Molecules **2010**, *15*, 8177-8192; doi:10.3390/molecules15118177

OPEN ACCESS

molecules

ISSN 1420-3049

www.mdpi.com/journal/molecules

Article

Analysis of Protein Pathway Networks Using Hybrid Properties

Lei Chen [1,2], Tao Huang [3,4], Xiao-He Shi [5], Yu-Dong Cai [6,7,]* and Kuo-Chen Chou [7]

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; E-Mail: chen_lei1@163.com (L.C.)

[2] Centre for Computational Systems Biology, Fudan University, Shanghai 200433, China

[3] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; E-Mail: tohuangtao@126.com (T.H.)

[4] Shanghai Center for Bioinformation Technology, Shanghai 200235, China

[5] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China; E-Mail: xiaoheshi@163.com (X-H.S.)

[6] Institute of Systems Biology, Shanghai University, Shanghai 200444, China

[7] Gordon Life Science Institute, San Diego, California 92130, USA; E-Mail: kcchou@gordonlifescience.org (K-C.C.)

* To whom correspondence should be addressed; E-Mail: caiyudong@staff.shu.edu.cn;
Tel.: 011-86-216613-6132; Fax: 011-86-216613-6133.

Received: 5 August 2010; in revised form: 11 November 2010 / Accepted: 12 November 2010 / Published: 12 November 2010

Abstract: Given a protein-forming system, *i.e.*, a system consisting of certain number of different proteins, can it form a biologically meaningful pathway? This is a fundamental problem in systems biology and proteomics. During the past decade, a vast amount of information on different organisms, at both the genetic and metabolic levels, has been accumulated and systematically stored in various specific databases, such as KEGG, ENZYME, BRENDA, EcoCyc and MetaCyc. These data have made it feasible to address such an essential problem. In this paper, we have analyzed known regulatory pathways in humans by extracting different (biological and graphic) features from each of the 17,069 protein-formed systems, of which 169 are positive pathways, *i.e.*, known regulatory pathways taken from KEGG; while 16,900 were negative, *i.e.*, not formed as a biologically meaningful pathway. Each of these protein-forming systems was represented by 352 features, of which 88 are graph features and 264 biological features. To analyze these features, the "Minimum Redundancy Maximum Relevance" and the "Incremental Feature

Selection" techniques were utilized to select a set of 22 optimal features to query whether a protein-forming system is able to form a biologically meaningful pathway or not. It was found through cross-validation that the overall success rate thus obtained in identifying the positive pathways was 79.88%. It is anticipated that, this novel approach and encouraging result, although preliminary yet, may stimulate extensive investigations into this important topic.

Keywords: protein-forming system; regulatory pathway; minimum redundancy maximum relevance; gene ontology; biological graphic feature

1. Introduction

During the past decade, the continuous development of high-throughput experimental technologies has increased the sizes of large-scale datasets, including both metagenomes and personal genomes, which necessitate renewed efforts to develop computational technologies for better biological interpretation of all this data. A vast amount of information about different organisms, both on the genetic and metabolic levels, has been accumulated and systematically stored in specific databases that are available on various websites including KEGG [1,2], ENZYME [3], BRENDA [4,5], and EcoCyc and MetaCyc [6].

KEGG (Kyoto Encyclopedia of Genes and Genomes) [1,2,7] is a widely used knowledge database for the systematic analysis of gene functions in terms of the interactions between genes and molecules; it consists of graphical diagrams of biochemical pathways, including most of the known metabolic pathways and some of the known regulatory pathways. Nowadays, KEGG PATHWAY is supplemented with a new global map of metabolic pathways, which is essentially a combined map of about 120 existing pathway maps. KEGG BRITE is an ontology database, which represents functional hierarchies of various biological objects, including molecules, cells, organisms, diseases and drugs, as well as relationships among them [8,9]. In these databases, experimental knowledge is organized and diagramed as smaller networks, and web interfaces and visualization tools have been developed to overview and analyze computationally generated global networks [10-12].

Many studies from various research laboratories around the world have indicated that mathematical analysis, computational modeling, and the introduction of novel physical concepts to solve important problems in biology and medicine, such as protein structural class prediction [13,14], modeling of 3D structures of targeted proteins for drug design [15-18], diffusion-controlled reaction simulation [19-22], cellular responding kinetics [23,24], bio-macromolecular internal collective motion simulation [25-27], identification of proteases and their types [28,29], membrane protein type prediction [30,31], protein cleavage site prediction [32,33], and signal peptide prediction [34,35], can provide very useful and timely information and insights for both basic research and drug development. Encouraged by these promising outcomes, the present study was initiated to address a fundamental problem in system biology and proteomics.

For most pathways stored in the KEGG server, it is barely possible to acquire their graph characteristics by manual query execution. The present study was devoted to the development of a new approach to address this problem that maybe of use for in-depth study of the various pathway network systems.

2. Materials and Methods

2.1. Materials

The data of regulatory pathways was collected from the public available database KEGG (ftp://ftp.genome.jp/pub/kegg/xml). Those pathways without GO information or biological properties were removed. Pathways involving less than three proteins were also excluded. As a result, 169 regulatory pathways, or protein-forming systems, were obtained and they are termed as "positive pathways". The 169 positive pathways as well as the protein codes contained in each of such pathways are given in Online Supporting Information S1.

The negative pathways data was generated by the following two routes: first, proteins were randomly picked as the nodes of a graph, followed by the creation of some arcs between these proteins in a random manner. The number of arcs in each pathway was assigned according to the size distribution of the arcs in the positive pathways. Second, about half of proteins were replaced by other proteins in each positive pathway, and the arcs between the proteins, including both the original and the replaced ones, left unchanged. Since positive pathways are very rare in comparison with the vast majority of negative pathways, in this study the number of negative pathways thus generated was 100 times as big as that of the positive ones. The 16,900 negative pathways thus obtained are given in Online Supporting Information S2.

