Int. J. Mol. Sci. **2014**, *15*, 4915-4937; doi:10.3390/ijms15034915

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

iNR-Drug: Predicting the Interaction of Drugs with Nuclear Receptors in Cellular Networking

Yue-Nong Fan [1], Xuan Xiao [1,2,4,*], Jian-Liang Min [1] and Kuo-Chen Chou [3,4]

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333046, Jiangxi, China; E-Mails: yuenong.f@163.com (Y.-N.F.); minjianliang@126.com (J.-L.M.)
[2] Information School, ZheJiang Textile & Fashion College, Ningbo 315211, China
[3] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia; E-Mail: kcchou@gordonlifescience.org
[4] Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478, USA

* Author to whom correspondence should be addressed; E-Mail: xxiao@gordonlifescience.org or jdzxiaoxuan@163.com; Tel.: +86-138-7980-9729; Fax: +86-798-873-2324.

Received: 13 January 2014; in revised form: 12 February 2014 / Accepted: 16 February 2014 / Published: 19 March 2014

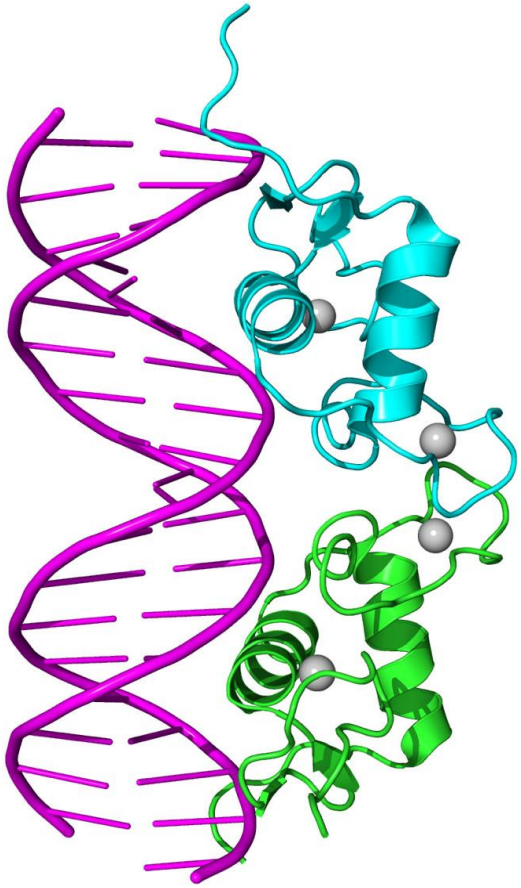
Abstract: Nuclear receptors (NRs) are closely associated with various major diseases such as cancer, diabetes, inflammatory disease, and osteoporosis. Therefore, NRs have become a frequent target for drug development. During the process of developing drugs against these diseases by targeting NRs, we are often facing a problem: Given a NR and chemical compound, can we identify whether they are really in interaction with each other in a cell? To address this problem, a predictor called "iNR-Drug" was developed. In the predictor, the drug compound concerned was formulated by a 256-D (dimensional) vector derived from its molecular fingerprint, and the NR by a 500-D vector formed by incorporating its sequential evolution information and physicochemical features into the general form of pseudo amino acid composition, and the prediction engine was operated by the SVM (support vector machine) algorithm. Compared with the existing prediction methods in this area, iNR-Drug not only can yield a higher success rate, but is also featured by a user-friendly web-server established at http://www.jci-bioinfo.cn/iNR-Drug/, which is particularly useful for most experimental scientists to obtain their desired data in a timely manner. It is anticipated that the iNR-Drug server may become a useful high throughput tool for both basic research and drug development, and that the current approach may be easily extended to study the interactions of drug with other targets as well.

Keywords: nuclear receptors (NRs); molecular fingerprints; pseudo amino acid composition; support vector machines (SVMs)

1. Introduction

With the ability to directly bind to DNA (Figure 1) and regulate the expression of adjacent genes, nuclear receptors (NRs) are a class of ligand-inducible transcription factors. They regulate various biological processes, such as homeostasis, differentiation, embryonic development, and organ physiology [1–3]. The NR superfamily has been classified into seven families: NR0 (knirps or DAX like) [4,5]; NR1 (thyroid hormone like), NR2 (HNF4-like), NR3 (estrogen like), NR4 (nerve growth factor IB-like), NR5 (fushi tarazu-F1 like), and NR6 (germ cell nuclear factor like). Since they are involved in almost all aspects of human physiology and are implicated in many major diseases such as cancer, diabetes and osteoporosis, nuclear receptors have become major drug targets [6,7], along with G protein-coupled receptors (GPCRs) [8–17], ion channels [18–20], and kinase proteins [21–24].

Figure 1. An illustration to show a nuclear receptor binding to DNA.



Identification of drug-target interactions is one of the most important steps for the new medicine development [25,26]. The method usually adopted in this step is molecular docking simulation [27–43]. However, to make molecular docking study feasible, a reliable 3D (three dimensional) structure of the target protein is the prerequisite condition. Although X-ray crystallography is a powerful tool in

determining protein 3D structures, it is time-consuming and expensive. Particularly, not all proteins can be successfully crystallized. For example, membrane proteins are very difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane protein 3D structures have been determined. Although NMR (Nuclear Magnetic Resonance) is indeed a very powerful tool in determining the 3D structures of membrane proteins as indicated by a series of recent publications (see, e.g., [44–51] and a review article [20]), it is also time-consuming and costly. To acquire the 3D structural information in a timely manner, one has to resort to various structural bioinformatics tools (see, e.g., [37]), particularly the homologous modeling approach as utilized for a series of protein receptors urgently needed during the process of drug development [19,52–57]. Unfortunately, the number of dependable templates for developing high quality 3D structures by means of homology modeling is very limited [37].

To overcome the aforementioned problems, it would be of help to develop a computational method for predicting the interactions of drugs with nuclear receptors in cellular networking based on the sequences information of the latter. The results thus obtained can be used to pre-exclude the compounds identified not in interaction with the nuclear receptors, so as to timely stop wasting time and money on those unpromising compounds [58].

Actually, based on the functional groups and biological features, a powerful method was developed recently [59] for this purpose. However, further development in this regard is definitely needed due to the following reasons. (a) He *et al*. [59] did not provide a publicly accessible web-server for their method, and hence its practical application value is quite limited, particularly for the broad experimental scientists; (b) The prediction quality can be further enhanced by incorporating some key features into the formulation of NR-drug (nuclear receptor and drug) samples via the general form of pseudo amino acid composition [60].

The present study was initiated with an attempt to develop a new method for predicting the interaction of drugs with nuclear receptors by addressing the two points.

As demonstrated by a series of recent publications [10,18,61–70] and summarized in a comprehensive review [60], to establish a really effective statistical predictor for a biomedical system, we need to consider the following steps: (a) select or construct a valid benchmark dataset to train and test the predictor; (b) represent the statistical samples with an effective formulation that can truly reflect their intrinsic correlation with the object to be predicted; (c) introduce or develop a powerful algorithm or engine to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us elaborate how to deal with these steps.

