tumor progression is often linked with the loss of E-cadherin function, leading to a more motile and invasive phenotype [42]. PREX1 (phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor) is highly expressed in prostate cancer, indicating a relationship between the cell invasion and its expression [30]. In melanomas, PREX1 over-expression was connected to the activation of ERK-MAPK signaling and required for efficient melanoblast metastasis as well as for migration [43]. Claudin-7, a common transmembrane protein, plays a vital role in the formation and maintenance of the permeability in polarized epithelial cells [44]. The aberrant Claudin-7 expression profile has been found in various tumors, such as highly induced Claudin-7 expression in both primary and metastatic breast tumors, [26] yet it is down-regulated in head and neck carcinomas [27]. These previous studies further supported our findings that Claudin-7 could be used as a biomarker for the differentiation and classification of various tumors. Rab-25, as a member of the Rab family of GTPases, Rab-25 is a constitutively active Rab GTPase that plays a crucial role in apical recycling and transcytosis pathways in polarized epithelial cells. Because loss of cell polarity is an essential hallmark of cancer, Rab-25 related trafficking has an important impact on epithelial cell polarity program in cancer progression [45].

Anomalous cancer cell energy metabolism was first observed by Otto Warbugy in 1930 and has been accepted as a hallmark of cancer. Abnormal fatty-acid synthesis as one type of energy metabolism is found in many cancer cells [46]. Here, several important fatty acid and glycolytic metabolism-related genes are found in the selected 23 proteins: FASN is a key enzyme which is required for de novo synthesis of fatty acid. It has been found that the FASN expression and activity are abnormally elevated in many types of human cancers, which may contribute to cellular resistance to drug- and radiation-induced apoptosis [46]. ACC1 is a rate-limiting enzyme in de novo fatty acids synthesis. It seems to be the limiting enzyme in proliferating cancer cells. ACC1 has been found to be up-regulated in proliferating cancer cell lines such as prostate, breast and liver. Indeed, it has been shown that knock-down of ACC1 by siRNA promotes apoptosis in prostate cancer and breast tumor cells but not in control noncancerous cells, underlining cancer cells' higher reliance on this enzyme than normal tissue [47]. AMPK (AMP-activated protein kinase, encoded by the gene PRKAA1/2) plays a crucial role in sensing available energy and coordinating external growth signals with cellular metabolism [48]. A decrease of AMPK signaling, mostly caused by the loss of function gene STK11, could lead to increased activation of mTOR and a shift toward glycolytic metabolism, which is found in a variety of cancers, including NSCLC [49] and cervical cancer [50].

Abnormal expression of hormone receptors are often shown in sex-related cancers, such as breast cancer and prostate cancer. Three hormone receptors are also reported in the selected proteins: Progestin receptor (PR), as a nuclear steroid receptor, has a high specificity for binding progesterone [51]. It has been shown in literature that PR inhibits the transition from G1 to S in the cell cycle and promote apoptosis in endometrial cancer cells [52]. In the GOG119 phase II trial, an estrogen surrogate named tamoxifen could enhance progestin activity in order to induce PR and cure endometrial patients [53]. Estrogen receptor (ER, activated by the hormone estrogen) is one of the most important therapeutic targets in breast cancers, given that the correlation between ER expression and cellular response to estrogen [54]. It has been reported that gene amplification of ESR1 frequently occur with breast cancer [29]. Androgen receptor (AR; NR3C4) is believed to solely mediate all the biological actions of endogenous, functioning mainly in regulating male development. Due to the strong connection between ARs and prostate cancer, androgen antagonists or androgen deprivation therapy has been applied to impede cancer cell proliferation of patients with androgen-dependent prostate cancer in clinical treatment [55].