2.2. Features

The use of graphic approaches to study biological systems can provide useful intuitive insights, as indicated by many previous studies on a variety of important biological topics, such as enzyme-catalyzed reactions [36-40], protein folding kinetics [41], inhibition of HIV-1 reverse transcriptase [42-44], inhibition kinetics of processive nucleic acid polymerases and nucleases [45], and drug metabolism systems [46]. Recently, graphical methods have also been utilized to deal with various biological and medical related problems [47-50].

In this study, both graphic features and biological properties were used to code each pathway. We downloaded the human KGML (KEGG XML) files from KEGG FTP site (ftp://ftp.genome.jp/pub/kegg/xml) and parsed them into graphs using KEGGgraph [51], an interface between KEGG pathway and graph objects in R. The vertices in graphs parsed from KGML files are proteins and the arcs indicate the relations between the protein vertices. Each graph is a directed graph or digraph [39,41], since the relation between two proteins is directional, *i.e.* one protein $\mathbf{P}_1$ can regulate another protein $\mathbf{P}_2$ while $\mathbf{P}_2$ cannot always regulate $\mathbf{P}_1$. In this study, 88 graph features were extracted from each directed graph that represents a pathway, and 264 features of biological properties were derived from biochemical properties and physicochemical properties, including amino acid compositions, hydrophobicity, normalized van der Waals volume, polarity, polarizability, solvent accessibility and secondary structure. Thus, we have a total of (88 + 264) = 352 features altogether. For the codes of the 352 features and how they were used to quantitatively define each of the 169 positive pathways, see

Online Supporting Information S3. Similarly, we can also uniquely define each of the 16,900 negative pathways in a 352-D (dimensional) space as done for the 169 positive pathways. Here, the detailed results for the 16,900 negative pathways are not shown because the corresponding file is too large to be submitted. However, it is available upon request.

Actually, many graph features were derived in [52-54], where the features were extracted from an undirected graph. In this study, every pathway can be deemed as a directed graph, where vertices denote proteins and arcs denote relations. The arcs are weighted by the likelihood that they may interact with each other, as will be further explained in Section 2.3. The 352 features were divided into the following groups.

(1) Graph size and graph density. Suppose the graph of a pathway is formulated by $G = (V, E)$ where V represent the vertices and E the arcs. The size of the graph is the number of proteins in the pathway. Suppose $|E|_{\max} = |V|^2$ is the theoretical maximum number of possible arcs in G. The graph density is defined as $|E|$ divided by $|E|_{\max}$ [52].

(2) Degree statistics. The in-degree (out-degree) of a vertex is defined as the number of in-neighbors (out-neighbors) of the vertex. Considered in this study were the mean in-degree, variance of in-degree, median in-degree, maximum in-degree, mean out-degree, variance of out-degree, median out-degree and maximum out-degree as features [53].

(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]. It is possible when $w(e) = 0$ for some arc $e \in E$; we extracted features in two cases: (a) all arcs in graph were considered including those with zero weights, and that mean and variance of those weights being taken as the features; (b) arcs with non-zero weights were considered so as to take mean and variance of the non-zero weights as features [52].

(4) Topological change. Let $G = (V, w(E))$ be a weighted pathway graph. This group of features was to measure the topological changes when different cutoffs of the weights were 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. Let $G_i = (V, E_i)$ (i = 1,2,3,4,5,6,7,8) be the graph that only includes arcs with weights higher than $i/10$ remained; *i.e.* $E_i=\{e \mid w(e) > i/10\}$. Topology changes are measured as $T_i = (|E_i|-|E_{i+1}|)/|E_i|$ for i = 1,2,3,4,5,6,7 ($T_i = 0$ if $|E_i| = 0$).

(5) Degree correlation. Let $G = (V, E)$ be a pathway graph with $V = \{v_1,v_2,\cdots,v_n\}$. For each vertex v_i, denote its in-neighbors as $V'_{,i} = \{v_{i_1},v_{i_2},\cdots, v_{i_k}\}$ and out-neighbors as $V''_{,i}=\{v_{j_1},v_{j_2},\cdots, v_{j_l}\}$. Let $H'_{,i} = (V'_{,i}, E'_{,i})$ and $H''_{,i} = (V''_{,i}, E''_{,i})$ be two subgraphs of G induced by $V'_{,i}$ and $V''_{,i}$, respectively. Define $D'_{,i} = |E'_{,i}| / k$ ($D'_{,i} = 0$ if $k = 0$) and $D''_{,i} = |E''_{,i}| / l$ ($D''_{,i} = 0$ if $l = 0$). Take the mean, variance and maximum of $D'_{,1}, \cdots, D'_{,n}$ and $D''_{,1}, \cdots, D''_{,n}$, respectively, as features in this group [54].

(6) Clustering. Let $G = (V, E)$ be a pathway graph with $V = \{v_1,v_2,\cdots,v_n\}$. For each vertex v_i, let its in-neighbors be $V'_{,i} = \{v_{i_1},v_{i_2},\cdots, v_{i_k}\}$ and out-neighbors be $V''_{,i} = \{v_{j_1},v_{j_2},\cdots, v_{j_l}\}$. Let $H'_{,i} = (V'_{,i}, E'_{,i})$ and $H''_{,i} = (V''_{,i}, E''_{,i})$ be two subgraphs of G induced by $V'_{,i}$ and $V''_{,i}$, respectively. Define $C'_{,i} = |E'_{,i}| / k^2$ ($C'_{,i} = 0$ if $k = 0$) and $C''_{,i} = |E''_{,i}| / l^2$ ($C''_{,i} = 0$ if $l = 0$). Take the mean, variance and maximum of $C'_{,1}, \cdots, C'_{,n}$ and $C''_{,1}, \cdots, C''_{,n}$, respectively, as features in this group [53].