2. Results and Discussion

2.1. Benchmark Dataset

The data used in the current study were collected from KEGG (Kyoto Encyclopedia of Genes and Genomes) [71] at http://www.kegg.jp/kegg/. KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing

and other high-throughput experimental technologies. Here, the benchmark dataset $\mathbb{S}$ can be formulated as

$$\mathbb{S} = \mathbb{S}^+ \bigcup \mathbb{S}^- \tag{1}$$

where $\mathbb{S}^+$ is the positive subset that consists of the interactive drug-NR pairs only, while $\mathbb{S}^-$ the negative subset that contains of the non-interactive drug-NR pairs only, and the symbol $\bigcup$ represents the union in the set theory. The so-called "interactive" pair here means the pair whose two counterparts are interacting with each other in the drug-target networks as defined in the KEGG database [71]; while the "non-interactive" pair means that its two counterparts are not interacting with each other in the drug-target networks. The positive dataset $\mathbb{S}^+$ contains 86 drug-NR pairs, which were taken from He *et al*. [59]. The negative dataset $\mathbb{S}^-$ contains 172 non-interactive drug-NR pairs, which were derived according to the following procedures: (a) separating each of the pairs in $\mathbb{S}^+$ into single drug and NR; (b) re-coupling each of the single drugs with each of the single NRs into pairs in a way that none of them occurred in $\mathbb{S}^+$; (c) randomly picking the pairs thus formed until reaching the number two times as many as the pairs in $\mathbb{S}^+$. The 86 interactive drug-NR pairs and 172 non-interactive drug-NR pairs are given in Supplementary Information S1, from which we can see that the 86 + 172 = 258 pairs in the current benchmark dataset $\mathbb{S}$ are actually formed by 25 different NRs and 53 different compounds.

2.2. Sample Representation

Since each of the samples in the current network system contains a drug (compound) and a NR (protein), the following procedures were taken to represent the drug-NR pair sample.

2.2.1. Use 2D Molecular Fingerprints to Represent Drugs

First, for the drug part in the current benchmark dataset, we can use a 256-D vector to formulate it as given by

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_i & \cdots & d_{256} \end{bmatrix}^{\mathbf{T}} \tag{2}$$

where $\mathbf{D}$ represents the vector for a drug compound, and d_i its i-th ($i = 1,2,\cdots,256$) component that can be derived by following the "2D molecular fingerprint procedure" as elaborated in [10]. The 53 molecular fingerprint vectors thus obtained for the 53 drugs in $\mathbb{S}$ are, respectively, given in Supplementary Information S2.

2.2.2. Use Pseudo Amino Acid Composition to Represent the Nuclear Receptors

The protein sequences of the 25 different NRs in $\mathbb{S}$ are listed in Supplementary Information S3. Suppose the sequence of a nuclear receptor protein $\mathbf{P}$ with L residues is generally expressed by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\mathrm{R}_8\cdots\mathrm{R}_L \tag{3}$$

where R_1 represents the 1st residue of the protein sequence $\mathbf{P}$, R_2 the 2nd residue, and so forth. Now the problem is how to effectively represent the sequence of Equation (3) with a non-sequential or discrete model [72]. This is because all the existing operation engines, such as covariance discriminant

(CD) [17,65,73–79], neural network [80–82], support vector machine (SVM) [62–64,83], random forest [84,85], conditional random field [66], nearest neighbor (NN) [86,87]; *K*-nearest neighbor (KNN) [88–90], OET-KNN [91–94], and Fuzzy *K*-nearest neighbor [10,12,18,69,95], can only handle vector but not sequence samples. However, a vector defined in a discrete model may completely lose all the sequence-order information and hence limit the quality of prediction. Facing such a dilemma, can we find an approach to partially incorporate the sequence-order effects?

Actually, one of the most challenging problems in computational biology is how to formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. To avoid completely losing the sequence-order information for proteins, the pseudo amino acid composition [96,97] or Chou's PseAAC [98] was proposed. Ever since the concept of PseAAC was proposed in 2001 [96], it has penetrated into almost all the areas of computational proteomics, such as predicting anticancer peptides [99], predicting protein subcellular location [100–106], predicting membrane protein types [107,108], predicting protein submitochondria locations [109–112], predicting GABA(A) receptor proteins [113], predicting enzyme subfamily classes [114], predicting antibacterial peptides [115], predicting supersecondary structure [116], predicting bacterial virulent proteins [117], predicting protein structural class [118], predicting the cofactors of oxidoreductases [119], predicting metalloproteinase family [120], identifying cysteine *S*-nitrosylation sites in proteins [66], identifying bacterial secreted proteins [121], identifying antibacterial peptides [115], identifying allergenic proteins [122], identifying protein quaternary structural attributes [123,124], identifying risk type of human papillomaviruses [125], identifying cyclin proteins [126], identifying GPCRs and their types [15,16], discriminating outer membrane proteins [127], classifying amino acids [128], detecting remote homologous proteins [129], among many others (see a long list of papers cited in the References section of [60]). Moreover, the concept of PseAAC was further extended to represent the feature vectors of nucleotides [65], as well as other biological samples (see, e.g., [130–132]). Because it has been widely and increasingly used, recently two powerful soft-wares, called "PseAAC-Builder" [133] and "propy" [134], were established for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server "PseAAC" [135] built in 2008.

According to a comprehensive review [60], the general form of PseAAC for a protein sequence **P** is formulated by

$$\mathbf{P} = \left[\psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_\Omega\right]^{\mathrm{T}} \tag{4}$$

where the subscript Ω is an integer, and its value as well as the components ψ_u ($u = 1, 2, \cdots, \Omega$) will depend on how to extract the desired information from the amino acid sequence of **P** (*cf.* Equation (3)). Below, let us describe how to extract useful information to define the components of PseAAC for the NR samples concerned.

First, many earlier studies (see, e.g., [136–141]) have indicated that the amino acid composition (AAC) of a protein plays an important role in determining its attributes. The AAC contains 20 components with each representing the occurrence frequency of one of the 20 native amino acids in the protein concerned. Thus, such 20 AAC components were used here to define the first 20 elements in Equation (4); *i.e.*,

$$\psi_i = f_i^{(1)} \qquad (i = 1, 2, \cdots, 20) \tag{5}$$

where $f_i^{(1)}$ is the normalized occurrence frequency of the i-th type native amino acid in the nuclear receptor concerned. Since AAC did not contain any sequence order information, the following steps were taken to make up this shortcoming.

To avoid completely losing the local or short-range sequence order information, we considered the approach of dipeptide composition. It contained 20 × 20 = 400 components [142]. Such 400 components were used to define the next 400 elements in Equation (4); *i.e.*,

$$\psi_{j+20} = f_j^{(2)} \quad (j = 1, 2, \cdots, 400) \tag{6}$$

where $f_j^{(2)}$ is the normalized occurrence frequency of the j-th dipeptides in the nuclear receptor concerned.

To incorporate the global or long-range sequence order information, let us consider the following approach. According to molecular evolution, all biological sequences have developed starting out from a very limited number of ancestral samples. Driven by various evolutionary forces such as mutation, recombination, gene conversion, genetic drift, and selection, they have undergone many changes including changes of single residues, insertions and deletions of several residues [143], gene doubling, and gene fusion. With the accumulation of these changes over a long period of time, many original similarities between initial and resultant amino acid sequences are gradually faded out, but the corresponding proteins may still share many common attributes [37], such as having basically the same biological function and residing at a same subcellular location [144,145]. To extract the sequential evolution information and use it to define the components of Equation (4), the PSSM (Position Specific Scoring Matrix) was used as described below.

