Classification and Analysis of Regulatory Pathways Using Graph Property, Biochemical and Physicochemical Property, and Functional Property

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

1 Institute of Systems Biology, Shanghai University, Shanghai, People's Republic of China, **2** Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China, **3** Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China, **4** College of Information Engineering, Shanghai Maritime University, Shanghai, People's Republic of China, **5** Gordon Life Science Institute, San Diego, California, United States of America

Abstract

Given a regulatory pathway system consisting of a set of proteins, can we predict which pathway class it belongs to? Such a problem is closely related to the biological function of the pathway in cells and hence is quite fundamental and essential in systems biology and proteomics. This is also an extremely difficult and challenging problem due to its complexity. To address this problem, a novel approach was developed that can be used to predict query pathways among the following six functional categories: (i) "Metabolism", (ii) "Genetic Information Processing", (iii) "Environmental Information Processing", (iv) "Cellular Processes", (v) "Organismal Systems", and (vi) "Human Diseases". The prediction method was established trough the following procedures: (i) according to the general form of pseudo amino acid composition (PseAAC), each of the pathways concerned is formulated as a 5570-D (dimensional) vector; (ii) each of components in the 5570-D vector was derived by a series of feature extractions from the pathway system according to its graphic property, biochemical and physicochemical property, as well as functional property; (iii) the minimum redundancy maximum relevance (mRMR) method was adopted to operate the prediction. A cross-validation by the jackknife test on a benchmark dataset consisting of 146 regulatory pathways indicated that an overall success rate of 78.8% was achieved by our method in identifying query pathways among the above six classes, indicating the outcome is quite promising and encouraging. To the best of our knowledge, the current study represents the first effort in attempting to identity the type of a pathway system or its biological function. It is anticipated that our report may stimulate a series of follow-up investigations in this new and challenging area.

Citation: Huang T, Chen L, Cai Y-D, Chou K-C (2011) Classification and Analysis of Regulatory Pathways Using Graph Property, Biochemical and Physicochemical Property, and Functional Property. PLoS ONE 6(9): e25297. doi:10.1371/journal.pone.0025297

Editor: Cathal Seoighe, National University of Ireland Galway, Ireland

Received March 9, 2011; **Accepted** August 31, 2011; **Published** September 28, 2011

Funding: This work was supported by grants from National Basic Research Program of China (2011CB510102, 2011CB510101). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

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

These authors contributed equally to this work.

Introduction

During the past decade, much information on different organisms has been accumulated at both the genetic and metabolic levels; meanwhile, many specific databases, such as KEGG/LIGAND [1,2,3,4], ENZYME [5], BRENDA [6], EcoCyc and MetaCyc [7,8], have been developed. However, biological meaningful pathways, such as the regulatory pathway and metabolic pathway, are still poorly understood. As one of the most important pathways in systems biology, the regulatory pathway includes two kinds of interactions: direct protein–protein interactions (such as physical binding and phosphorylation) and indirect protein–protein interactions (such as the relations between transcription factors and downstream gene products) [2].

KEGG (Kyoto Encyclopedia of Genes and Genomes) [1,2,3,4] is a collection of online databases for dealing with genomes, enzymatic pathways, and biological chemicals. KEGG contains five main databases [4]: (i) KEGG Atlas, (ii) KEGG Pathway, (iii) KEGG Genes, (iv) KEGG Ligand, and (v) KEGG BRITE. The KEGG BRITE database (http://www.genome.jp/kegg/brite.html) includes some known regulatory pathways. It is an ontology database for representing functional hierarchies of various biological objects. The database also includes molecules, cells, organisms, diseases and drugs, as well as the relationships among them [9,10]. In this database, experimental knowledge is collected and diagramed as pathways, i.e. smaller networks of specific function. Several visualization tools have been developed to view and analyze the global networks through web interfaces [11,12,13].

According to the data in KEGG BRITE, regulatory pathways are classified into six pathway classes. Since different class pathway represents different biological function, developing a successful classifier to identify the pathway class is very useful in system biology. Some efforts have been made in this regard. Dale et al. [14] tried to predict whether a metabolic pathway is present or absent in an organism. In our previous work [15], we developed a model to predict whether a regulatory pathway can be formed for a system consisting of certain number of different proteins. But predicting the biological function of regulatory pathway is still an untouched problem. It is a big challenge in both systems biology and proteomics because this kind of information is very hard to recover and transform into the data that can be processed by computers. The purpose of this study is not to achieve a high accuracy, but to analyze some features, which may provide useful information for characterizing a meaningful regulatory pathway.

To realize this, some feature selection methods, such as the minimum redundancy maximum relevance [16] and incremental feature selection approaches, were employed to analyze the relevant features, while Nearest Neighbor Algorithm (NNA) [17,18], Sequential Minimal Optimization (SMO) [19,20] and Bayesian network (BayesNet) [21] were used to classify the pathways. Finally, the jackknife cross-validation [22] was adopted to evaluate the prediction performance. As a result, 49 features were selected as the optimal features and the overall accuracy by using these features was 78.8%.

It was suggested by analyzing the optimized features that biochemical and physicochemical property and functional property are important to determine the biological function of each regulatory pathway. Although it represents the first work ever in predicting the classification of regulatory pathways and it is still quite preliminary, we believe that our exploration can stimulate a series of follow-up studies in this area important to both system biology and proteomics.

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

Materials and Methods

Benchmark dataset

We downloaded the human KGML (KEGG XML) files from KEGG FTP site (ftp://ftp.genome.jp/pub/kegg/xml) in April 2009. We reduced the original data by the following two steps: (i) remove proteins without GO information or biochemical and physicochemical properties in each pathway; (ii) exclude pathways with less than three proteins. As a result, 146 regulatory pathways were obtained. According to the data in KEGG BRITE (http://www.genome.jp/kegg/brite.html), these pathways belong to the following six functional categories: (i) Metabolism, (ii) Genetic Information Processing, (iii) Environmental Information Processing, (iv) Cellular Processes, (v) Organismal Systems, and (vi) Human Diseases. Shown in **Table 1** is the distribution of the six classes of regulatory pathways in this study.

Table 1. The distribution of the 146 regulatory pathways.