(7) Topological. Let $G = (V, E)$ be a pathway graph with $V = \{v_1,v_2,\cdots,v_n\}$. For each pair of vertices v_i, v_j $(i \neq j)$, denote $n^1_{,ij}$ as the number of both in-neighbor of v_i and in-neighbor of v_j, $n^2_{,ij}$ as the number of both in-neighbor of v_i and out-neighbor of v_j, $n^3_{,ij}$ as the number of both out-neighbor of v_i and in-neighbor of v_j and $n^4_{,ij}$ as the number of both out-neighbor of v_i and out-neighbor of v_j. For each vertex v_i, denote $n^1_{,i}$ and $n^2_{,i}$ as the number of in-neighbors and out-neighbors of v_i. Let $T^1_{,ij} = n^1_{,ij} / n^1_{,i}$ ($T^1_{,ij}$

$= 0$ if $n^{1}{}_{,i} = 0$), $T^{2}{}_{,ij} = n^{2}{}_{,ij} / n^{1}{}_{,i}$ ($T^{2}{}_{,ij} = 0$ if $n^{1}{}_{,i} = 0$), $T^{3}{}_{,ij} = n^{3}{}_{,ij} / n^{2}{}_{,i}$ ($T^{3}{}_{,ij} = 0$ if $n^{2}{}_{,i} = 0$), and $T^{4}{}_{,ij} = n^{4}{}_{,ij} / n^{2}{}_{,i}$ ($T^{4}{}_{,ij} = 0$ if $n^{2}{}_{,i} = 0$). For each vertex v_i, let $T^{k}{}_{,i}$ be the mean of $T^{k}{}_{,i1},\cdots, T^{k}{}_{,in}$ for $k = 1,2,3,4$. Features in this group are defined as the mean, variance and maximum of $T^{k}{}_{,1},\cdots, T^{k}{}_{,n}$ for $k = 1,2,3,4$ [54].

(8) Singular values. Let $G = (V, E)$ be a pathway graph and A be its adjacent matrix. Take the first three largest singular values as the features [52].

(9) Local density change. Let $G = (V, E)$ be a pathway graph with $V = \{v_1,v_2,\cdots,v_n\}$. This group of features was to measure the similarity of the in-neighbors and out-neighbors of a protein in the pathway. For each vertex v_i, suppose $V'{}_{,i} = \{v_{i_1},v_{i_2},\cdots, v_{i_k}\}$ and $V''{}_{,i} = \{v_{j_1},v_{j_2},\cdots, v_{j_l}\}$ be the in-neighbors and out-neighbors of v_i, respectively. We only show how to gain features from the in-neighbors of each vertex under different cutoffs, which included 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 vertex $v_{i_1},v_{i_2},\cdots, v_{i_k}$ and the weight of each pair of vertices is the likelihood of the corresponding proteins (see Section 2.3). Suppose the cutoff is w, which may be 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 or 0.9. Extract a spanning subgraph $G'{}_{,i}(w)$ of $K'{}_{,i}$ with edges whose weights are greater than w. Compute $L'{}_{,i}(w) = 2|E(G'{}_{,i}(w))|/(k(k\text{-}1))$ ($L'{}_{,i}(w) = 0$ if $k \leq 1$). Take the mean and maximum of $L'{}_{,1}(w), L'{}_{,2}(w),\cdots, L'{}_{,n}(w)$ as features under cutoff w.

The above features are for the pathway graph representation. The following are for the biochemical properties and physicochemical properties, where biochemical properties include amino acid compositions and secondary structure, while physicochemical properties include hydrophobicity, normalized van der Waals volume, polarity, polarizability and solvent accessibility. These properties have been widely applied in the field of computational biology [55-63]. Suppose a pathway consists of n proteins, the mean and maximum values of biological properties of the n proteins are taken as the features.

(10) Hydrophobicity, normalized van der Waals volume, polarity and polarizability: 42 features can be extracted from each of these physicochemical properties [64,65]. Here we will only describe how to obtain features from the hydrophobicity property, as features from other properties can be obtained in a similar way. Each amino acid is assigned into one of the three categories, polar (P), neutral (N) and hydrophobic (H). For a given protein sequence, we use P, N or H to substitute each amino acid in the sequence, and the resulting sequence is called a protein pseudo-sequence. Composition (C) is defined as the percentage of P, N and H in the whole pseudo-sequence. Transition (T) is defined as the changing frequency between any two characters (such as P and N, P and H, N and H). Distribution (D) is defined as the sequence segment (in percentage) of the pseudo-sequence that 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 15 properties for (C), (T) and (D), respectively. Totally $21 \times 2 = 42$ features are obtained.

(11) Solvent accessibility: each amino acid can be predicted by ACCpro [66] 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, resulting in totally $7 \times 2 = 14$ features.

(12) Secondary structure: each amino acid in the protein sequence is substituted by one of three letters like hydrophobicity property. For details, please see [67,68]. $21 \times 2 = 42$ features can be derived from this property.

(13) Amino acid compositions: the percentage of each amino acid in the whole sequence. Totally, 20 × 2 = 40 features about amino acid composition are extracted.

Table 1. Amount of properties in feature group 10–13.

Properties	C	T	D	Total
Hydrophobicity	3	3	15	21
Normalized van der Waals volume	3	3	15	21
Polarity	3	3	15	21
Polarizability	3	3	15	21
Secondary structure	3	3	15	21
Solvent accessibility	1	1	5	7
Amino acid composition	20	---	---	20
Total	---	---	---	132

Table 2. The distribution of 352 features.