According to Schaffer [146], the sequence evolution information of a nuclear receptor protein **P** with L amino acid residues can be expressed by a $L \times 20$ matrix, as given by

$$\mathbf{P}_{\mathrm{PSSM}}^{(0)} = \begin{bmatrix} E_{1\to1}^0 & E_{1\to2}^0 & \cdots & E_{1\to20}^0 \\ E_{2\to1}^0 & E_{2\to2}^0 & \cdots & E_{2\to20}^0 \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to1}^0 & E_{L\to2}^0 & \cdots & E_{L\to20}^0 \end{bmatrix} \tag{7}$$

where $E_{i\to j}^0$ represents the original score of the i-th amino acid residue (i = 1, 2,…, L) in the nuclear receptor sequence changed to amino acid type j (j = 1, 2,…, 20) in the process of evolution. Here, the numerical codes 1, 2,…, 20 are used to respectively represent A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, the 20 single-letter codes for the 20 native amino acids. The L × 20 scores in Equation (7) were generated by using PSI-BLAST [147] to search the UniProtKB/Swiss-Prot database (The Universal Protein Resource (UniProt); http://www.uniprot.org/) through three iterations with 0.001 as the E-value cutoff for multiple sequence alignment against the sequence of the nuclear receptor concerned. In order to make every element in Equation (7) be scaled from their original score ranges into the region of [0, 1], we performed a conversion through the standard sigmoid function to make it become

$$\mathbf{P}_{\mathrm{PSSM}}^{(0)} = \begin{bmatrix} E_{1\to1}^{1} & E_{1\to2}^{1} & \cdots & E_{1\to20}^{1} \\ E_{2\to1}^{1} & E_{2\to2}^{1} & \cdots & E_{2\to20}^{1} \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to1}^{1} & E_{L\to2}^{1} & \cdots & E_{L\to20}^{1} \end{bmatrix} \tag{8}$$

where

$$E_{i\to j}^{1} = \frac{1}{1+\mathrm{e}^{-E_{i\to j}^{0}}} \quad (1 \le i \le L, \;\; 1 \le j \le 20) \tag{9}$$

Now we extract the useful information from Equation (8) to define the next 20 components of Equation (4) via the following equation

$$\psi_{j+400} = \ell_j \quad (j = 1, 2, \cdots, 20) \tag{10}$$

where

$$\ell_j = \frac{1}{L} \times \sum\nolimits_{k=1}^{L} E_{k\to j}^{1} \;\; (j = 1, 2, \cdots, 20) \tag{11}$$

Moreover, we used the grey system model approach as elaborated in [68] to further define the next 60 components of Equation (4); *i.e.*,

$$\psi_{j+440} = \phi_j \quad (j = 1, \; 2, \cdots, \; 60) \tag{12}$$

where

$$\begin{cases} \phi_{3j-2} = w_1 f_j^{(1)} a_1^j \\ \phi_{3j-1} = w_1 f_j^{(1)} a_2^j \\ \phi_{3j} \;\; = w_3 f_j^{(1)} b^j \end{cases} \quad (j = 1, 2, \cdots, 20) \tag{13}$$

In the above equation, w_1, w_2, and w_3 are weight factors, which were all set to 1 in the current study; $f_j^{(1)}$ has the same meaning as in Equation (5); a_1^j, a_2^j, and b^j are given by

$$\begin{bmatrix} a_1^j \\ a_2^j \\ b^j \end{bmatrix} = \left(\mathbf{B}_j^{\mathrm{T}} \mathbf{B}_j\right)^{-1} \mathbf{B}_j^{\mathrm{T}} \mathbf{U}_j \quad (j = 1, 2, \cdots, 20) \tag{14}$$

where

$$\mathbf{B}_j = \begin{bmatrix} -E_{2\to j}^{1} & -\left(E_{1\to j}^{1} + 0.5E_{2\to j}^{1}\right) & 1 \\ -E_{3\to j}^{1} & -\left(\sum\limits_{i=1}^{2} E_{i\to j}^{1} + 0.5E_{3\to j}^{1}\right) & 1 \\ \vdots & \vdots & \vdots \\ -E_{L\to j}^{1} & -\left(\sum\limits_{i=1}^{L-1} E_{i\to j}^{1} + 0.5E_{L\to j}^{1}\right) & 1 \end{bmatrix} \tag{15}$$

and

$$\mathbf{U}_j = \begin{bmatrix} E^1_{2\to j} - E^1_{1\to j} \\ E^1_{3\to j} - E^1_{2\to j} \\ \vdots \\ E^1_{L\to j} - E^1_{L-1\to j} \end{bmatrix} \tag{16}$$

Combining Equations (5), (6), (10) and (12), we found that the total number of the components obtained via the current approach for the PseAAC of Equation (4) is

$$\Omega = 20 + 400 + 20 + 60 = 500 \tag{17}$$

and each of the 500 components is given by

$$\psi_u = \begin{cases} f_u^{(1)} & \text{if } 1 \le u \le 20 \\ f_u^{(2)} & \text{if } 21 \le u \le 420 \\ \ell_u & \text{if } 421 \le u \le 440 \\ \phi_u & \text{if } 441 \le u \le 500 \end{cases} \tag{18}$$

2.2.3. Formulate the Pair of Drugs with Nuclear Receptor

Since the elements in Equations (2) and (4) are well defined, we can now formulate the drug-NR pair by combining the two equations as given by

$$\mathbf{G} = \mathbf{D} \oplus \mathbf{P} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{256} & \psi_1 & \psi_2 & \cdots & \psi_{500} \end{bmatrix} \tag{19}$$

where **G** represents the drug-NR pair, $\oplus$ the orthogonal sum, and the 256 + 500 = 756 components are defined by Equations (2) and (18).

For the sake of convenience, let us use x_i ($i = 1, 2, \cdots, 756$) to represent the 756 components in Equation (19); *i.e.*,

$$\mathbf{G} = \begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_{756} \end{bmatrix}^{\mathbf{T}} \tag{20}$$

To optimize the prediction quality with a time-saving approach, similar to the treatment [148–150], let us convert Equation (20) to

$$\mathbf{G} = \begin{bmatrix} y_1 & y_2 & \cdots & y_i & \cdots & y_{756} \end{bmatrix}^{\mathbf{T}} \tag{21}$$

where

$$y_i = \frac{x_i - \langle x_i \rangle}{\mathrm{SD}(x)} \tag{22}$$

where the symbol $\langle\ \rangle$ means taking the average of the quantity therein, and SD means the corresponding standard derivation.