Pathway class	Number of pathway
Metabolism	73
Genetic Information Processing	2
Environmental Information Processing	15
Cellular Processes	9
Organismal Systems	19
Human Diseases	28
Total	146

doi:10.1371/journal.pone.0025297.t001

Features construction

To develop a powerful predictor for classifying a protein system or pathway consisting of a set of proteins, one of the keys is to formulate the protein system with an effective mathematical expression that can truly reflect its intrinsic correlation with the attribute to be predicted [23]. In this regard, we can utilize the concept of pseudo amino acid composition (PseAAC) [24]. For a brief introduction about Chou's PseAAC, visit the Wikipedia web-page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems (see, e.g., [25,26,27,28,29,30,31,32,33,34]). For various different modes of PseAAC, see [35]. Actually, the general form of PseAAC can be formulated as (see Eq.6 of [23]):

$$\mathbf{P}=[\psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_\Omega]^{\mathbf{T}} \tag{1}$$

where $\mathbf{T}$ is a transpose operator, while the subscript Ω is an integer and its value as well as the components ψ_1, ψ_2, … will depend on how to extract the desired information from the amino acid sequence of $\mathbf{P}$. Likewise, a pathway $\mathbb{P}$ consisting a set of proteins can also be generally formulated as vector with Ω components; i.e.,

$$\mathbb{P}=[\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_u \quad \cdots \quad \Psi_\Omega]^{\mathbf{T}} \tag{2}$$

where Ψ_1 represents the 1st feature of the pathway, Ψ_2 the 2nd feature, and so forth. Below, let us elaborate how to define Ω as well as the components in Eq.2.

1. Graph property. Graphic approaches are deemed as useful tools to study complex biological systems as they can provide intuitive insights and the overall structure property, as indicated by various studies on a series of important biological topics [36,37,38,39,40,41,42,43,44,45,46,47,48]. To use the graphic approach for the current study, each regulatory pathway was represented as a graph, where the vertices represent proteins and the arcs represent the relations between the corresponding proteins. In fact, it is a directed graph or digraph [38,39]. This is because the relation between two proteins is directional; i.e., one protein, say $\mathbf{P}_1$, can regulate another protein, say $\mathbf{P}_2$, while $\mathbf{P}_2$ cannot always regulate $\mathbf{P}_1$. In this paper, we extracted 88 graph features from each directed graph that represents a regulatory pathway. Most of the graph features were derived in [49,50,51,52,53] where, however, the graphs are undirected. In this study, we extended them into directed graphs. The features of our directed graphs can be briefed as follows.

1. Graph size and graph density. Let $G=(V, E)$ be a pathway graph, where V denotes vertex set and E arcs set. The graph size is the number of vertices in the graph. $|E|_{\max}=|V|^2$ is the theoretical maximum number of arcs in G with $|V|$ vertices. The graph density is calculated by $|\mathrm{E}|/|E|_{\max}$ [49].
2. Degree statistics. The in-degree (out-degree) of a vertex is the number of its in-neighbors (out-neighbors). The mean, variance, median, and maximum of in-degree and out-degree, respectively, were taken as features in this feature group [50].

3. Edge weight statistics. Let $G=(V, w(E))$ be a weighted pathway graph where each arc is weighted by a weight w in the range of [0,1]. The symbol e is called a missing edge if $w(e)=0$. In this study, the mean and variance of the arc weights were considered as features, including two different cases (with and without missing edges) [49].
4. Topological change. Let $G=(V, w(E))$ be a weighted pathway graph. This group of features is to measure the topological changes when different cutoffs of the weights are applied to the graph. The weight cutoffs included 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8. Topology changes were defined as the change rate of the number of arcs in subgraphs under two consecutive cutoffs.
5. Degree correlation. Let $G=(V, E)$ be a pathway graph with $V=\{u_1,u_2,\ldots,u_n\}$. For each vertex u_i, calculate the average number of arcs of its in-neighbors and out-neighbors, respectively. Considered as features in this study were the mean, variance and maximum of the two kinds of property, respectively [51].
6. Clustering. Let $G=(V, E)$ be a pathway graph with $V=\{u_1,u_2,\ldots,u_n\}$. For each vertex u_i, calculate the graph density of the subgraph induced by its in-neighbors and out-neighbors, respectively. Take the mean, variance and maximum of the two kinds of property [50], respectively, as the features for the current study.
7. Topological. Let $G=(V, E)$ be a pathway graph with $V=\{u_1,u_2,\ldots,u_n\}$. Define four function as follows: (i) $in\text{-}in(u_i, u_j)$ for the number of both in-neighbors of u_i and in-neighbors of u_j; (ii) $in\text{-}out(u_i, u_j)$ for the number of both in-neighbors of u_i and out-neighbors of u_j; (iii) $out\text{-}in(u_i, u_j)$ for the number of both out-neighbors of u_i and in-neighbors of u_j; (iv) $out\text{-}out(u_i, u_j)$ for the number of both out-neighbors of u_i and out-neighbors of u_j. For each vertex u_i, calculate the four values T_{i1}, T_{i2}, T_{i3}, and T_{i4} as follows: (i) T_{i1} is the mean of $in\text{-}in(u_i, u_j)/n_{i1}$; (ii) T_{i2} the mean of $in\text{-}out(u_i, u_j)/n_{i1}$; (iii) T_{i3} the mean of $out\text{-}in(u_i, u_j)/n_{i2}$; (iv) T_{i4} the mean of $out\text{-}out(u_i, u_j)/n_{i1}$. In the above, n_{i1} and n_{i2} are the number of in-neighbors and out-neighbors of u_i, respectively. Take the mean, variance and maximum of T_{i1}, T_{i2}, T_{i3}, and T_{i4}, respectively, as the features [51] for the current study.
8. Singular values. Let A be the adjacent matrix of the pathway graph. Take the first three largest singular values [49] as the features for this study.
9. Local density change. Let $G=(V, E)$ be a pathway graph with $V=\{u_1,u_2,\ldots,u_n\}$. For each vertex u_i, let $V_{i1}=\{u_{i_1},u_{i_2},\ldots u_{i_k}\}$ and $V_{i2}=\{u_{j_1},u_{j_2},\ldots,u_{j_l}\}$ be its in-neighbors and out-neighbors, respectively. Here we only introduce how to extract features from out-neighbors of each vertex under the cutoff w, which may be 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. Construct a weighted undirected complete graph K_i with vertices $u_{j_1},u_{j_2},\ldots,u_{j_l}$ and the weights of each edge can be calculated by Eq. 2 in Section 2 "Gene ontology". Extract a spanning subgraph $G_i(w)$ of K_i with edges whose weights are greater than w. Calculate $L_i(w)=2|E(G_i(w))|/(l(l-1))$ ($L_i(w)=0$ if $l\le 1$). Take the mean and maximum of $L_1(w)$, $L_2(w)$,…, $L_n(w)$ under cutoff w as the features for the current study.