Group ID	**Group Name**	**Number of features**
1	Graph size and graph density	2
2	Degree statistic	8
3	Edge weight statistics	4
4	Topological change	7
5	Degree correlation	6
6	Clustering	6
7	Topological	12
8	Singular values	3
9	Local density change	40
10	Hydrophobicity, normalized van der Waals volume, polarity and polarizability	4 × 2 × 21 = 168
11	Solvent accessibility	7 × 2 = 14
12	Secondary structure	2 × 21 = 42
13	Amino acid compositions	2 × 20 = 40

Shown in Table 1 are the numbers of the properties in the above feature group 10–13. Before taking the mean and maximum values of properties in these groups, the following conversion was taken to adjust their values according to a standard scale:

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

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

$$\begin{aligned}\Omega &= 2+8+4+7+6+6+12+3+40+2\times(4\times 21+21+7+20) \\ &= 88+2\times 132 = 352.\end{aligned} \tag{2}$$

As for the detailed distribution of the 352 features, see Table 2.

2.3. Gene ontology

As mentioned above, some features need the arc weight to indicate how likely it is that an interaction may happen between two proteins. In order to generate the edge weight of two interacting proteins, we used gene ontology consortium (GO) [69] to represent each protein. "Ontology" is a specification of a conceptualization and refers to the subject of existence. GO is established by the following three criteria: molecular function, biological process, and cellular component. GO consortium is considered to be a very powerful and helpful vehicle for investigating protein-protein interactions [70], because these three criteria reflect the attribute of gene, gene product, gene-product groups and core features reflecting the subcellular localization [71,72]. The steps of using GO (gene ontology) encoding are described as following:

(1) By using Uniprot2GO mapping provided by GOA Uniprot 34.0 on November 21st 2005 (http://www.ebi.ac.uk/GOA/) [69] which contains 9525 GO items, the functional annotations of proteins provided by GO were obtained.

(2) Each protein can be represented in a 9,525-dimensional vector using each of the 9525 GO items as the vector base, e.g., if a given protein hits a GO item which is the i-th entry of the 9525 GO items, then the i-th component of the 9,525-dimensional vector is set to be 1, otherwise 0.

(3) Thus, each protein sample can be formulated as a 9,525-D vector:

$$\mathbf{P} = \left[p_1\ p_2\ \cdots\ p_i\ \cdots\ p_{9525}\right]^{\mathrm{T}} \tag{3}$$

where $p_i = 1$ if the sample hit the i-th GO item; 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 computed by the following formula:

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

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

2.4. Minimum redundancy maximum relevance (mRMR)

Feature selection can reduce the feature dimensions so as to improve the efficiency of a learning machine. The concrete procedure can be realized by utilizing the mRMR approach, which was first proposed by Peng [73]. This is because it can balance the minimum redundancy and the maximum relevance. The maximum relevance would guarantee selection of those features contributing most to the classification, while the minimum redundancy would guarantee exclusion of those already been covered by the selected features. During the selecting process, one feature at a time was selected by mRMR into the selected list. In each round, a feature with maximum relevance and minimum redundancy was selected. As a result, we obtained a complete list of the selected features with some order. When computing the redundancy and relevance, the mutual information (MI) was adopted, as defined below:

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dx\,dy \tag{5}$$

where *x* and *y* are two random variables; *p*(*x*,*y*) is the joint probabilistic distribution of *x* and *y*; while *p*(*x*) and *p*(*y*) the marginal probabilities of *x* and *y*, respectively.

Let Ω denote the whole feature set. The selected feature set with *m* features is denoted by Ω_s, and the rest of *n* features is denoted by Ω_r. The relevance of a feature *f* and the target variable *h* can be computed as *I*(*f*, *h*), the redundancy between a feature *f* and the selected Ω_s is computed as:

$$r(f,\Omega_s)=\frac{1}{m}\sum_{f_i\in\Omega_s}I(f,f_i)\ \ (r(f,\Omega_s)=0 \text{ if } m=0) \tag{6}$$

For each feature *f* in Ω_r, compute the following equation:

$$R(f,\Omega_s)=I(f,h)-r(f,\Omega_s) \tag{7}$$

To maximize the relevance and minimize redundancy, select a feature $f' \in\Omega_r$ such that:

$$R(f',\Omega_s)=\max_{f\in\Omega_r}R(f,\Omega_s) \tag{8}$$

Then take f' into Ω_s and remove f' from Ω_r. For the rest features, in each round the most relevant and least redundant feature is removed from Ω_r and put into Ω_s, until all features are in Ω_s. Thus, for a feature pool Ω with *N*(*N*=*n*+*m*) features, mRMR program will execute *N* rounds and provide an ordered feature list:

$$F=[f_0 f_1 \ldots f_k \ldots f_{N\text{-}1}] \tag{9}$$

where *k* denotes the round at which the feature is selected.

2.5. Nearest neighbor algorithm

In this study, the NN (nearest neighbor) algorithm [74] was adopted to predict the class of pathway (positive or negative). The "nearness" is defined by the Euclidian distance:

$$d(c_1,c_2)=1-\frac{c_1\cdot c_2}{\|c_1\|\cdot\|c_2\|} \tag{10}$$

where $c_1{\cdot}c_2$ is dot product of two vectors c_1 and c_2, $\|c_1\|$ and $\|c_2\|$ are the modulus of vector c_1 and c_2, respectively. The smaller the $d(c_1, c_2)$, the nearer the two vectors are [75].