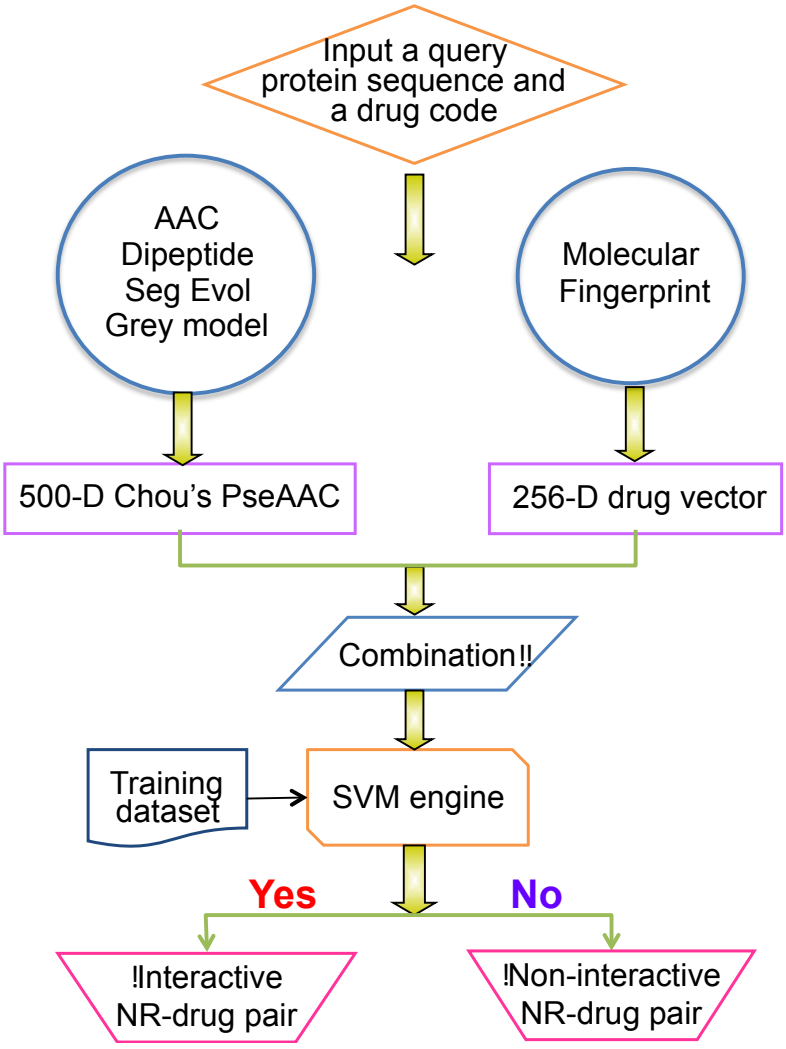
2.2.4. Operation Engine or Algorithm

In this study, the SVM (support vector machine) was used as the operation engine. SVM has been widely used in the realm of bioinformatics (see, e.g., [62–64,151–154]). The basic idea of SVM is to transform the data into a high dimensional feature space, and then determine the optimal separating hyperplane using a kernel function. For a brief formulation of SVM and how it works, see the papers [155,156]; for more details about SVM, see a monograph [157].

In this study, the LIBSVM package [158] was used as an implementation of SVM, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/, the popular radial basis function (RBF) was taken as the kernel function. For the current SVM classifier, there were two uncertain parameters: penalty parameter C and kernel parameter γ. The method of how to determine the two parameters will be given later.

The predictor obtained via the aforementioned procedure is called iNR-Drug, where "*i*" means identify, and "NR-Drug" means the interaction between nuclear receptor and drug compound. To provide an intuitive overall picture, a flowchart is provided in Figure 2 to show the process of how the predictor works in identifying the interactions between nuclear receptors and drug compounds.

Figure 2. A flowchart to show the operation process of the iNR-Drug predictor.



3. Experimental Section

3.1. Metrics for Measuring Prediction Quality

To provide a more intuitive and easier-to-understand method to measure the prediction quality, the following set of metrics based on the formulation used by Chou [159–161] in predicting signal peptides was adopted. According to Chou's formulation, the sensitivity, specificity, overall accuracy, and Matthew's correlation coefficient can be respectively expressed as [62,65–67]

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-} \right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+} \right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-} \right)}} \end{cases} \tag{23}$$

where N^+ is the total number of the interactive NR-drug pairs investigated while N_-^+ the number of the interactive NR-drug pairs incorrectly predicted as the non-interactive NR-drug pairs; N^- the total number of the non-interactive NR-drug pairs investigated while N_+^- the number of the non-interactive NR-drug pairs incorrectly predicted as the interactive NR-drug pairs.

According to Equation (23) we can easily see the following. When $N_-^+ = 0$ meaning none of the interactive NR-drug pairs was mispredicted to be a non-interactive NR-drug pair, we have the sensitivity $\mathrm{Sn} = 1$; while $N_-^+ = N^+$ meaning that all the interactive NR-drug pairs were mispredicted to be the non-interactive NR-drug pairs, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning none of the non-interactive NR-drug pairs was mispredicted, we have the specificity $\mathrm{Sp} = 1$; while $N_+^- = N^-$ meaning all the non-interactive NR-drug pairs were incorrectly predicted as interactive NR-drug pairs, we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of the interactive NR-drug pairs in the dataset $\mathbb{S}^+$ and none of the non-interactive NR-drug pairs in $\mathbb{S}^-$ was incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the interactive NR-drug pairs in the dataset $\mathbb{S}^+$ and all the non-interactive NR-drug pairs in $\mathbb{S}^-$ were mispredicted, we have the overall accuracy $\mathrm{Acc} = 0$. The Matthews correlation coefficient MCC is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$ meaning that none of the interactive NR-drug pairs in the dataset $\mathbb{S}^+$ and none of the non-interactive NR-drug pairs in $\mathbb{S}^-$ was mispredicted, we have $\mathrm{MCC} = 1$; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have $\mathrm{MCC} = 0$ meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$ we have $\mathrm{MCC} = 0$ meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier to understand when using Equation (23) to examine a predictor for its four metrics, particularly for its Mathew's correlation coefficient. It is

instructive to point out that the metrics as defined in Equation (23) are valid for single label systems; for multi-label systems, a set of more complicated metrics should be used as given in [162].

3.2. Jackknife Test Approach

How to properly test a predictor for its anticipated success rates is very important for its development as well as its potential application value. Generally speaking, the following three cross-validation methods are often used to examine the quality of a predictor and its effectiveness in practical application: independent dataset test, subsampling or *K*-fold (such as five-fold, seven-fold, or 10-fold) crossover test and jackknife test [163]. However, as elaborated by a penetrating analysis in [164], considerable arbitrariness exists in the independent dataset test. Also, as demonstrated in [165], the subsampling (or *K*-fold crossover validation) test cannot avoid arbitrariness either. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset [73,74,156,166–168]. Therefore, the jackknife test has been widely recognized and increasingly utilized by investigators to examine the quality of various predictors (see, e.g., [14,15,68,99,106,107,124,169,170]). Accordingly, in this study the jackknife test was also adopted to evaluate the accuracy of the current predictor.

As mentioned above, the SVM operation engine contains two uncertain parameters C and γ. To find their optimal values, a 2-D grid search was conducted by the jackknife test on the benchmark dataset $\mathbb{S}$. The results thus obtained are shown in Figure 3, from which it can be seen that the iNR-Drug predictor reaches its optimal status when $C = 2^3$ and $\gamma = 2^{-9}$. The corresponding rates for the four metrics (*cf.* Equation (23)) are given in Table 1, where for facilitating comparison, the overall accuracy Acc reported by He *et al*. [59] on the same benchmark dataset is also given although no results were reported by them for Sn, Sp and MCC. It can be observed from the table that the overall accuracy obtained by iNR-Drug is remarkably higher that of He *et al*. [59], and that the rates achieved by iNR-Drug for the other three metrics are also quite higher. These facts indicate that the current predictor not only can yield higher overall prediction accuracy but also is quite stable with low false prediction rates.

Figure 3. A 3-D graph showing how to optimize the two parameters γ and C in SVM via the jackknife success rates.