2. Gene ontology. As mentioned before, some features need the arc weight to evaluate the relation between two proteins. Thus, we used the information from gene ontology consortium (GO) [54] to represent each of the proteins concerned and evaluate its relation with the other proteins. "Ontology" is a specification of a conceptualization and refers to the subject of existence. GO is established according to the following three criteria: molecular function, biological process, and cellular component. Using GO information to represent protein samples can catch their core features [23] as proved by significantly enhancing the success rate in predicting their subcellular localization [55,56,57]. The GO approach has also been used to study protein-protein interactions [58,59]. Here, using the similar method as in [52], each protein sample can be formulated as a 5218-D vector:

$$\mathbf{P}=[p_1,p_2,\ldots,p_{5218}]^{\mathbf{T}} \tag{3}$$

where $p_i=1$ if the sample hit the i-th GO number; otherwise, $p_i=0$. The interaction between $\mathbf{P}_i$ and $\mathbf{P}_j$, *i.e.* the weight of arc between the two proteins, is defined by

$$w(\mathbf{P}_i,\mathbf{P}_j)=\frac{\mathbf{P}_i\cdot\mathbf{P}_j}{\|\mathbf{P}_i\|\cdot\|\mathbf{P}_j\|} \tag{4}$$

where $\mathbf{P}_i\cdot\mathbf{P}_j$ is the dot product of $\mathbf{P}_i$ and $\mathbf{P}_j$, and $\|\mathbf{P}_i\|$ and $\|\mathbf{P}_j\|$ are their modulus.

3. Biochemical and physicochemical property. Beside the graph property, the biological property of each pathway is also indispensable to characterize meaningful regulatory pathways. In this study, the biochemical and physicochemical properties, which have been used to study various biological problems [60,61,62], were employed to represent the biological property of each pathway. These properties included hydrophobicity, normalized van der Waals volume, polarity, polarizability, secondary structure, solvent accessibility, and amino acid compositions. For a regulatory pathway involving n proteins, both the mean and maximum values of their biological properties were taken for the features of the pathway, as detailed below.

1. Hydrophobicity, normalized van der Waals volume, polarity and polarizability: 42 features can be extracted from each of these properties [63,64], respectively. Here we only describe how to obtain the features from the hydrophobicity property, while features from the other properties can be obtained in a similar way. Each amino acid is substituted by one of the three letters, polar (P), neutral (N) and hydrophobic (H). Given a protein sequence, use P, N or H to substitute each amino acid in the sequence, and the sequence thus obtained is called a protein pseudo-sequence. Composition (C) is the percentage of P, N and H in the whole pseudo-sequence. Transition (T) is the changing frequency between any two characters. Distribution (D) is the sequence segment (in percentage) of the pseudo-sequence which is needed to contain the first, 25%, 50%, 75% and the last of the Ps, Ns and Hs, respectively. In conclusion, there are three, three, and fifteen properties for (C), (T) and (D), respectively. Accordingly, we have $(3+3+15)=21$ features for the "mean" category, $(3+3+15)=21$ feature for the "maximum" category, and hence a total of $(21\times 2)=42$ features by considering the "hydrophobicity" property alone. Similarly, we also have $(21\times 2)=42$ features by considering each of the other three properties, i.e., the "normalized van der Waals volume", "polarity", and "polarizability". Thus, we have a total of $42\times 4=168$ features by considering the above four properties.
2. Secondary structure: according to the secondary structural propensity of amino acids, each protein sequence can also be coded with three letters [65,66]. Thus, like the case in considering hydrophobicity, we also have $21\times 2=42$ features by considering the "secondary structure" property (or propensity).

Table 2. A breakdown of the 264 features for a pathway system by considering its biochemical and physicochemical properties.

Properties	C	T	D	Mean category	Maximum category	Pathway system
Hydrophobicity	3	3	15	21	21	42
Normalized van der Waals volume	3	3	15	21	21	42
Polarity	3	3	15	21	21	42
Polarizability	3	3	15	21	21	42
Secondary structure	3	3	15	21	21	42
Solvent accessibility	1	1	5	7	7	14
Amino acid composition	20	N/A	N/A	20	20	40
Total	36	36	80	132	132	**264**

doi:10.1371/journal.pone.0025297.t002

3. Solvent accessibility: ACCpro [67] can be used to predict each amino acid as hidden (H) or exposed (E) to solvent. Then the protein sequence is coded with letters H and E. Use composition (C) for H, transition (T) between H and E, and five distributions (D) for H in this property. Thus we have $(1+1+5)\times2=14$ features by considering the "solvent accessibility" property.
4. Amino acid compositions: it contains 20 components with each representing the percentage of each amino acid in a protein sequence [68]. Thus, we have 20 features for the "mean" category, and 20 features for the "maximum" category. Totally, we have $20\times2=40$ features for a pathway system by considering the amino acid composition.

Shown in **Table 2** is a breakdown of the 264 features for a pathway system by considering its biochemical and physicochemical properties. Before taking the mean and maximum values of each property into account, the following equations were used to adjust them according to a standard scale [61]:

$$\begin{cases} U_{ij}=(u_{ij}-u_j)/T_j \\ T_j=\sqrt{\sum_{i=1}^{N}(u_{ij}-u_j)/(N-1)} \\ u_j=\sum_{i=1}^{N}u_{ij}/N \end{cases} \quad (5)$$

where T_j is the standard deviation of the j-th feature and u_j the mean value of the j-th feature.

4. Functional property. The last category of features is about the functional property of each regulatory pathway. The gene ontology enrichment score of pathway i on gene ontology item j was defined as the $-\log_{10}$ of the hypergeometric test p value [15,69,70,71] of proteins in pathway i and can be computed by the following equation:

Table 3. A breakdown of the of 5570 features.

Categories	Group name	Number of features
Graph property	Graph size and graph density	2
	Degree statistics	8
	Edge weight statistics	4
	Topological change	7
	Degree correlation	6
	Clustering	6
	Topological	12
	Singular values	3
	Local density change	40
Biochemical and physicochemical property	Amino acid compositions	40
	Hydrophobicity, normalized van der Waals volume, polarity and polarizability	168
	Solvent accessibility	14
	Secondary structure	42
Functional property	Gene ontology enrichment score	5218
Total	N/A	5570

doi:10.1371/journal.pone.0025297.t003

$$\text{Score}_i^j = -\log_{10}(\text{p-value}) = -\log_{10}\left[\sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right] \quad (6)$$

where N is the number of overall proteins in KEGG of human, M is the number of proteins annotated to gene ontology item j, n is the number of proteins in pathway i, m is the number of proteins in pathway i that are annotated to gene ontology item j. The larger the enrichment score of one gene ontology item, the more overrepresented this item is. There were a total of 5,218 gene ontology (GO) enrichment score features.