In the NN algorithm, suppose there are *m* training pathways, each of them is either positive or negative, and a query protein system needs to be determined as forming either a positive or negative pathway. The distances between each of the *m* pathways and the new pathway are computed, and the nearest neighbor of the new pathway is found. If the nearest neighbor is positive or negative, then the query protein system is assigned to be with positive or negative pathway, respectively.

2.6. Jackknife cross-validation

The prediction model was examined by the jackknife test. In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its accuracy: independent dataset test, subsampling (K-fold cross-validation) test, and the jackknife test [14]. However, as elucidated by [76] and demonstrated by Eq. (50) in [75], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (see, e.g., [77,88]). Accordingly, in this study the jackknife test was adopted to examine the quality of our prediction method as well. During the jackknifing process, each of the statistical samples in the benchmark dataset was in turn singled out as the prediction target and the rest of the samples were used to train the prediction model.

2.7. Incremental feature selection (IFS)

From mRMR, we obtained an ordered feature list $F = [f_0 f_1 \ldots f_k \ldots f_{N\text{-}1}]$. Let $F_i = \{f_0, f_1 \ldots f_i\}$ $(0 \le I \le N\text{-}1)$ be the *i*-th feature set taken from *F*. For every *i* $(0 \le i \le N\text{-}1)$, we executed NN algorithm with the features in F_i and obtained an accuracy of correctly predicting the positive pathways, evaluated by jackknife cross-validation. As a result, a curve named IFS curve, with identification accuracy as its y-axis and the index *i* of F_i as its x-axis, was obtained.

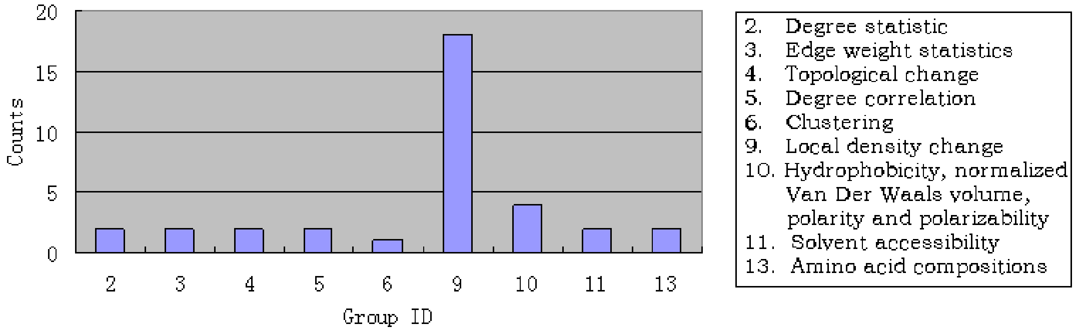
3. Results and Discussion

3.1. Results of mRMR

The mRMR program was downloaded from http://research.janelia.org/peng/proj/mRMR/. It was run with default parameters. The following two feature lists were obtained through the mRMR program: (1) MaxRel features list; (2) mRMR features list (see Online Supporting Information S4).

For the MaxRel feature list, we investigated the most relevant 10% of the features (35 in total). Shown in Figure 1 is the distribution of these features. It is straightforward to see that 27 (77.1%) features come from pathway graph, indicating that among the adopted features, graph features contribute most to the forming of regulatory pathways. Of the 27 features, 18 (51.43%) were from the 9-th feature group, which reflects the essence of the similarity concerned, implying that similar proteins can be regulated by the same protein.

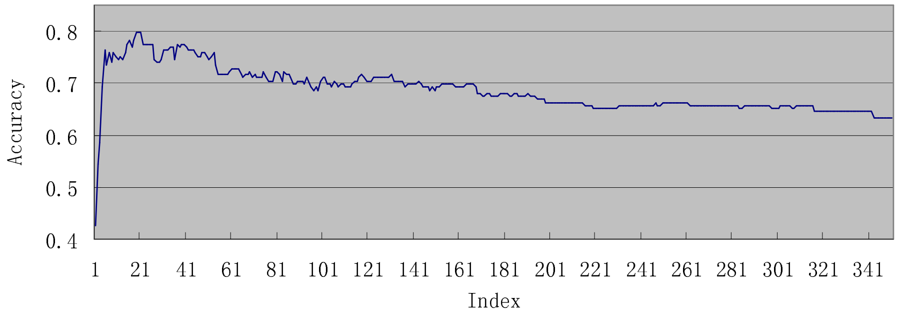
Figure 1. Illustration to show the distribution of features. See the text in Section 3.1 for further explanation.



3.2. Results of IFS

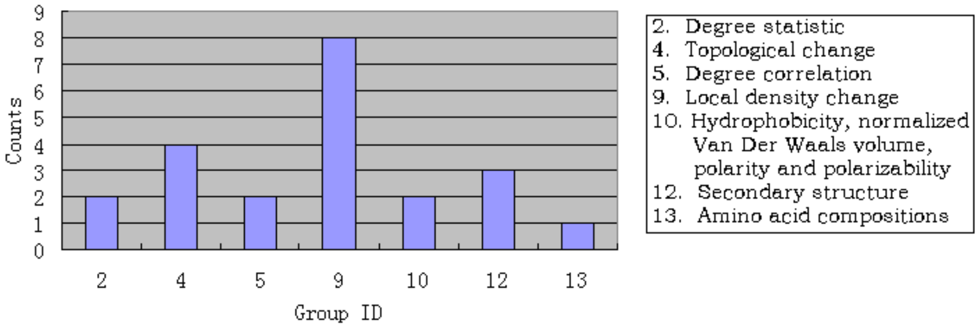
Shown in **Figure 2** is the IFS (incremental feature selection) curve. The highest accuracy of IFS for the positives is 79.88% using 22 features (see Online Supporting Information S4). When using these optimized 22 features, the accuracy of negative pathways and total accuracy were 99.69% and 99.49%, respectively. The detailed IFS data can be found in Online Supporting Information S5.