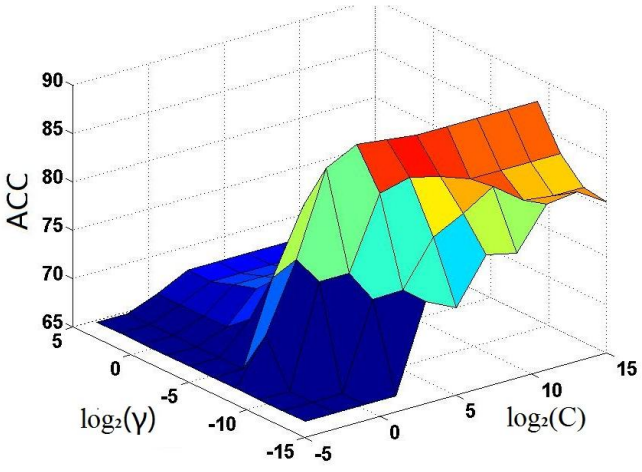


Table 1. The jackknife success rates obtained iNR-Drug in identifying the interactive NR-drug pairs and non-interactive NR-drug pairs for the benchmark dataset $\mathbb{S}$ (*cf.* Supplementary Information S1).

Metrics used for measuring prediction quality (*cf.* Equation (23))	iNR-Drug [a]	Method by He *et al.* [b]
Sn	$\frac{68}{86} = 79.07\%$	N/A
Sp	$\frac{162}{172} = 94.19\%$	N/A
Acc	$\frac{230}{258} = 89.15\%$	85.66%
MCC	75.19%	N/A

[a] The parameters used: $C = 2^3$ and $\gamma = 2^{-9}$ for the SVM operation engine; [b] See [59].

3.3. Independent Dataset Test

As mentioned above (Section 3.2), the jackknife test is the most objective method for examining the quality of a predictor. However, as a demonstration to show how to practically use the current predictor, we took 41 NR-drug pairs from the study by Yamanishi *et al*. [171] that had been confirmed by experiments as interactive pairs. For such an independent dataset, 34 were correctly identified by iNR-Drug as interactive pairs, *i.e.*, $\mathrm{Sn} = 34/41 = 82.92\%$, which is quite consistent with the rate of 79.07% achieved by the predictor on the benchmark dataset $\mathbb{S}$ via the jackknife test as reported in Table 1.

4. Conclusions

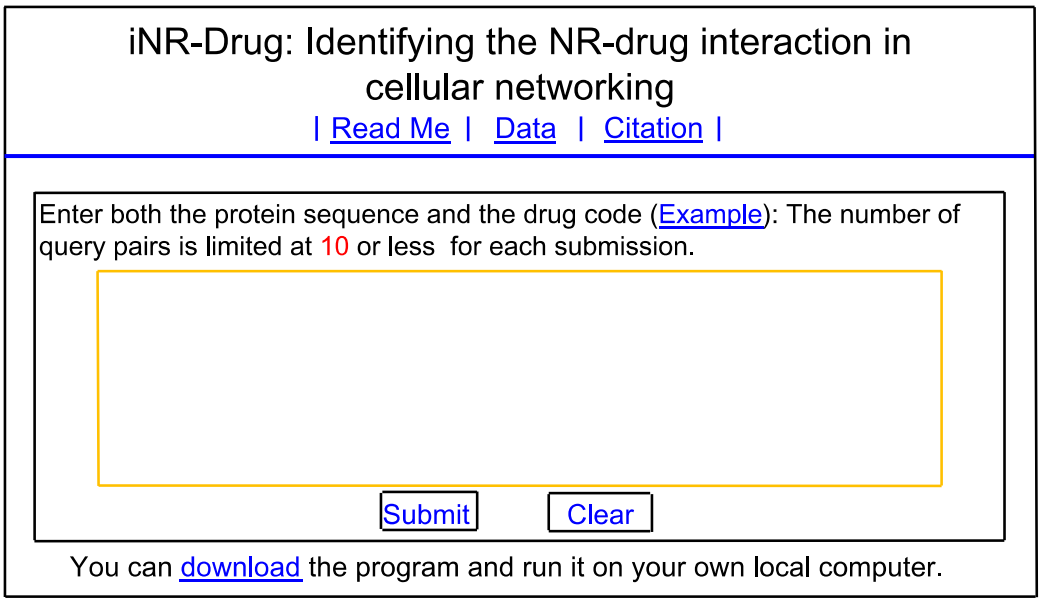
It is anticipated that the iNR-Drug predictor developed in this paper may become a useful high throughput tool for both basic research and drug development, and that the current approach may be easily extended to study the interactions of drug with other targets as well. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [98,172], a publicly accessible web-server for iNR-Drug was established.

For the convenience of the vast majority of biologists and pharmaceutical scientists, here let us provide a step-by-step guide to show how the users can easily get the desired result by using iNR-Drug web-server without the need to follow the complicated mathematical equations presented in this paper for the process of developing the predictor and its integrity.

Step 1. Open the web server at the site http://www.jci-bioinfo.cn/iNR-Drug/ and you will see the top page of the predictor on your computer screen, as shown in Figure 4. Click on the Read Me button to see a brief introduction about iNR-Drug predictor and the caveat when using it.

Step 2. Either type or copy/paste the query NR-drug pairs into the input box at the center of Figure 4. Each query pair consists of two parts: one is for the nuclear receptor sequence, and the other for the drug. The NR sequence should be in FASTA format, while the drug in the KEGG code beginning with the symbol #. Examples for the query pairs input and the corresponding output can be seen by clicking on the Example button right above the input box.

Figure 4. A semi-screenshot to show the top page of the iNR-Drug web-server. Its website address is at http://www.jci-bioinfo.cn/iNR-Drug.



Step 3. Click on the Submit button to see the predicted result. For example, if you use the three query pairs in the Example window as the input, after clicking the Submit button, you will see on your screen that the "hsa:2099" NR and the "D00066" drug are an interactive pair, and that the "hsa:2908" NR and the "D00088" drug are also an interactive pair, but that the "hsa:5468" NR and the "D00279" drug are not an interactive pair. All these results are fully consistent with the experimental observations. It takes about 3 minutes before each of these results is shown on the screen; of course, the more query pairs there is, the more time that is usually needed.

Step 4. Click on the Citation button to find the relevant paper that documents the detailed development and algorithm of iNR-Durg.

Step 5. Click on the Data button to download the benchmark dataset used to train and test the iNR-Durg predictor.

Step 6. The program code is also available by clicking the button download on the lower panel of Figure 4.