5. Representation of each pathway. It follows from the description in Section 1 "Graph property", 3 "Biochemical and physicochemical property" and 4 "Functional property" that the total number of features was $(88+264+5218)=5570$, as summarized **Table 3**. Thus, according to Eq.2, each of the 146 pathway samples in the benchmark dataset (Table S1) will be represented by a 5570-D vector.

mRMR method

Minimum Redundancy Maximum Relevance (mRMR), first proposed by Peng *et al.* [16], was employed in this study, as it is established according to two excellent criteria: Max-Relevance and Min-Redundancy. Max-Relevance guarantees that features giving most contribution to the classification will be selected, while Min-Redundancy guarantees that features whose classification ability has already been covered by selected features will be excluded. By mRMR program, we can obtain two feature lists: MaxRel features list and mRMR features list. MaxRel features list sort features only according to the Max-Relevance criteria, while mRMR features list is obtained in terms of both Max-Relevance and Min-Redundancy. Thus, for a feature set $\mathbf{\Omega}$ with N features, mRMR program will execute N rounds and a feature with maximum relevance and minimum redundancy is selected in each round. Finally, we can obtain an ordered feature list, i.e., mRMR features list:

$$F = [f_0, f_1, \ldots, f_{N-1}] \quad (7)$$

For detail description of the mRMR method, please refer to Peng *et al.*'s paper [16]. Now, mRMR method has been widely utilized to tackle various biological problems [45,52,72,73,74,75,76] and deemed as a powerful and useful tool to extract important information in complex systems. The mRMR program developed by Peng et al [16] is available at http://penglab.janelia.org/proj/mRMR/.

Table 4. The distribution of the most relevant 55 features.

Category	Number of features
Graph property	0
Biochemical and physicochemical property	32
Functional property	23
Total	55

doi:10.1371/journal.pone.0025297.t004

Prediction model

In this study, we tried three prediction methods: Nearest Neighbor Algorithm (NNA), Sequential Minimal Optimization (SMO) and Bayesian network (BayesNet). NNA using cosine similarity as "nearness" [15,61,62,71,77] was implemented with in-house script. The NNA program can be downloaded from http://pcal.biosino.org/NNA.html. SMO and BayesNet were implemented in Weka (Waikato Environment for Knowledge Analysis) [78]. Weka, which was developed by the University of

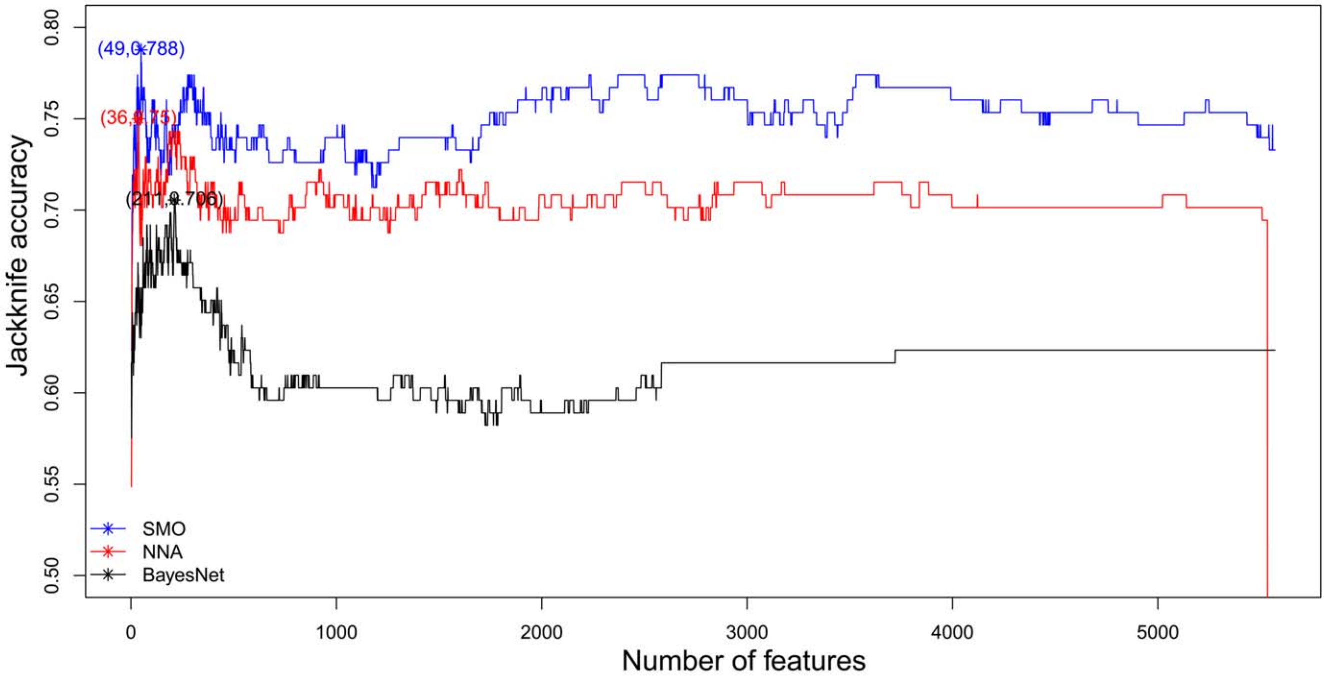

Figure 1. The IFS curve. The highest ACC value of IFS is 78.8% using 49 features and SMO model.
doi:10.1371/journal.pone.0025297.g001

Waikato in New Zealand, is software collecting a variety of state-of-art machine learning algorithms and data preprocessing tools. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning [78]. Weka can be downloaded from http://www.cs.waikato.ac.nz/ml/weka/.