Surprisingly, among these 23 selected proteins that are used to distinguish different cancers, α-tubulin and GAPDH are often used as controls in western blot analysis. In the following part, we will discuss known findings about α-tubulin and GAPDH that lend credence to the validity of our findings for their importance to distinguish cancers. For example, both α- and β-tubulin proteins are responsible for assembling microtubules (MTs, cytoskeletal polymeric structures), and certain posttranslational modifications. The acetylation of α-tubulin (Lys-40) [56] could alter dynamic behavior of MTs, which may lead to changes in biological functions that MTs perform during cell division, migration, and intracellular trafficking. Taking the dynamic parameters into account, MTs provide an attractive target for chemotherapy against rapidly growing tumor cells such as in lymphoma and leukemia, metastatic cancers, and slow growing tumors of the breast, ovary, and lung [57, 58]. Over the last decade, GAPDH (glyceraldehyde-3-phosphate dehydrogenase) was considered a housekeeping gene and was as a control for equal loading during the experimental process. However, it has been shown that GAPDH expression varies different types of tissues. Moreover, GAPDH expression varies due to oxygen tension [59], and the expression levels of GAPDH vary in fallopian tube cancers and ovarian cancers [60]. On the basis of GAPDH’s predilection for AU-rich elements, it has been shown that GAPDH can bind to the CSF-1 3'UTR that stabilize the mRNA [60]. To summarize, combining all the evidence, tubulin proteins and GAPDH may bring a new perspective on cancer studies, and it is suggested that they are not used as controls in western blot analysis of different types of cancer.

Other selected proteins include phosphatases, transcriptional activators, linker proteins and transferrin receptors: GATA3 is a transcriptional activator with high expression levels [61] and the third most frequently mutated gene in breast cancer [62]. Thus, GATA3 has proved to be a useful immunohistochemical marker to predict tumor recurrence early in the progression of breast cancer. PEA15, as a multifunctional linker protein predominantly expressed in the cells of the nervous system, such as astrocytes [63], controls a variety of cellular processes, such as cell survival, proliferation, migration and adhesion [64]. PEA15 functions in various cancers, concluding glioblastoma, astrocytoma, and mammary, as well as skin cancers. PEA15 can have both anti- (in ovarian carcinoma [65]) and pro- (glioblastoma [66]) tumorigenic functions, depending on its interactions. TFRC is a transferrin receptor. It is a major iron importer in most mammalian cells. It has been shown that TFRC proteins increase in breast, malignant pancreatic cancer, and other cancers [67, 68]. PKCα is encoded by PRKCA gene and is a serine- and threonine- specific kinase. This gene is highly expressed in multiple cancers, and the high activation of PKCα has been identified to promote the genesis of breast cancer [69]. The high abundance in serum makes this protein to be a good diagnostic biomarker of lung cancer [70] and gastric carcinoma [71]. TIGAR is a fructose-2-6-bisphosphatase that promotes the production of antioxidant (NADPH) and nucleotide synthesis material (ribose-5-phosphate) and seems to be important for tissue renewal and intestinal tumorigenesis. Up-regulated expression of TIGAR in human colon tumors along with other evidence suggest its importance in the development of cancer and metabolism regulation and may be used as a therapeutic target in diseases such as intestinal cancer [28]. CD20 (Membrane-Spanning 4-Domains, Subfamily A, Member 1, MS4A1) encodes a surface molecule B-lymphocyte during the differentiation of B-cells into plasma cells. Currently, a CD20 monoclonal antibody has been utilized in the treatment of cancer, even though its dosage is still under discussion [72]. GAB2 (GRB2-associated-binding protein 2) is a docking protein, which mainly interacts with signaling molecules. Research has shown that the oncogenesis of many cancers including gastric, colon, ovarian and breast cancer is related to GAB2 [73, 74]. For example, GAB2 can amplify the signal of receptor tyrosine kinases (RTKs), which plays roles in breast cancer development and progression [75].

As shown above, all of the top 23 proteins are closely related to certain types of cancers. Researchers have focused on common features of different cancer types for decades [31]. Admittedly, in theory, the hallmarks of cancer would help us develop drugs to treat all types of cancers as a whole. However, this “one size fits all” cancer treatment has disappointed us due to its treatment-related toxicity and inefficiency. Despite the fact that personalized treatments have been proposed, the theory still stays at a conceptual phase. Thus, having a better understanding of the potential values and the applied ranges of cancer drugs based on different biomarkers may be a more realistic way to treat different types of cancers.

Potential values of our findings

Previous experimental studies in the literature could consolidate our results showing that the selected 23 proteins could be used as biomarkers for certain cancers. They also can explain partially why the combination of these proteins could be used to accurately classify different cancer types. However, to our knowledge, reasons behind the varying expression patterns in different types of cancers have not been found. At least, by using our computational method, one could gain a better understanding of the similarities and differences among different cancers. This could help us identify proteins that could promote the development of cancers and proteins that might not be indispensable for cancer development. Further studies should be performed to determine whether the differential expression patterns of proteins in various cancers are influenced by their original tissues. Those proteins specifically expressed in certain types of cancers could be considered as potential specific cancer targets, which could be used to improve the target efficiency. Therefore, our results may help drug designers obtain a better understanding of the potential targets of drugs by shedding some light on the cancer type-specific biomarker discoveries.