Acknowledgments

The authors would like to express their gratitude to the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of the paper. This work was supported by the grants from the National Natural Science Foundation of China (No. 31260273), the Province National Natural Science Foundation of Jiangxi (No. 2010GZS0122, No. 20114BAB211013 and No. 20122BAB201020), the Department of Education of Jiangxi Province (GJJ12490), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No. 20120BDH80023), and the Jiangxi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Altucci, L.; Gronemeyer, H. Nuclear receptors in cell life and death. *Trends Endocrinol. Metab.* **2001**, *12*, 460–468.
2. Bates, M.K.; Kerr, R.M. *Nuclear Receptors*; Nova Science: Hauppauge, NY, USA, 2011.
3. Bunce, C.M.; Campbell, M.J. *Nuclear Receptors: Current Concepts and Future Challenges*; Springer: Dordrecht, The Netherlands; New York, NY, USA, 2010; p. xii, 457.
4. Robinson-Rechavi, M.; Garcia, H.E.; Laudet, V. The nuclear receptor superfamily. *J. Cell Sci.* **2003**, *116*, 585–586.
5. Kastner, P. Non-steroid nuclear receptors: What are genetic studies telling us their role in renal life? *Cell* **1995**, *83*, 859–869.
6. Chen, T. Nuclear receptor drug discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 418–426.
7. Tirona, R.G.; Kim, R.B. Nuclear receptors and drug disposition gene regulation. *J. Pharm. Sci.* **2005**, *94*, 1169–1186.
8. Lin, W.Z.; Xiao, X.; Chou, K.C. GPCR-GIA: A web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. *Protein Eng. Des. Sel.* **2009**, *22*, 699–705.
9. Chou, K.C.; Elrod, D.W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.* **2002**, *1*, 429–433.
10. Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS One* **2013**, *8*, e72234.
11. Xiao, X.; Wang, P.; Chou, K.C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput. Chem.* **2009**, *30*, 1414–1423.
12. Xiao, X.; Wang, P.; Chou, K.C. GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Mol. Biosyst.* **201**1, *7*, 911–919.
13. 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 Pept. Lett.* **2010**, *17*, 559–567.
14. 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.
15. Xie, H.L.; Fu, L.; Nie, X.D. Using ensemble SVM to identify human GPCRs *N*-linked glycosylation sites based on the general form of Chou's PseAAC. *Protein Eng. Des. Sel.* **2013**, *26*, 735–742.
16. Zia Ur, R.; Khan, A. Identifying GPCRs and their types with Chou's pseudo amino acid composition: An approach from multi-scale energy representation and position specific scoring matrix. *Protein Pept. Lett.* **2012**, *19*, 890–903.

17. Chou, K.C. Prediction of G-protein-coupled receptor classes. *J. Proteome Res.* **2005**, *4*, 1413–1418.
18. Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol.* **2013**, *337C*, 71–79.
19. Chou, K.C. Insights from modelling three-dimensional structures of the human potassium and sodium channels. *J. Proteome Res.* **2004**, *3*, 856–861.
20. Pielak, R.M.; Chou, J.J. Influenza M2 proton channels. *Biochim. Biophys. Acta* **2011**, *1808*, 522–529.
21. Chou, K.C.; Watenpaugh, K.D.; Heinrikson, R.L. A Model of the complex between cyclin-dependent kinase 5 (Cdk5) and the activation domain of neuronal Cdk5 activator. *Biochem. Biophys. Res. Commun.* **1999**, *259*, 420–428.
22. Schnell, J.R.; Zhou, G.P.; Zweckstetter, M.; Rigby, A.C.; Chou, J.J. Rapid and accurate structure determination of coiled-coil domains using NMR dipolar couplings: Application to cGMP-dependent protein kinase Ialpha. *Protein Sci.* **2005**, *14*, 2421–2428.
23. 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.
24. Zweckstetter, M.; Schnell, J.R.; Chou, J.J. Determination of the packing mode of the coiled-coil domain of cGMP-dependent protein kinase Ialpha in solution using charge-predicted dipolar couplings. *J. Am. Chem. Soc.* **2005**, *127*, 11918–11919.
25. Knowles, J.; Gromo, G. A guide to drug discovery: Target selection in drug discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 63–69.
26. Lindsay, M.A. Target discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 831–838.
27. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
28. Chou, K.C.; Wei, D.Q.; Zhong, W.Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid., 2003, Vol. 310, 675). *Biochem. Biophys. Res. Commun.* **2003**, *308*, 148–151.
29. 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.
30. Chou, K.C.; Wei, D.Q.; Du, Q.S.; Sirois, S.; Zhong, W.Z. Review: Progress in computational approach to drug development against SARS. *Curr. Med. Chem.* **2006**, *13*, 3263–3270.
31. Du, Q.S.; Wang, S.; Wei, D.Q.; Sirois, S.; Chou, K.C. Molecular modelling and chemical modification for finding peptide inhibitor against SARS CoV Mpro. *Anal. Biochem.* **2005**, *337*, 262–270.
32. Huang, R.B.; Du, Q.S.; Wang, C.H.; Chou, K.C. An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus. *Biochem. Biophys. Res. Commun.* **2008**, *377*, 1243–1247.
33. Du, Q.S.; Huang, R.B.; Wang, C.H.; Li, X.M.; Chou, K.C. Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus. *J. Theor. Biol.* **2009**, *259*, 159–164.

34. Wei, H.; Wang, C.H.; Du, Q.S.; Meng, J.; Chou, K.C. Investigation into adamantane-based M2 inhibitors with FB-QSAR. *Med. Chem.* **2009**, *5*, 305–317.
35. Du, Q.S.; Huang, R.B.; Wang, S.Q.; Chou, K.C. Designing inhibitors of M2 proton channel against H1N1 swine influenza virus. *PLoS One* **2010**, *5*, e9388.
36. Wang, S.Q.; Du, Q.S.; Huang, R.B.; Zhang, D.W.; Chou, K.C. Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus. *Biochem. Biophys. Res. Commun.* **2009**, *386*, 432–436.
37. Chou, K.C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105–2134.
38. Cai, L.; Wang, Y.; Wang, J.F.; Chou, K.C. Identification of proteins interacting with human SP110 during the process of viral infections. *Med. Chem.* **2011**, *7*, 121–126.
39. Liao, Q.H.; Gao, Q.Z.; Wei, J.; Chou, K.C. Docking and molecular dynamics study on the inhibitory activity of novel inhibitors on epidermal growth factor receptor (EGFR). *Med. Chem.* **2011**, *7*, 24–31.
40. Li, X.B.; Wang, S.Q.; Xu, W.R.; Wang, R.L.; Chou, K.C. Novel inhibitor design for hemagglutinin against H1N1 influenza virus by core hopping method. *PLoS One* **2011**, *6*, e28111.
41. Ma, Y.; Wang, S.Q.; Xu, W.R.; Wang, R.L.; Chou, K.C. Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach. *PLoS One* **2012**, *7*, e38546.
42. Wang, J.F.; Chou, K.C. Insights from modeling the 3D structure of New Delhi metallo-beta-lactamase and its binding interactions with antibiotic drugs. *PLoS One* **2011**, *6*, e18414.
43. Wang, J.F.; Chou, K.C. Insights into the mutation-induced HHH syndrome from modeling human mitochondrial ornithine transporter-1. *PLoS One* **2012**, *7*, e31048.
44. Berardi, M.J.; Shih, W.M.; Harrison, S.C.; Chou, J.J. Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. *Nature* **2011**, *476*, 109–113.
45. Schnell, J.R.; Chou, J.J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* **2008**, *451*, 591–595.
46. OuYang, B.; Xie, S.; Berardi, M.J.; Zhao, X.M.; Dev, J.; Yu, W.; Sun, B.; Chou, J.J. Unusual architecture of the p7 channel from hepatitis C virus. *Nature* **2013**, *498*, 521–525.
47. Oxenoid, K.; Chou, J.J. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10870–10875.
48. Call, M.E.; Wucherpfennig, K.W.; Chou, J.J. The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nat. Immunol.* **2010**, *11*, 1023–1029.
49. Pielak, R.M.; Chou, J.J. Solution NMR structure of the V27A drug resistant mutant of influenza A M2 channel. *Biochem. Biophys. Res. Commun.* **2010**, *401*, 58–63.
50. Pielak, R.M.; Jason, R.; Schnell, J.R.; Chou, J.J. Mechanism of drug inhibition and drug resistance of influenza A M2 channel. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 7379–7384.
51. Wang, J.; Pielak, R.M.; McClintock, M.A.; Chou, J.J. Solution structure and functional analysis of the influenza B proton channel. *Nat. Struct. Mol. Biol.* **2009**, *16*, 1267–1271.
52. Chou, K.C.; Jones, D.; Heinrikson, R.L. Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Lett.* **1997**, *419*, 49–54.