1. Nearest Neighbor Algorithm (NNA). Nearest Neighbor Algorithm (NNA) [17,18], which has been widely used in bioinformatics and computational biology [15,59,60,72,79,80], was adopted to predict the pathway class of each query pathway. The "nearness" is calculated as below

$$D(\mathbb{P}_1,\mathbb{P}_2)=1-\frac{\mathbb{P}_1,\mathbb{P}_2}{\|\mathbb{P}_1\|\cdot\|\mathbb{P}_2\|} \tag{8}$$

where $\mathbb{P}_1$ and $\mathbb{P}_2$ are two vectors representing two pathways, $\mathbb{P}_1\cdot\mathbb{P}_2$ is their dot product, $\|\mathbb{P}_1\|$ and $\|\mathbb{P}_2\|$ are the modulus of vector $\mathbb{P}_1$ and $\mathbb{P}_2$. The smaller the $D(\mathbb{P}_1,\mathbb{P}_2)$, the more similar the two pathways are [55]. In NNA, suppose there are m training pathways, each of them belongs to exact one pathway class, and a query pathway needs to be classified into one pathway class. The distances between each of the m training pathways and the query pathway can be calculated, and the nearest neighbor of the query pathway is found. If the nearest neighbor belongs to the i-th pathway class, the query pathway is classified into the i-th pathway class. For an intuitive illustration of how NNA works, see Fig.5 of [23].

2. SMO. SMO implements John Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial or Gaussian kernels [19,20]. All attributes are processed before using SMO to make prediction, for example nominal attributes are transformed into binary ones, and attributes are normalized [78].

3. BayesNet. BayesNet learns Bayesian networks under the assumptions that all attributes should be nominal (In particular, numeric ones should be prediscretized) and there are no missing values. Two different algorithms are used to estimate the conditional probability tables of the network [78] and several search algorithms are implemented for local score metrics, such as K2 [81], Hill Climbing [82], TAN [83,84] and so on. For more detailed description of this classifiers in Weka can be found in [21].

Jackknife test to examine the quality of the current prediction method

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [85]. However, of the three test methods, the jackknife test is deemed the most objective [56]. The reasons are as follows. **(i)** For the independent dataset test, although all the proteins used to test the predictor are outside the training dataset used to train it so as to exclude the "memory" effect or bias, the way of how to select the independent proteins to test the predictor could be quite arbitrary unless the number of independent proteins is sufficiently large. This kind of arbitrariness might result in completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent testing dataset might fail to keep so when tested by another independent testing dataset [85]. **(ii)** For the subsampling test, the concrete procedure usually used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset, as demonstrated by Eqs.28–30 in [23].

Table 5. The 49 optimized features.

Order	Featurename
1	secondary_structure_composition_P_max
2	solvent_accessibility_composition_H_mean
3	solvent_accessibility_distribution_H.0.75_max
4	GO:0043627 response to estrogen stimulus
5	GO:0045121 membrane raft
6	secondary_structure_distribution_H.0.25_max
7	AA_composition_S_mean
8	secondary_structure_distribution_N.0.25_max
9	VanDerWaal_composition_P_max
10	GO:0043330 response to exogenous dsRNA
11	VanDerWaal_distribution_H.0.75_max
12	AA_composition_T_max
13	AA_composition_D_max
14	secondary_structure_distribution_H.0.5_max
15	GO:0048519 negative regulation of biological process
16	GO:0002687 positive regulation of leukocyte migration
17	secondary_structure_composition_P_mean
18	polarity_composition_N_max
19	GO:0042088 T-helper 1 type immune response
20	polarity_transition_NH_max
21	AA_composition_S_max
22	GO:0042063 gliogenesis
23	polarizability_distribution_P.0.75_max
24	GO:0090068 positive regulation of cell cycle process
25	GO:0014829 vascular smooth muscle contraction
26	secondary_structure_distribution_H.0.75_max
27	AA_composition_Q_mean
28	GO:0030225 macrophage differentiation
29	GO:0046661 male sex differentiation
30	hydrophobicity_composition_N_max
31	solvent_accessibility_distribution_H.0.0_max
32	polarity_distribution_P.0.5_max
33	polarizability_distribution_H.0.75_max
34	GO:0031594 neuromuscular junction
35	GO:0031330 negative regulation of cellular catabolic process
36	AA_composition_P_max
37	GO:0042953 lipoprotein transport
38	GO:0048523 negative regulation of cellular process
39	GO:0030217 T cell differentiation
40	GO:0007517 muscle organ development
41	GO:0009913 epidermal cell differentiation
42	GO:0042177 negative regulation of protein catabolic process
43	GO:0048641 regulation of skeletal muscle tissue development
44	hydrophobicity_distribution_N.0.75_max
45	hydrophobicity_distribution_H.0.75_max
46	GO:0022408 negative regulation of cell-cell adhesion
47	GO:0048608 reproductive structure development
48	GO:0045638 negative regulation of myeloid cell differentiation
49	GO:0006897 endocytosis

doi:10.1371/journal.pone.0025297.t005

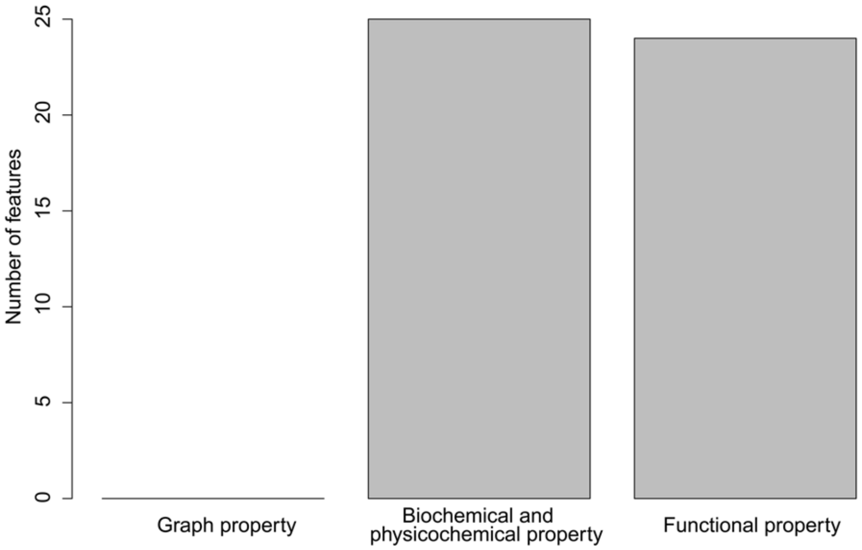

Figure 2. Distribution of the optimized 49 features. It is straightforward to see that 25 (25/49, 51.0%) features were from the biochemical and physicochemical property and 24 (24/49, 49.0%) features were from the functional property, while none of features in graph property was selected into the optimized feature set.
doi:10.1371/journal.pone.0025297.g002

Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as a good one. **(iii)** In the jackknife test, all the proteins in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining protein samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each protein sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. Accordingly, the jackknife test has been increasingly and widely used by those investigators with strong math background to examine the quality of various predictors (see, e.g., [25,26,27,28,29,30,31,32,33,34,86,87,88,89,90]). In view of this, here the jackknife test was also used to examine the quality of the current predictor in identifying the pathway class.