Supporting Information

S1 File. The dataset used in this study. There were 3467 cancer patient samples in 10 cancer types, with 187 proteins for each sample. The 3467 samples were randomly divided into 2775 training samples and 692 independent test samples. The first column is the sample ID, the second column is the cancer types whose description can be found in Table 1. The third to the 189th columns were proteins.
(XLSX)

S2 File. The mRMR table. All the 187 protein features were ranked from the most important to the least by using the mRMR method on training set. The top 23 proteins were regarded as composing the optimal feature set because by using the 23 protein features, the MCC on the training set evaluated by 10-fold cross validation reached 0.904 which was the first reach above 0.900, and with more protein features, the MCC did not increase much.
(XLSX)

S3 File. The classification MCCs of four prediction methods, SMO (Sequential minimal optimization), IB1 (Nearest Neighbor Algorithm), Dagging and RandomForest (Random Forest), on the training set evaluated by 10-fold cross validation and the MCC of SMO with 23 features on test set.
(XLSX)

Author Contributions

Conceived and designed the experiments: TH XYK YDC. Performed the experiments: PWZ TH. Analyzed the data: PWZ LC TH. Contributed reagents/materials/analysis tools: YDC. Wrote the paper: PWZ TH NZ LC.

References

1. Viet CT, Schmidt BL. Understanding oral cancer in the genome era. Head & neck. 2010; 32(9):1246–68.
2. Mazumder A, Palma AJF, Wang Y. Validation and integration of gene-expression signatures in cancer. 2008.
3. Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V, et al. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. Molecular & Cellular Proteomics. 2005; 4(4):346–55.
4. Spurrier B, Ramalingam S, Nishizuka S. Reverse-phase protein lysate microarrays for cell signaling analysis. Nature protocols. 2008; 3(11):1796–808. Epub 2008/11/01. doi: doi: 10.1038/nprot.2008.179. PubMed PMID: PMID: 18974738.
5. Gundisch S, Grundner-Culemann K, Wolff C, Schott C, Reischauer B, Machatti M, et al. Delayed times to tissue fixation result in unpredictable global phosphoproteome changes. Journal of proteome research. 2013; 12(10):4424–34. Epub 2013/08/30. doi: doi: 10.1021/pr400451z. PubMed PMID: PMID: 23984901.
6. Sonntag J, Bender C, Soons Z, der Heyde Sv, König R, Wiemann S, et al. Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer. Translational Proteomics. 2014; 2:52–9.
7. Liu H, Setiono R. Incremental feature selection. Applied Intelligence. 1998; 9(3):217–30.
8. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. Nature methods. 2013.
9. Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407):330–7. doi: 10.1038/nature11252 PMID: 22810696

10. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2005; 27(8):1226–38.

11. Li B-Q, Hu L-L, Chen L, Feng K-Y, Cai Y-D, Chou K-C. Prediction of protein domain with mRMR feature selection and analysis. PLoS One. 2012; 7(6):e39308. doi: 10.1371/journal.pone.0039308 PMID: 22720092

12. Li B-Q, Huang T, Liu L, Cai Y-D, Chou K-C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PloS one. 2012; 7(4):e33393. doi: 10.1371/journal.pone.0033393 PMID: 22496748

13. Jiang Y, Li B-Q, Zhang Y, Feng Y-M, Gao Y-F, Zhang N, et al. Prediction and Analysis of Post-Translational Pyruvoyl Residue Modification Sites from Internal Serines in Proteins. PloS one. 2013; 8(6): e66678. PMID: 23805260

14. He Z, Zhang J, Shi X-H, Hu L-L, Kong X, Cai Y-D, et al. Predicting drug-target interaction networks based on functional groups and biological features. PloS one. 2010; 5(3):e9603. doi: 10.1371/journal.pone.0009603 PMID: 20300175

15. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009; 37(1):1–13. doi: doi: 10.1093/nar/gkn923. PubMed PMID: PMID: 19033363; PubMed Central PMCID: PMC2615629.

16. Xu Z, Dai M, Meng D. Fast and efficient strategies for model selection of Gaussian support vector machine. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on. 2009; 39(5):1292–307.

17. Hastie T, Tibshirani R. Classification by pairwise coupling. Proceedings of the 1997 conference on Advances in neural information processing systems 10; Denver, Colorado, USA. 302744: MIT Press; 1998. p. 507–13.

18. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Machine learning. 1991; 6(1):37–66.

19. Ting KM, Witten IH, editors. Stacking bagged and dagged models. Fourteenth international Conference on Machine Learning; 1997; San Francisco, CA.

20. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.

21. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975; 405(2):442–51. PMID: 1180967

22. Chen L, Feng KY, Cai YD, Chou KC, Li HP. Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. Bmc Bioinformatics. 2010; 11:293. doi: Artn 293 Doi doi: 10.1186/1471-2105-11-293. PubMed PMID: ISI:000279734000001. PMID: 20513238

23. Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000; 16(5):412–24. PMID: 10871264

24. Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. Computational Biology and Chemistry. 2004; 28(5):367–74. PMID: 15556477

25. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE. 2012; 7(8):e41882. doi: 10.1371/journal.pone.0041882 PMID: 22905111

26. Erin N, Wang N, Xin P, Bui V, Weisz J, Barkan GA, et al. Altered gene expression in breast cancer liver metastases. International Journal of Cancer. 2009; 124(7):1503–16. doi: 10.1002/ijc.24131 PMID: 19117052

27. Ginos MA, Page GP, Michalowicz BS, Patel KJ, Volker SE, Pambuccian SE, et al. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. Cancer research. 2004; 64(1):55–63. PMID: 14729608

28. Bensaad K, Tsuruta A, Selak MA, Vidal M, Nakano K, Bartrons R, et al. TIGAR, a p53-inducible regulator of glycolysis and apoptosis. Cell. 2006; 126(1):107–20. PMID: 16839880

29. Holst F, Stahl PR, Ruiz C, Hellwinkel O, Jehan Z, Wendland M, et al. Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. Nature genetics. 2007; 39(5):655–60. PMID: 17417639

30. Qin J, Xie Y, Wang B, Hoshino M, Wolff DW, Zhao J, et al. Upregulation of PIP3-dependent Rac exchanger 1 (P-Rex1) promotes prostate cancer metastasis. Oncogene. 2009; 28(16):1853–63. doi: 10.1038/onc.2009.30 PMID: 19305425

31. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144(5):646–74. doi: doi: 10.1016/j.cell.2011.02.013. PubMed PMID: PMID: 21376230.

32. Bae D-H, Jansson PJ, Huang ML, Kovacevic Z, Kalinowski D, Lee CS, et al. The role of NDRG1 in the pathology and potential treatment of human cancers. Journal of clinical pathology. 2013; 66(11):911–7. doi: 10.1136/jclinpath-2013-201692 PMID: 23750037

33. Stein S, Thomas EK, Herzog B, Westfall MD, Rocheleau JV, Jackson RS, et al. NDRG1 is necessary for p53-dependent apoptosis. Journal of Biological Chemistry. 2004; 279(47):48930–40. PMID: 15377670

34. Kovacevic Z, Richardson DR. The metastasis suppressor, Ndrg-1: a new ally in the fight against cancer. Carcinogenesis. 2006; 27(12):2355–66. PMID: 16920733

35. Moore JD. In the wrong place at the wrong time: does cyclin mislocalization drive oncogenic transformation? Nature Reviews Cancer. 2013; 13(3):201–8. doi: 10.1038/nrc3468 PMID: 23388618

36. Nozoe T, Korenaga D, Kabashima A, Ohga T, Saeki H, Sugimachi K. Significance of cyclin B1 expression as an independent prognostic indicator of patients with squamous cell carcinoma of the esophagus. Clinical cancer research. 2002; 8(3):817–22. PMID: 11895914

37. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nature genetics. 2010; 42(11):978–84. doi: 10.1038/ng.687 PMID: 20972438

38. Gutman DA, Cooper L, Hwang SN, Holder CA, Gao J, Aurora TD, et al. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. Radiology. 2013; 267(2):560–9. doi: 10.1148/radiol.13120118 PMID: 23392431

39. Hirao A, Kong Y-Y, Matsuoka S, Wakeham A, Ruland J, Yoshida H, et al. DNA damage-induced activation of p53 by the checkpoint kinase Chk2. Science. 2000; 287(5459):1824–7. PMID: 10710310

40. Antoni L, Sodha N, Collins I, Garrett MD. CHK2 kinase: cancer susceptibility and cancer therapy–two sides of the same coin? Nature Reviews Cancer. 2007; 7(12):925–36. PMID: 18004398

41. Lountos GT, Jobson AG, Tropea JE, Self CR, Zhang G, Pommier Y, et al. Structural characterization of inhibitor complexes with checkpoint kinase 2 (Chk2), a drug target for cancer therapy. Journal of structural biology. 2011; 176(3):292–301. doi: 10.1016/j.jsb.2011.09.008 PMID: 21963792

42. Canel M, Serrels A, Frame MC, Brunton VG. E-cadherin–integrin crosstalk in cancer invasion and metastasis. Journal of cell science. 2013; 126(2):393–401.