53. Chou, K.C.; Tomasselli, A.G.; Heinrikson, R.L. Prediction of the tertiary structure of a caspase-9/inhibitor complex. *FEBS Lett.* **2000**, *470*, 249–256.
54. Chou, K.C.; Howe, W.J. Prediction of the tertiary structure of the beta-secretase zymogen. *Biochem. Biophys. Res. Commun.* **2002**, *292*, 702–708.
55. Chou, K.C. Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. *J. Proteome Res.* **2005**, *4*, 1681–1686.
56. Chou, K.C. Insights from modeling the 3D structure of DNA-CBF3b complex. *J. Proteome Res.* **2005**, *4*, 1657–1660.
57. Chou, K.C. Modeling the tertiary structure of human cathepsin-E. *Biochem. Biophys. Res. Commun.* **2005**, *331*, 56–60.
58. Sirois, S.; Hatzakis, G.E.; Wei, D.Q.; Du, Q.S.; Chou, K.C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29*, 55–67.
59. He, Z.; Zhang, J.; Shi, X.H.; Hu, L.L.; Kong, X.; Cai, Y.D.; Chou, K.C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One* **2010**, *5*, e9603.
60. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247.
61. Qiu, W.R.; Xiao, X.; Chou, K.C. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **2014**, *15*, 1746–1766.
62. Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition *Nucleic Acids Res.* **2013**, *41*, e69.
63. Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **2013**, *442*, 118–125.
64. Liu, B.; Zhang, D.; Xu, R.; Xu, J.; Wang, X.; Chen, Q.; Dong, Q.; Chou, K.C. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **2013**, doi:10.1093/bioinformatics/btt709.
65. Chen, W.; Lin, H.; Feng, P.M.; Ding, C.; Zuo, Y.C.; Chou, K.C. iNuc-PhysChem: A sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One* **2012**, *7*, e47843.
66. Xu, Y.; Ding, J.; Wu, L.Y.; Chou, K.C. iSNO-PseAAC: Predict cysteine *S*-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* **2013**, *8*, e55844.
67. Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine *S*-nitrosylation sites in proteins. *Peer J.* **2013**, *1*, e171.
68. Min, J.L.; Xiao, X.; Chou, K.C. iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed Res. Int.* **2013**, *2013*, 701317.
69. Xiao, X.; Wang, P.; Lin, W.Z.; Jia, J.H.; Chou, K.C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **2013**, *436*, 168–177.

70. Guo, S.H.; Deng, E.Z.; Xu, L.Q.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo *k*-tuple nucleotide composition. *Bioinformatics* **2014**, doi:10.1093/bioinformatics/btu083.
71. Kotera, M.; Hirakawa, M.; Tokimatsu, T.; Goto, S.; Kanehisa, M. The KEGG databases and tools facilitating omics analysis: Latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.* **2012**, *802*, 19–39.
72. Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1–16.
73. Zhou, G.P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729–738.
74. Zhou, G.P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 57–59.
75. Chou, K.C.; Elrod, D.W. Prediction of enzyme family classes. *J. Proteome Res.* **2003**, *2*, 183–190.
76. Wang, M.; Yang, J.; Xu, Z.J.; Chou, K.C. SLLE for predicting membrane protein types. *J. Theor. Biol.* **2005**, *232*, 7–15.
77. Xiao, X.; Wang, P.; Chou, K.C. Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image. *J. Theor. Biol.* **2008**, *254*, 691–696.
78. Chou, K.C. A novel approach to predicting protein structural classes in a (20–1)-*D* amino acid composition space. *Proteins: Struct. Funct. Genet.* **1995**, *21*, 319–344.
79. Zhou, G.P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins: Struct. Funct. Genet.* **2003**, *50*, 44–48.
80. Feng, K.Y.; Cai, Y.D.; Chou, K.C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 213–217.
81. Cai, Y.D.; Chou, K.C. Artificial neural network for predicting alpha-turn types. *Anal. Biochem.* **1999**, *268*, 407–409.
82. Thompson, T.B.; Chou, K.C.; Zheng, C. Neural network prediction of the HIV-1 protease cleavage sites. *J. Theor. Biol.* **1995**, *177*, 369–379.
83. Xiao, X.; Wang, P.; Chou, K.C. iNR-PhysChem: A sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *PLoS One* **2012**, *7*, e30869.
84. Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. iDNA-Prot: Identification of DNA binding proteins using random forest with grey model. *PLoS One* **2011**, *6*, e24756.
85. Kandaswamy, K.K.; Chou, K.C.; Martinetz, T.; Moller, S.; Suganthan, P.N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.* **2011**, *270*, 56–62.
86. Cai, Y.D.; Chou, K.C. Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics* **2004**, *20*, 1151–1156.
87. Chou, K.C.; Cai, Y.D. Prediction of protease types in a hybridization space. *Biochem. Biophys. Res. Commun.* **2006**, *339*, 1015–1020.
88. Chou, K.C.; Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers. *J. Proteome Res.* **2006**, *5*, 1888–1897.

89. Chou, K.C.; Shen, H.B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **2006**, *347*, 150–157.
90. Chou, K.C.; Shen, H.B. Large-scale predictions of Gram-negative bacterial protein subcellular locations. *J. Proteome Res.* **2006**, *5*, 3420–3428.
91. Chou, K.C.; Shen, H.B. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728–1734.
92. 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.
93. Shen, H.B.; Chou, K.C. Using optimized evidence-theoretic *K*-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 288–292.
94. Shen, H.B.; Chou, K.C. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.* **2009**, *394*, 269–274.
95. Shen, H.B.; Yang, J.; Chou, K.C. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *J. Theor. Biol.* **2006**, *240*, 9–13.
96. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct. Funct. Genet.* **2001**, *43*, 246–255.
97. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19.
98. Lin, S.X.; Lapointe, J. Theoretical and experimental biology in one. *J. Biomed. Sci. Eng. (JBiSE)* **2013**, *6*, 435–442.
99. Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40.
100. Mei, S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.* **2012**, *310*, 80–87.
101. Chang, T.H.; Wu, L.C.; Lee, T.Y.; Chen, S.P.; Huang, H.D.; Horng, J.T. EuLoc: A web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 91–103.
102. Fan, G.L.; Li, Q.Z. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2012**, *304*, 88–95.
103. Huang, C.; Yuan, J. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *Biosystems* **2013**, *113*, 50–57.
104. Lin, H.; Wang, H.; Ding, H.; Chen, Y.L.; Li, Q.Z. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. *Acta Biotheor.* **2009**, *57*, 321–330.
105. Wan, S.; Mak, M.W.; Kung, S.Y. GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2013**, *323*, 40–48.