Incremental feature selection (IFS)

As described in Section "mRMR method", mRMR features list $F = [f_0, f_1, \ldots, f_{N-1}]$ can be obtained by mRMR program. Denote the i-th feature set by $F_i = \{ f_0, f_1, \ldots, f_i\}$ $(0 \le i \le N-1)$. For each i $(0 \le i \le N-1)$, execute NNA, SMO and BayesNet with the features in F_i, then the overall accuracy of the classification (ACC), defined by "the number of correctly predicted pathways"/"the total number of pathways", evaluated by jackknife test, was obtained. As a result, we can plot a curve named IFS curve with ACC as its y-axis and the index i of F_i as its x-axis.

Results and Discussion

Results of mRMR

The mRMR program was achieved from http://penglab.janelia.org/proj/mRMR. It was run with default parameters and two feature lists were obtained by executing mRMR program: (i) MaxRel features list; (ii) mRMR features list (see Table S2).

MaxRel features list was obtained by sorting features according to their contribution to the classification. We investigated the most relevant 1% of the features (totally 55) and **Table 4** shows the distribution of these features. It is clear that 32 (32/55, 58.18%) features come from biochemical and physicochemical property and 23 (23/55, 41.82%) features come from functional property. All of these indicate that among the adopted features the biochemical and physicochemical property of each pathway provide the most contribution to classification and functional property also gives important contribution. It is startling that none of the features about graph property was the most relevant 1% feature, while they were considered as important factors to form some biological meaningful systems, such as protein complex [45,53]. In this study, we only take care of classifying a regulatory pathway into correct pathway class but not to analyze which feature is more important to form a regulatory pathway. In this stage, graph property may be not very important while biological and functional properties are more important to determine the biological function of each pathway.

Results of IFS

Shown in **Figure 1** are the IFS curves of NNA, SMO and BayesNet. The highest ACC value of IFS is 78.8% using 49 features and SMO models (See **Table 5** for the detail 49 features). The detailed IFS data can be found in Table S3.

Figure 2 shows the distribution of the optimized 49 features. It is straightforward to see that 25 (25/49, 51.0%) features were from the biochemical and physicochemical property and 24 (24/49, 49.0%) features were from the functional property, while none of features in graph property was selected into the optimized feature set. All of these indicate the same conclusion as described in Section "Results of mRMR".

Analysis of optimal features for pathway classification

It was seen from **Table 5** and **Figure 2** that the biochemical and physicochemical properties and Gene Ontology functional properties were important for pathway classification.

Table 6. Hypergeometric test of overlap between KEGG pathway classes and GO terms in optimized features.

	Metabolism	Genetic Information Processing	Environmental Information Processing	Cellular Processes	Organismal Systems	Human Diseases
GO:0043627 response to estrogen stimulus	0.032588	1	5.15E-16	1.86E-08	0.004826	2.30E-19
GO:0045121 membrane raft	0.681728	0.018851	2.68E-13	7.52E-15	1.09E-22	8.64E-15
GO:0043330 response to exogenous dsRNA	1	1	0.106165	0.003522	0.000117	0.001727
GO:0048519 negative regulation of biological process	1	1	1.86E-59	8.01E-39	4.20E-12	1.90E-51
GO:0002687 positive regulation of leukocyte migration	1	1	2.11E-09	0.001789	0.013702	0.000707
GO:0042088 T-helper 1 type immune response	1	1	3.50E-06	0.471266	0.094723	0.001178
GO:0042063 gliogenesis	0.993714	1	5.20E-11	1.30E-05	0.019525	1.32E-13
GO:0090068 positive regulation of cell cycle process	0.911776	1	9.12E-08	3.49E-06	0.024096	3.29E-08
GO:0014829 vascular smooth muscle contraction	1	1	0.000189	0.049965	0.023416	0.002415
GO:0030225 macrophage differentiation	1	1	0.003204	0.022913	0.00372	0.001178
GO:0046661 male sex differentiation	0.664515	1	4.00E-10	0.036323	0.938207	3.85E-07
GO:0031594 neuromuscular junction	1	1	0.001106	4.49E-06	1.97E-05	0.00224
GO:0031330 negative regulation of cellular catabolic process	1	1	0.006858	0.527536	0.137844	0.00224
GO:0042953 lipoprotein transport	1	1	0.127363	0.312566	0.023416	0.031663
GO:0048523 negative regulation of cellular process	0.999997	1	1.89E-56	1.93E-38	1.57E-08	4.91E-50
GO:0030217 T cell differentiation	0.957773	1	1.26E-16	0.023685	0.000397	1.82E-10
GO:0007517 muscle organ development	0.998366	1	6.32E-12	6.49E-09	0.32379	2.38E-09
GO:0009913 epidermal cell differentiation	1	1	0.123185	0.55964	0.968491	0.395449
GO:0042177 negative regulation of protein catabolic process	1	1	0.019214	0.002942	0.021538	0.001178
GO:0048641 regulation of skeletal muscle tissue development	1	1	5.03E-05	0.001284	0.447341	2.50E-06
GO:0022408 negative regulation of cell-cell adhesion	1	1	0.015685	0.040951	0.017213	0.001727
GO:0048608 reproductive structure development	0.431739	1	2.90E-16	0.036125	0.271969	4.81E-12
GO:0045638 negative regulation of myeloid cell differentiation	1	1	0.032936	0.289118	0.009817	1.09E-06
GO:0006897 endocytosis	0.995474	1	0.000121	0.012134	0.09916	0.006247

doi:10.1371/journal.pone.0025297.t006

Within the selected 25 biochemical and physicochemical properties, there were 6 secondary structure features, 6 amino acid composition features, 3 solvent accessibility features, 3 polarity features, 3 hydrophobicity features, 2 vanderWaal features and 2 polarizability features. Obviously, secondary structure features and amino acid composition features were more important than other biochemical and physicochemical properties. The correct secondary structure of protein is essential to its function. Structural incorrect proteins are associated with many different kinds of disease such as Alzheimer's disease, Huntington's and Parkinson's disease [91]. In KEGG pathway classification, there are 28 disease pathways. Some of the disease pathways, such as neurodegenerative disease pathways and cancer pathways, are caused by or associated with protein misfolding [91]. Amino acid composition has been used to explain a lot of biological phenomenon, such as translation rate [62] and metabolic stability of proteins [61]. Amino acid composition has a close relationship with protein synthesis and degradation [62,70]. In KEGG pathway classification, there are 73 metabolism pathways. The amino acid composition features may affect these metabolism pathways.