Figure 2. The IFS (incremental feature selection) curve. See the text in Section 3.2 for further explanation.



Shown in Figure 3 is the distribution of the optimized 22 features. It is again straightforward to see that 16 (72.72%) features were from the pathway graph, among which 8 (36.36%) features were from the 9-th feature group, reaching the same conclusion as that in Section 3.1.

Figure 3. Distribution of the optimized 22 features. See the text in Section 3.2 for further explanation.



3.3. Analysis of the important features

In this work, we present a novel KEGG pathway network analysis method based on hybrid properties, the graph properties and biochemical and physicochemical properties. It was found that the features contributing most in forming pathways were the “out_local_density” and “in_local_density”, both of which were involved with the change of the number of the edges when different weight cutoffs were applied to the graph. Therefore, more edges might remain in the positive graph when higher weight cutoffs were applied. The other graph feature with more contribution to the pathway is the “topological mean”, reflecting various proteins topologies in the regulatory pathway. For a non-broken graph, linear graph (proteins in the graph form a linear path) has a minimum topological mean, while a complete graph has a maximum topological mean. A densely-connected graph always has higher topological mean, indicating a higher likelihood to form a regulatory pathway. The “in_degree_variance”, “out_degree_variance”, and “out_degree_correlation_max” represent the difference of similarity between each of the protein pairs. Most of the forefront features with the dominant contribution are graph features, indicating that graph features are the most important ones. The biochemical and physicochemical properties, including "polarity_composition_P_max", "secondary_structure_distribution_P-1.0_mean", "secondary_structure_distribution_P-1.0_max", "secondary_structure_distribution_P-0.0_max", "polarizability_distribution_N-1.0_max", and "AA_composition_ C_mean" also had considerable contributions in determining the regulatory networks. The distribution of the polarity of proteins structures had strong impact on the conformation of proteins, and hence their interactions as well as their binding sites.

4. Conclusions

We analyzed 352 features extracted from each of the generated positive pathways and negative pathways. Of the 352 features, 88 were graph ones, meaning that each pathway was treated as a graph; and 264 were derived from protein biological properties. The mRMR (minimum redundancy maximum relevance) and IFS (incremental feature selection) techniques were employed to analyze these features. Nearest neighbor algorithm and jackknife test were used to evaluate the accuracy of our model in searching for the positive pathways. As a result, 22 features were found to be the important features for the classification. These findings might be of use for stimulating further studies on such an important and challenging topic.

Acknowledgements

The authors wish to thank the four anonymous reviewers for their constructive comments that were very helpful for strengthening the presentation of this paper. This study was supported by the Grant from Shanghai Commission for Science and Technology (KSCX2-YW-R-112).

References

1 Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375-376.

2. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acid. Res.* **1999**, *27*, 29-34.

3. Bairoch, A. The ENZYME data bank. *Nucl. Acid. Res*. **1994**, *22*, 3626-3627.

4. Schomburg, I.; Chang, A.; Hofmann, O.; Ebeling, C.; Ehrentreich, F.; Schomburg, D. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* **2002**, *27*, 54-56.

5. Schomburg, I.; Chang, A.; Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucl. Acid. Res.*, **2002**, *30*, 47-49.

6. Krieger, C.; Zhang, P.; Mueller, L.; Wang, A.; Paley, S.; Arnaud, M.; Pick, J.; Rhee, S.; Karp, P. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucl. Acid. Res.*, **2004**, *32*, D438-D442.

7. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acid. Res.*, **2000**, *28*, 27-30.

8. Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T. KEGG for linking genomes to life and the environment. *Nucl. Acid. Res.* **2008**, *36*, D480-D484.
9. Klukas, C.; Schreiber, F. Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics* **2007**, *23*, 344-350.
10. Pharkya, P.; Nikolaev, E.; Maranas, C. Review of the BRENDA Database. *Metab. Eng.* **2003**, *5*, 71-73.
11. Caspi, R.; Foerster, H.; Fulcher, C.; Hopkinson, R.; Ingraham, J.; Kaipa, P.; Krummenacker, M.; Paley, S.; Pick, J.; Rhee, S. MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucl. Acid. Res.* **2006**, *34*, D511-D516.
12. Caspi, R.; Foerster, H.; Fulcher, C.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S.; Shearer, A.; Tissier, C. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucl. Acid. Res.* **2008**, *36*, D623-D631.
13. Zhou, G.P.; Assa-Munt, N. Some insights into protein structural class prediction. *Protein. Struct. Funct. Genet.* **2001**, *44*, 57-59.
14. Chou, K.C.; Zhang, C.T. Prediction of protein structural classes. *Crit. Rev. Biochem. Molec. Biol.* **1995**, *30*, 275-349.
15. Zhou, G.P.; Troy, F.A. NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. *Curr. Protein Pept. Sci.* **2005**, *6*, 399-411.
16. Chou, K.C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105-2134.
17. Sharma, A.K.; Zhou, G.P.; Kupferman, J.; Surks, H.K.; Christensen, E.N.; Chou, J.J.; Mendelsohn, M.E.; Rigby, A.C. Probing the interaction between the coiled coil leucine zipper of cGMP-dependent protein kinase Ialpha and the C terminus of the myosin binding subunit of the myosin light chain phosphatase. *J. Biol. Chem.* **2008**, *283*, 32860-32869.
18. Zhou, G.P.; Surks, H.K.; Schnell, J.R.; Chou, J.J.; Mendelsohn, M.E.; Rigby, A.C. The Three-Dimensional Structure of the cGMP-Dependent Protein Kinase I-α Leucine Zipper Domain and Its Interaction with the Myosin Binding Subunit. *Blood* **2004**, *104*, 963a.
19. Zhou, G.Q.; Zhong, W.Z. Diffusion-controlled reactions of enzymes. A comparison between Chou's model and Alberty-Hammes-Eigen's model. *Eur. J. Biochem.* **1982**, *128*, 383-387.
20. Chou, K.C.; Zhou, G.P. Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J. Amer. Chem. Soc.* **1982**, *104*, 1409-1413.
21. Zhou, G.P.; Li, T.T.; Chou, K.C. The flexibility during the juxtaposition of reacting groups and the upper limits of enzyme reactions. *Biophys. Chem.* **1981**, *14*, 277-281.
22. Zhou, G.Z.; Wong, M.T.; Zhou, G.Q. Diffusion-controlled reactions of enzymes. An approximate analytic solution of Chou's model. *Biophys. Chem.* **1983**, *18*, 125-132.
23. Qi, J.P.; Ding, Y.S.; Shao, S.H.; Zeng, X.H.; Chou, K.C. Cellular responding kinetics based on a model of gene regulatory networks under radiotherapy. *Health* **2010**, *2*, 137-146 (openly accessible at http://www.scirp.org/journal/Health/).
24. Qi, J.P.; Shao, S.H.; Li, D.D.; Zhou, G.P. A dynamic model for the p53 stress response networks under ion radiation. *Amino Acids* **2007**, *33*, 75-83.