106. Huang, C.; Yuan, J.Q. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. *J. Theor. Biol.* **2013**, *335*, 205–212.
107. Chen, Y.K.; Li, K.B. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *318*, 1–12.
108. Huang, C.; Yuan, J.Q. A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. *J. Membr. Biol.* **2013**, *246*, 327–334.
109. Nanni, L.; Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* **2008**, *34*, 653–660.
110. Fan, G.L.; Li, Q.Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. *Amino Acids* **2012**, *43*, 545–555.
111. Mei, S. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *J. Theor. Biol.* **2012**, *293*, 121–130.
112. 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. Theor. Biol.* **2009**, *259*, 366–372.
113. Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2011**, *281*, 18–23.
114. Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **2007**, *248*, 546–551.
115. Khosravian, M.; Faramarzi, F.K.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting antibacterial peptides by the concept of Chou;s pseudo-amino acid composition and machine learning methods. *Protein Pept. Lett.* **2013**, *20*, 180–186.
116. Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.* **2011**, *32*, 271–278.
117. Nanni, L.; Lumini, A.; Gupta, D.; Garg, A. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 467–475.
118. Sahu, S.S.; Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **2010**, *34*, 320–327.
119. Zhang, G.Y.; Fang, B.S. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. *J. Theor. Biol.* **2008**, *253*, 310–315.
120. Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics* **2011**, *12*, 191–197.

121. Yu, L.; Guo, Y.; Li, Y.; Li, G.; Li, M.; Luo, J.; Xiong, W.; Qin, W. SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2010**, *267*, 1–6.
122. Mohabatkar, H.; Beigi, M.M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.* **2013**, *9*, 133–137.
123. Zhang, S.W.; Chen, W.; Yang, F.; Pan, Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: A sequence-segmented PseAAC approach. *Amino Acids* **2008**, *35*, 591–598.
124. Sun, X.Y.; Shi, S.P.; Qiu, J.D.; Suo, S.B.; Huang, S.Y.; Liang, R.P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. BioSyst.* **2012**, *8*, 3178–3184.
125. Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* **2010**, *263*, 203–209.
126. Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2010**, *17*, 1207–1214.
127. Hayat, M.; Khan, A. Discriminating outer membrane proteins with fuzzy *K*-nearest neighbor algorithms based on the general form of Chou's PseAAC. *Protein Pept. Lett.* **2012**, *19*, 411–421.
128. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* **2009**, *257*, 17–26.
129. Liu, B.; Wang, X.; Zou, Q.; Dong, Q.; Chen, Q. Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Inform.* **2013**, *32*, 775–782.
130. Li, B.Q.; Huang, T.; Liu, L.; Cai, Y.D.; Chou, K.C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS One* **2012**, *7*, e33393.
131. Huang, T.; Wang, J.; Cai, Y.D.; Yu, H.; Chou, K.C. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. *PLoS One* **2012**, *7*, e34460.
132. Jiang, Y.; Huang, T.; Lei, C.; Gao, Y.F.; Cai, Y.D.; Chou, K.C. Signal propagation in protein interaction network during colorectal cancer progression. *BioMed Res. Int.* **2013**, *2013*, 287019.
133. Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **2012**, *425*, 117–119.
134. Cao, D.S.; Xu, Q.S.; Liang, Y.Z. Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, *29*, 960–962.
135. Shen, H.B.; Chou, K.C. PseAAC: A flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008**, *373*, 386–388.
136. Nakashima, H.; Nishikawa, K.; Ooi, T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* **1986**, *99*, 153–162.
137. Zhang, C.T.; Chou, K.C. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.* **1992**, *1*, 401–408.

138. Zhang, C.T.; Chou, K.C. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophys. J.* **1992**, *63*, 1523–1529.
139. Chou, K.C.; Zhang, C.T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.* **1994**, *269*, 22014–22020.
140. Zhang, C.T.; Chou, K.C. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. II. correlative effect. *J. Protein Chem.* **1995**, *14*, 251–258.
141. Chou, K.C. Does the folding type of a protein depend on its amino acid composition? *FEBS Lett.* **1995**, *363*, 127–131.
142. Liu, W.; Chou, K.C. Protein secondary structural content prediction. *Protein Eng.* **1999**, *12*, 1041–1050.
143. Chou, K.C. The convergence-divergence duality in lectin domains of the selectin family and its implications. *FEBS Lett.* **1995**, *363*, 123–126.
144. Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One* **2011**, *6*, e18258.
145. Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **2012**, *8*, 629–641.
146. Schaffer, A.A.; Aravind, L.; Madden, T.L.; Shavirin, S.; Spouge, J.L.; Wolf, Y.I.; Koonin, E.V.; Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **2001**, *29*, 2994–3005.
147. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
148. Bolstad, B.M.; Irizarry, R.A.; Astrand, M.; Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **2003**, *19*, 185–193.
149. Schadt, E.E.; Li, C.; Ellis, B.; Wong, W.H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell Biochem. Suppl.* **2001**, *37*, 120–125.
150. Shi, J.Y.; Zhang, S.W.; Pan, Q.; Cheng, Y.M.; Xie, J. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids* **2007**, *33*, 69–74.
151. Liu, H.; Wang, M.; Chou, K.C. Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *336*, 737–739.
152. Wang, S.Q.; Yang, J.; Chou, K.C. Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. *J. Theor. Biol.* **2006**, *242*, 941–946.
153. Chen, J.; Liu, H.; Yang, J.; Chou, K.C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **2007**, *33*, 423–428.
154. Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition via grey system model. *PLoS One* **2012**, *7*, e49040.
155. Chou, K.C.; Cai, Y.D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277*, 45765–45769.

156. 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.
157. Cristianini, N.; Shawe-Taylor, J. *An Introduction of Support Vector Machines and Other Kernel-Based Learning Methodds*; Cambridge University Press: Cambridge, UK, 2000.
158. Chang, C.; Lin, C. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, doi:10.1145/1961189.1961199.
159. Chou, K.C. Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 136–139.
160. Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.* **2001**, *14*, 75–79.
161. Chou, K.C. Prediction of signal peptides using scaled window. *Peptides* **2001**, *22*, 1973–1979.
162. Chou, K.C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* **2013**, *9*, 1092–1100.
163. Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.
164. 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.
165. Chou, K.C.; Shen, H.B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci.* **2010**, *2*, 1090–1103.
166. Cai, Y.D.; Zhou, G.P.; Chou, K.C. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. *J. Theor. Biol.* **2005**, *234*, 145–149.
167. Cai, Y.D.; Zhou, G.P.; Jen, C.H.; Lin, S.L.; Chou, K.C. Identify catalytic triads of serine hydrolases by support vector machines. *J. Theor. Biol.* **2004**, *228*, 551–557.
168. Shi, J.Y.; Zhang, S.W.; Pan, Q.; Zhou, G.P. Using pseudo amino acid composition to predict protein subcellular location: Approached with amino acid composition distribution. *Amino Acids* **2008**, *35*, 321–327.
169. Fan, G.L.; Li, Q.Z. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *334*, 45–51.
170. Wu, Z.C.; Xiao, X.; Chou, K.C. iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein Pept. Lett.* **2012**, *19*, 4–14.
171. Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **2010**, *26*, i246–i254.
172. Chou, K.C.; Shen, H.B. Review: Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.* **2009**, *2*, 63–92.