To investigate the association between KEGG pathway classes and GO terms in optimized features, we calculated their hypergeometric test p values which were shown in **Table 6**. As shown from the table, "Metabolism" pathways were associated with GO term "GO:0043627 response to estrogen stimulus", "Genetic Information Processing" pathways were associated with GO term "GO:0045121 membrane raft", "Environmental Information Processing" pathways, "Cellular Processes" pathways, "Organismal Systems" pathways and "Human Diseases" pathways were associated with many GO terms in optimized features. Some associations are obvious and well-known, such as the association between "Environmental Information Processing" pathways and GO term "GO:0043627 response to estrogen stimulus", the association between "Cellular Processes" pathways and GO terms "GO:0048519 negative regulation of biological process" and "GO:0048523 negative regulation of cellular process", the association between "Organismal Systems" pathways and GO terms "GO:0030217 T cell differentiation", "GO:0030225 macrophage differentiation" etc., the association between "Human Diseases" pathways and GO terms "GO:0048519 negative regulation of biological process", "GO:0048523 negative regulation of cellular process" and "GO:0042063 gliogenesis". The relationship between "Metabolism" pathways and GO term "GO:0043627 response to estrogen stimulus" may be indirect. Estrogen can introduce dramatic changes of cell, such as apoptosis and carcinogenesis [92,93]. During these cellular changes, the metabolism pathways will change as well. "Genetic Information Processing" pathways include many biological processes, such as transcription, translation, folding, sorting, degradation, replication and repair. All these steps require translocation of big molecular which needs the assistant of membrane systems. Membrane raft involves in biosynthetic traffic, endocytosis and signal transduction [94].

Combining the 25 biochemical and physicochemical properties and 24 Gene Ontology functional properties together, most KEGG pathways can correctly classified with reasonable biological meanings. The prediction model can be used to classify new pathway into existing pathway function groups. This means predicting the function of new pathways which is one of the ultimate goals of biology research.

We have analyzed 5570 features extracted from each of known regulatory pathway in humans. Of the 5570 features, 88 were derived from the graph property, 264 from the biochemical and physicochemical property of proteins, and 5218 from the functional property. Subsequently, the mRMR method and IFS techniques were employed to analyze and identify the the important features. Nearest neighbor algorithm and jackknife test were utilized to evaluate the accuracy of the classifier. As a result, 49 features were found to be as the important features for classifying the pathway groups according to their biological functions. These findings might provide useful insights, stimulating in-depth investigation into such an important and challenging problem.

Supporting Information

Table S1 The pathway benchmark dataset. It contains 146 pathways classified into six classes or groups according their biological functions.
(XLS)

Table S2 Two lists obtained by mRMR program.
(PDF)

Table S3 The IFS results for NNA, SMO and BayesNet.
(XLS)

Acknowledgments

The authors would like to take this opportunity to express our gratitude to the editor and the anonymous reviewer for their constructive comments, which were very helpful in strengthening the presentation of this study.

Author Contributions

Conceived and designed the experiments: Y-DC. Performed the experiments: TH LC. Analyzed the data: TH. Contributed reagents/materials/analysis tools: LC. Wrote the paper: TH LC KC-C.

References

1. Kanehisa M (1997) A database for post-genome analysis. Trends in genetics: TIG 13: 375–376.
2. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27–30.
3. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 27: 29–34.
4. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic acids research 32: D277–D280.
5. Bairoch A (1994) The ENZYME data bank. Nucleic acids research 22: 3626–3627.
6. Schomburg I, Chang A, Hofmann O, Ebeling C, Ehrentreich F, et al. (2002) BRENDA: a resource for enzyme data and metabolic information. Trends in biochemical sciences 27: 54–56.
7. Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic acids research 30: 47–49.
8. Krieger C, Zhang P, Mueller L, Wang A, Paley S, et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic acids research 32: D438–D442.
9. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480–484.
10. Klukas C, Schreiber F (2007) Dynamic exploration and editing of KEGG pathway diagrams. Bioinformatics 23: 344–350.
11. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, et al. (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 34: D511–516.
12. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36: D623–631.
13. Pharkya P, Nikolaev EV, Maranas CD (2003) Review of the BRENDA Database. Metab Eng 5: 71–73.
14. Dale JM, Popescu L, Karp PD (2010) Machine learning methods for metabolic pathway prediction. BMC Bioinformatics 11: 15.
15. Chen L, Huang T, Shi XH, Cai YD, Chou KC (2010) Analysis of protein pathway networks using hybrid properties. Molecules 15: 8177–8192.
16. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence. pp 1226–1238.
17. Salzberg S, Cost S (1992) Predicting protein secondary structure with a nearest-neighbor algorithm* 1. Journal of molecular biology 227: 371–374.
18. Denoeux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems Man and Cybernetics 25: 804–813.
19. Platt J, ed. Fast training of support vector machines using sequential minimal optimization. Cambridge, MA: MIT Press.
20. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13: 637–649.
21. Bouckaert RR (2004) Bayesian network classifiers in Weka. Department of Computer Science, University of Waikato, New Zealand.
22. Chou KC, Zhang CT (1995) Critical Reviews in Biochemistry and Molecular. Biology 30: 275–349.
23. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273: 236–247.
24. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol44, 60) 43: 246–255.
25. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
26. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263: 203–209.
27. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, et al. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. Journal of Theoretical Biology 259: 366–372.
28. Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein & Peptide Letters 16: 27–31.