25. Zhou, G.P. Biological functions of soliton and extra electron motion in DNA structure. *Phys. Scr.* **1989**, *40*, 698-701.
26. Chou, K.C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.* **1988**, *30*, 3-48.
27. Chou, K.C. The biological functions of low-frequency phonons: 6. A possible dynamic mechanism of allosteric transition in antibody molecules. *Biopolymers* **1987**, *26*, 285-295.
28. Zhou, G.P.; Cai, Y.D. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. *Protein. Struct. Funct. Genet.* **2006**, *63*, 681-684.
29. Chou, K.C.; Shen, H.B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Comm.* **2008**, *376*, 321-325.
30. Cai, Y.D.; Zhou, G.P.; Chou, K.C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.* **2003**, *84*, 3257-3263.
31. Chou, K.C.; Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339-345.
32. Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **1993**, *268*, 16938-16948.
33. Chou, K.C. Review: Prediction of HIV protease cleavage sites in proteins, *Anal. Biochem.* **1996**, *233*, 1-14.
34. Chou, K.C. Review: Prediction of protein signal sequences. *Curr. Protein Pept. Sci.* **2002**, *3*, 615-622.
35. Chou, K.C.; Shen, H.B. Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 633-640.
36. Chou, K.C.; Forsen, S. Graphical rules for enzyme-catalyzed rate laws. *Biochem. J.* **1980**, *187*, 829-835.
37. Myers, D.; Palmer, G. Microcomputer tools for steady-state enzyme kinetics. *Bioinformatics* **1985**, *1*, 105-110.
38. Zhou, G.P.; Deng, M.H. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem. J.* **1984**, *222*, 169-176.
39. Chou, K.C. Graphic rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.* **1989**, *264*, 12074-12079.
40. Andraos, J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: New methods based on directed graphs. *Can. J. Chem.* **2008**, *86*, 342-357.
41. Chou, K.C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.* **1990**, *35*, 1-24.
42. Althaus, I.W.; Chou, J.J.; Gonzales, A.J.; Diebel, M.R.; Chou, K.C.; Kezdy, F.J.; Romero, D.L.; Aristoff, P.A.; Tarpley, W.G.; Reusser, F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry* **1993**, *32*, 6548-6554.

43. Althaus, I.W.; Chou, J.J.; Gonzales, A.J.; Diebel, M.R.; Chou, K.C.; Kezdy, F.J.; Romero, D.L.; Aristoff, P.A.; Tarpley, W.G.; Reusser, F. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J. Biol. Chem.* **1993**, *268*, 6119-6124.
44. Althaus, I.W.; Gonzales, A.J.; Chou, J.J.; Diebel, M.R.; Chou, K.C.; Kezdy, F.J.; Romero, D.L.; Aristoff, P.A.; Tarpley, W.G.; Reusser, F. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *J. Biol. Chem.* **1993**, *268*, 14875-14880.
45. Chou, K.C.; Kezdy, F.J.; Reusser, F. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.* **1994**, *221*, 217-230.
46. Chou, K.C. Graphic rule for drug metabolism systems. *Curr. Drug Metabol.* **2010**, *11*, 369-378.
47. Perez-Montoto, L.G.; Santana, L.; Gonzalez-Diaz, H. Scoring function for DNA-drug docking of anticancer and antiparasitic compounds based on spectral moments of 2D lattice graphs for molecular dynamics trajectories. *Eur. J. Medicinal Chem.* **2009**, *44*, 4461-4469.
48. Gonzalez-Diaz, H.; Perez-Montoto, L.G.; Duardo-Sanchez, A.; Paniagua, E.; Vazquez-Prieto, S.; Vilas, R.; Dea-Ayuela, M.A.; Bolas-Fernandez, F.; Munteanu, C.R.; Dorado, J.; Costas, J.; Ubeira, F.M. Generalized lattice graphs for 2D-visualization of biological information. *J. Theor. Biol.* **2009**, *261*, 136-147.
49. Munteanu, C.R.; Magalhaes, A.L.; Uriarte, E.; Gonzalez-Diaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.* **2009**, *257*, 303-311.
50. Perez-Bello, A.; Munteanu, C.R.; Ubeira, F.M.; De Magalhaes, A.L.; Uriarte, E.; Gonzalez-Diaz, H. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. *J. Theor. Biol.* **2009**, *256*, 458-466.
51. Zhang, J.; Wiemann, S. KEGGgraph: A graph approach to KEGG PATHWAY in R and Bioconductor. *Bioinformatics* **2009**, *25*, 1470-1471.
52. Chakrabarti, D. *Tools for Large Graph Mining*; PhD Thesis, School of Computer Science, Carnegie Mellon University, 2005.
53. Barabási, A.; Oltvai, Z. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101-113.
54. Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S. A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **2005**, *122*, 957-968.
55. Niu, B.; Jin, Y.H.; Lu, L.; Fen, K.Y.; Gu, L.; He, Z.S.; Lu, W.L.; Li, Y.X.; Cai, Y.D. Prediction of small molecule and enzyme interaction-ness using AdaBoost. *Mol. Divers.* **2009**, *13*, 313-320.
56. Chen, L.; Shi, X.; Kong, X.; Zeng, Z.; Cai, Y. Identifying Protein Complexes Using Hybrid Properties. *J. Proteome Res.* **2009**, *8*, 5212-5218.
57. Li, W.; Lin, K.; Feng, K.; Cai, Y. Prediction of protein structural classes using hybrid properties. *Mo. Divers.* **2008**, *12*, 171-179.
58. Yu, X.; Cao, J.; Cai, Y.; Shi, T.; Li, Y. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.* **2006**, *240*, 175-184.
59. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Protein. Struct. Funct. Genet.* **2001**, *43*, 246-255.