43. Lindsay CR, Lawn S, Campbell AD, Faller WJ, Rambow F, Mort RL, et al. P-Rex1 is required for efficient melanoblast migration and melanoma metastasis. Nature communications. 2011; 2:555. doi: 10.1038/ncomms1560 PMID: 22109529

44. Turksen K, Troy T-C. Junctions gone bad: claudins and loss of the barrier in cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2011; 1816(1):73–9.

45. Agarwal R, Jurisica I, Mills GB, Cheng KW. The emerging role of the RAB25 small GTPase in cancer. Traffic. 2009; 10(11):1561–8. doi: 10.1111/j.1600-0854.2009.00969.x PMID: 19719478

46. Kuhajda FP. Fatty acid synthase and cancer: new application of an old pathway. Cancer research. 2006; 66(12):5977–80. PMID: 16778164

47. Brusselmans K, De Schrijver E, Verhoeven G, Swinnen JV. RNA Interference–Mediated Silencing of the Acetyl-CoA-Carboxylase-α Gene Induces Growth Inhibition and Apoptosis of Prostate Cancer Cells. Cancer research. 2005; 65(15):6719–25. PMID: 16061653

48. Macheda ML, Rogers S, Best JD. Molecular and cellular regulation of glucose transporter (GLUT) proteins in cancer. Journal of cellular physiology. 2005; 202(3):654–62. PMID: 15389572

49. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012; 150(6):1107–20. doi: 10.1016/j.cell.2012.08.029 PMID: 22980975

50. Wingo SN, Gallardo TD, Akbay EA, Liang M-C, Contreras CM, Boren T, et al. Somatic LKB1 mutations promote cervical cancer progression. PLoS One. 2009; 4(4):e5137. doi: 10.1371/journal.pone.0005137 PMID: 19340305

51. Yang S, Thiel KW, Leslie KK. Progesterone: the ultimate endometrial tumor suppressor. Trends in Endocrinology & Metabolism. 2011; 22(4):145–52.

52. Dai D, Wolf DM, Litman ES, White MJ, Leslie KK. Progesterone inhibits human endometrial cancer cell growth and invasiveness Down-regulation of cellular adhesion molecules through progesterone B receptors. Cancer research. 2002; 62(3):881–6. PMID: 11830547

53. Singh M, Zaino RJ, Filiaci VJ, Leslie KK. Relationship of estrogen and progesterone receptors to clinical outcome in metastatic endometrial carcinoma: a Gynecologic Oncology Group study. Gynecologic oncology. 2007; 106(2):325–33. PMID: 17532033

54. Weigel MT, Dowsett M. Current and emerging biomarkers in breast cancer: prognosis and prediction. Endocrine-related cancer. 2010; 17(4):R245-R62. doi: 10.1677/ERC-10-0136 PMID: 20647302

55. Matsumoto T, Sakari M, Okada M, Yokoyama A, Takahashi S, Kouzmenko A, et al. The androgen receptor in health and disease. Annual review of physiology. 2013; 75:201–24. doi: 10.1146/annurev-physiol-030212-183656 PMID: 23157556

56. Weber K, Schneider A, Westermann S, Müller N, Plessmann U. Posttranslational modifications of α- and β-tubulin in< i> Giardia lamblia, an ancient eukaryote. FEBS letters. 1997; 419(1):87–91. PMID: 9426225

57. Honore S, Pasquier E, Braguer D. Understanding microtubule dynamics for improved cancer therapy. Cellular and Molecular Life Sciences CMLS. 2005; 62(24):3039–56. PMID: 16314924

58. Pasquier E, Kavallaris M. Microtubules: a dynamic target in cancer therapy. IUBMB life. 2008; 60 (3):165–70. doi: 10.1002/iub.25 PMID: 18380008

59. Tricarico C, Pinzani P, Bianchi S, Paglierani M, Distante V, Pazzagli M, et al. Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. Analytical biochemistry. 2002; 309(2):293–300. PMID: 12413463

60. Barber RD, Harmer DW, Coleman RA, Clark BJ. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiological genomics. 2005; 21(3):389–95. PMID: 15769908

61. Mehra R, Varambally S, Ding L, Shen R, Sabel MS, Ghosh D, et al. Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. Cancer research. 2005; 65 (24):11259–64. PMID: 16357129

62. Jiang YZ, Yu KD, Zuo WJ, Peng WT, Shao ZM. GATA3 mutations define a unique subtype of luminal-like breast cancer with improved survival. Cancer. 2014.