29. Ding H, Luo L, Lin H (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein & Peptide Letters 16: 351–355.
30. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. Journal of Theoretical Biology 257: 17–26.
31. Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology 281: 18–23.
32. Yu L, Guo Y, Li Y, Li G, Li M, et al. (2010) SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. Journal of Theoretical Biology 267: 1–6.
33. Gu Q, Ding YS, Zhang TL (2010) Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. Protein & Peptide Letters 17: 559–567.
34. Qiu JD, Huang JH, Shi SP, Liang RP (2010) Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein & Peptide Letters 17: 715–722.
35. Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6: 262–274.
36. Chou K (1980) A new schematic method in enzyme kinetics. European Journal of Biochemistry 113: 195–198.
37. Zhou GP, Deng MH (1984) An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochemical Journal 222: 169–176.
38. Chou KC (1989) Graphic rules in steady and non-steady enzyme kinetics. Journal of Biological Chemistry 264: 12074–12079.
39. Chou K (1990) Review: Applications of graph theory to enzyme kinetics and protein folding kinetics: Steady and non-steady-state systems. Biophysical chemistry 35: 1–24.
40. Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Canadian Journal of Chemistry 86: 342–357.
41. Chou K (2010) Graphic rule for drug metabolism systems. Current Drug Metabolism 11: 369–378.
42. Althaus I, Chou J, Gonzales A, Deibel M, Chou K, et al. (1993) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. Journal of Biological Chemistry 268: 6119–6124.
43. Althaus I, Gonzales A, Chou J, Romero D, Deibel M, et al. (1993) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. Journal of Biological Chemistry 268: 14875–14880.
44. Althaus I, Chou J, Gonzales A, Deibel M, Chou K, et al. (1993) Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32: 6548–6554.
45. Chen C, Chen L, Zou X, Cai P (2009) Prediction of Protein Secondary Structure Content by Using the Concept of Chous Pseudo Amino Acid Composition and Support Vector Machine. Protein and Peptide Letters 16: 27–31.
46. Chou KC, Zhang CT, Maggiora GM (1997) Disposition of amphiphilic helices in heteropolar environments. PROTEINS: Structure, Function, and Genetics 28: 99–108.
47. Zhou GP (2011) The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. Journal of Theoretical Biology 284: 142–148.
48. Wu ZC, Xiao X, Chou KC (2010) 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol 267: 29–34.
49. Chakrabarti D (2005) Tools for large graph mining Carnegie Mellon University.
50. Barabasi A, Oltvai Z (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5: 101–113.
51. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck F, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968.
52. Chen L, Lu L, Feng KR, Li WJ, Song J, et al. (2009) Multiple Classifier Integration for the Prediction of Protein Structural Classes. Journal of Computational Chemistry 30: 2248–2254.
53. Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z (2008) Protein complex identification by supervised graph local clustering. Bioinformatics 24: i250–i258.
54. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, et al. (2003) The gene ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Research 13: 662–672.
55. Chou K, Shen H (2007) Recent progress in protein subcellular location prediction. Analytical Biochemistry 370: 1–16.
56. Chou KC, Shen HB (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science, 2010, 2, 1090–1103). Nature Protocols 3: 153–162.
57. Chou KC, Wu ZC, Xiao X (2011) iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. PLoS One 6: e18258.
58. Chou K, Cai Y (2006) Predicting Protein-Protein interactions from sequences in a hybridization space. J Proteome Res 5: 316–322.
59. Chen L, Shi XH, Kong XY, Zeng ZB, Cai YD (2009) Identifying Protein Complexes Using Hybrid Properties. Journal of Proteome Research 8: 5212–5218.
60. Chen L, Feng KY, Cai YD, Chou KC, Li HP (2010) Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. BMC bioinformatics 11: 293.
61. Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 5: e10972.
62. Huang T, Wan S, Xu Z, Zheng Y, Feng KY, et al. (2011) Analysis and prediction of translation rate based on sequence and functional features of the mRNA. PLoS ONE 6: e16036.
63. Dubchak I, Muchnik I, Holbrook S, Kim S (1995) Prediction of protein folding class using global description of amino acid sequence. Proceedings of the National Academy of Sciences of the United States of America 92: 8700–8704.
64. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim S (1999) Recognition of a protein fold in the context of the SCOP classification. Proteins: Structure, Function, and Bioinformatics 35: 401–407.
65. Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics 27: 329–335.
66. Cheng J, Randall A, Sweredoski M, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic acids research 33: W72–W76.
67. Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins: Structure, Function, and Bioinformatics 47: 142–153.
68. Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function & Genetics 21: 319–344.
69. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 8: R3.
70. Huang T, Wang P, Ye ZQ, Xu H, He Z, et al. (2010) Prediction of Deleterious Non-Synonymous SNPs Based on Protein Interaction Network and Hybrid Properties. PLoS ONE 5: e11900.
71. Huang T, Xu Z, Chen L, Cai YD, Kong X (2011) Computational Analysis of HIV-1 Resistance Based on Gene Expression Profiles and the Virus-Host Interaction Network. PLoS ONE 6: e17291.
72. He Z, Zhang J, Shi X, Hu L, Kong X, et al. (2010) Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features. PLoS ONE 5: e9603.
73. Cai Y, Lu L (2008) Predicting n-terminal acetylation based on feature selection method. Biochemical and biophysical research communications 372: 862–865.
74. Cai Y, Lu L, Chen L, He J (2010) Predicting subcellular location of proteins using integrated-algorithm method. Molecular Diversity 14: 551–558.
75. Lu L, Niu B, Zhao J, Liu L, Lu W, et al. (2009) GalNAc-transferase specificity prediction based on feature selection method. Peptides 30: 359–364.
76. Lu L, Shi X, Li S, Xie Z, Feng Y, et al. (2010) Protein sumoylation sites prediction based on two-stage feature selection. Molecular Diversity 14: 81–86.
77. Huang T, Cui W, Hu L, Feng K, Li YX, et al. (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS ONE 4: e8126.
78. Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Pub.
79. Chen L, Qian ZL, Fen KY, Cai YD (2010) Prediction of Interactiveness Between Small Molecules and Enzymes by Combining Gene Ontology and Compound Similarity. Journal of Computational Chemistry 31: 1766–1776.
80. Cai Y, Chou K (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochemical and biophysical research communications 305: 407–411.
81. Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Machine learning 9: 309–347.
82. Buntine W (1996) A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering 8: 195–210.
83. Cheng J, Greiner R. Comparing Bayesian network classifiers; 1999. Proceedings UAI, 101–107.
84. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Machine learning 29: 131–163.
85. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30: 275–349.
86. Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. Journal of Theoretical Biology 252: 350–356.
87. Xiao X, Wu ZC, Chou KC (2011) A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 6: e20592.
88. Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. Journal of Theoretical Biology 253: 310–315.

89. Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology 248: 546–551.
90. Gao CF, Qiu ZX, Wu XJ, Tian FW, Zhang H, et al. (2011) A Novel Fuzzy Fisher Classifier for Signal Peptide Prediction. Protein Peptide Letters 18: 831–838.
91. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75: 333–366.
92. Lobanova YS, Scherbakov AM, Shatskaya VA, Krasil'nikov MA (2007) Mechanism of estrogen-induced apoptosis in breast cancer cells: role of the NF-kappaB signaling pathway. Biochemistry (Mosc) 72: 320–327.
93. Chang M (2011) Dual roles of estrogen metabolism in mammary carcinogenesis. BMB Rep 44: 423–434.
94. Chazal N, Gerlier D (2003) Virus entry, assembly, budding, and membrane rafts. Microbiol Mol Biol Rev 67: 226–237, table of contents.