60. He, Z.S.; Zhang, J.; Shi, X.H.; Hu, L.L.; Kong, X.G.; Cai, Y.D.; Chou, K.C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE* **2010**, *5*, e9603.
61. Huang, T.; Shi, X.H.; Wang, P.; He, Z.; Feng, K.Y.; Hu, L.; Kong, X.; Li, Y.X.; Cai, Y.D.; Chou, K.C. Analysis and Prediction of the Metabolic Stability of Proteins Based on Their Sequential Features, Subcellular Locations and Interaction Networks. *PLoS ONE* **2010**, *5*, e10972.
62. Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics* **2009**, *6*, 262-274.
63. Xiao, X.; Chou, K.C. Digital coding of amino acids based on hydrophobic index. *Protein Peptide Lett.* **2007**, *14*, 871-875.
64. Dubchak, I.; Muchnik, I.; Holbrook, S.; Kim, S. Prediction of protein folding class using global description of amino acid sequence. *Proc. Nat. Acad. Sci.* **1995**, *92*, 8700-8704.
65. Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S. Recognition of a protein fold in the context of the SCOP classification. *Protein. Struct. Funct. Genet.* **1999**, *35*, 401-407.
66. Pollastri, G.; Baldi, P.; Fariselli, P.; Casadio, R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* **2002**, *47*, 142-153.
67. Cheng, J.; Randall, A.; Sweredoski, M.; Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucl. Acid. Res.*, **2005**, *33*, W72-W76.
68. Frishman, D.; Argos, P. Seventy-five percent accuracy in protein secondary structure prediction. *Protein. Struct. Funct. Genet.* **1997**, *27*, 329-335.
69. Camon, E.; Magrane, M.; Barrell, D.; Binns, D.; Fleischmann, W.; Kersey, P.; Mulder, N.; Oinn, T.; Maslen, J.; Cox, A. The gene ontology annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **2003**, *13*, 662-672.
70. Chou, K.; Cai, Y. Predicting Protein- Protein Interactions from Sequences in a Hybridization Space. *J. Proteome Res.* **2006**, *5*, 316-322.
71. Chou, K.C.; Shen, H.B. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE* **2010**, *5*, e9931.
72. Chou, K.C.; Shen, H.B. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE* **2010**, *5*, e11335.
73. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patt. Anal. Mach. Int.* **2005**, *27*, 1226-1238.
74. Salzberg, S.; Cost, S. Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.* **1992**, *227*, 371-374.
75. Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1-16.
76. Chou, K.C.; Shen, H.B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153-162.
77. Zhou, G.P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729-738.

78. Munteanu, C.B.; Gonzalez-Diaz, H.; Magalhaes, A.L. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *J. Theoret. Biol.* **2008**, *254*, 476-482.
79. Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *J. Theoret. Biol.* **2009**, *261*, 449-458.
80. Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Protein. Struct. Funct. Genet.* **2003**, *50*, 44-48.
81. Qiu, J.D.; Huang, J.H.; Liang, R.P.; Lu, X.Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: An approach from discrete wavelet transform. *Anal. Biochem.* **2009**, *390*, 68-73.
82. Jahandideh, S.; Sarvestani, A.S.; Abdolmaleki, P.; Jahandideh, M.; Barfeie, M. gamma-Turn types prediction in proteins using the support vector machines. *J. Theor. Biol.* **2007**, *249*, 785-790.
83. Shao, X.; Tian, Y.; Wu, L.; Wang, Y.; Jing, L.; Deng, N. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *J. Theoret. Biol.* **2009**, *258*, 289-293.
84. Yang, J.Y.; Peng, Z.L.; Yu, Z.G.; Zhang, R.J.; Anh, V.; Wang, D. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. *J. Theoret. Biol.* **2009**, *257*, 618-626.
85. Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theoret. Biol.* **2008**, *252*, 350-356.
86. Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z.; Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theoret. Biol.* **2009**, *259*, 366–372.
87. Lin, H.; Ding, H.; Feng-Biao Guo, F.B.; Zhang, A.Y.; Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Peptide Lett.* **2008**, *15*, 739-744.
88. Gu, Q.; Ding, Y.S.; Zhang, T.L. 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 Lett.* **2010**, *17*, 559-567.

Sample Availability: Samples of the compounds are available from the authors.