63. Danziger N, Yokoyama M, Jay T, Cordier J, Glowinski J, Chneiweiss H. Cellular Expression, Developmental Regulation, and Phylogenic Conservation of PEA-15, the Astrocytic Major Phosphoprotein and Protein Kinase C Substrate. Journal of neurochemistry. 1995; 64(3):1016–25. PMID: 7861130

64. Ramos JW, Kojima TK, Hughes PE, Fenczik CA, Ginsberg MH. The death effector domain of PEA-15 is involved in its regulation of integrin activation. Journal of Biological Chemistry. 1998; 273(51):33897–900. PMID: 9852038

65. Bartholomeusz C, Rosen D, Wei C, Kazansky A, Yamasaki F, Takahashi T, et al. PEA-15 induces autophagy in human ovarian cancer cells and is associated with prolonged overall survival. Cancer research. 2008; 68(22):9302–10. doi: 10.1158/0008-5472.CAN-08-2592 PMID: 19010903

66. Xiao C, Yang BF, Asadi N, Beguinot F, Hao C. Tumor necrosis factor-related apoptosis-inducing ligand-induced death-inducing signaling complex and its modulation by c-FLIP and PED/PEA-15 in glioma cells. Journal of Biological Chemistry. 2002; 277(28):25020–5. PMID: 11976344

67. Miller LD, Coffman LG, Chou JW, Black MA, Bergh J, D'Agostino R, et al. An iron regulatory gene signature predicts outcome in breast cancer. Cancer research. 2011; 71(21):6728–37. doi: 10.1158/0008-5472.CAN-11-1870 PMID: 21875943

68. Ryschich E, Huszty G, Knaebel H, Hartel M, Büchler M, Schmidt J. Transferrin receptor is a marker of malignant phenotype in human pancreatic cancer and in neuroendocrine carcinoma of the pancreas. European Journal of Cancer. 2004; 40(9):1418–22. PMID: 15177502

69. Tam WL, Lu H, Buikhuisen J, Soh BS, Lim E, Reinhardt F, et al. Protein kinase C α is a central signaling node and therapeutic target for breast cancer stem cells. Cancer cell. 2013; 24(3):347–64. doi: 10.1016/j.ccr.2013.08.005 PMID: 24029232

70. Hudler P, Kocevar N, Komel R. Proteomic Approaches in Biomarker Discovery: New Perspectives in Cancer Diagnostics. The Scientific World Journal. 2014; 2014.

71. Masuda T-a, Inoue H, Sonoda H, Mine S, Yoshikawa Y, Nakayama K, et al. Clinical and biological significance of S-phase kinase-associated protein 2 (Skp2) gene expression in gastric carcinoma modulation of malignant phenotype by Skp2 overexpression, possibly via p27 proteolysis. Cancer research. 2002; 62(13):3819–25. PMID: 12097295

72. Taylor RP, Lindorfer MA. Analyses of CD20 Monoclonal Antibody-Mediated Tumor Cell Killing Mechanisms: Rational Design of Dosing Strategies. Molecular pharmacology. 2014. doi: doi: 10.1124/mol.114.092684. PubMed PMID: PMID: 24944188.

73. Simister PC, Feller SM. Order and disorder in large multi-site docking proteins of the Gab family—implications for signalling complex formation and inhibitor design strategies. Molecular BioSystems. 2012; 8 (1):33–46. doi: 10.1039/c1mb05272a PMID: 21935523

74. Vaughan TY, Verma S, Bunting KD. Grb2-associated binding (Gab) proteins in hematopoietic and immune cell biology. American journal of blood research. 2011; 1(2):130. PMID: 22163099

75. Wohrle F, Daly RJ, Brummer T. Function, regulation and pathological roles of the Gab/DOS docking proteins. Cell Commun Signal. 2009; 7(22):10.1186. doi: 10.1186/1478-811X-7-22 PMID: 19737390