Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 701317, 13 pages
http://dx.doi.org/10.1155/2013/701317

Research Article

iEzy-Drug: A Web Server for Identifying the Interaction between Enzymes and Drugs in Cellular Networking

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

[1] *Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046, China*
[2] *Information School, ZheJiang Textile & Fashion College, NingBo 315211, China*
[3] *Gordon Life Science Institute, Belmont, MA 02478, USA*
[4] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia*

Correspondence should be addressed to Xuan Xiao; xxiao@gordonlifescience.org

Received 7 August 2013; Accepted 17 September 2013

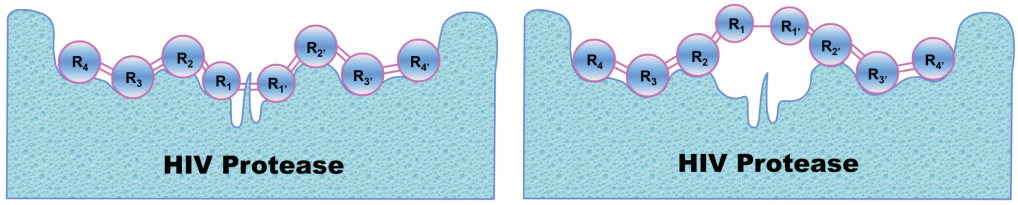
Academic Editor: Tatsuya Akutsu

With the features of extremely high selectivity and efficiency in catalyzing almost all the chemical reactions in cells, enzymes play vitally important roles for the life of an organism and hence have become frequent targets for drug design. An essential step in developing drugs by targeting enzymes is to identify drug-enzyme interactions in cells. It is both time-consuming and costly to do this purely by means of experimental techniques alone. Although some computational methods were developed in this regard based on the knowledge of the three-dimensional structure of enzyme, unfortunately their usage is quite limited because three-dimensional structures of many enzymes are still unknown. Here, we reported a sequence-based predictor, called "iEzy-Drug," in which each drug compound was formulated by a molecular fingerprint with 258 feature components, each enzyme by the Chou's pseudo amino acid composition generated via incorporating sequential evolution information and physicochemical features derived from its sequence, and the prediction engine was operated by the fuzzy K-nearest neighbor algorithm. The overall success rate achieved by iEzy-Drug via rigorous cross-validations was about 91%. Moreover, to maximize the convenience for the majority of experimental scientists, a user-friendly web server was established, by which users can easily obtain their desired results.

1. Introduction

Enzymes are biomacromolecules that catalyze almost all the chemical reactions essential for the life of a cell [1]. Most enzymes are proteins although some RNA molecules have been identified to possess the function of enzyme as well. As catalysts, enzymes possess two exceptional features: one is of high efficiency and the other of high selectivity. For instance, the second-order rate constant between some enzymes and their substrates [2] was surprisingly high [3], which could almost reach the upper limit of diffusion-controlled reaction rate according to the calculation and analysis by Chou and coworkers [4–6]. The high selectivity or specificity of enzymes was likened to the "lock-and-key" model, implying that an accurate fit is required between the active site of an enzyme and its substrate for the catalytic reaction to occur. Owing to the previous unique features, enzymes play a crucial role in controlling and regulating the order of chemical reactions in cells that is vitally important for their survival. It is also because of this that enzymes are excellent drug targets, and actually many drugs are enzyme inhibitors. For example, some peptide inhibitors against HIV/AIDS [7–10] and SARS (severe acute respiratory syndrome) [11–13] were based on the Chou's distorted key theory [14], as illustrated in Figure 1, where (a) shows a good fit for a cleavable octapeptide with the active site of HIV-protease and (b) shows that the peptide has become an ideal inhibitor or "distorted key" after its scissile bond is modified. For a brief introduction about the Chou's distorted key theory and its application for designing peptide drugs, see a Wikipedia article at http://en.wikipedia.org/wiki/Chou's_distorted_key_theory_for_peptide_drugs.

To develop enzyme-targeting drugs, an essential step is to identify drug-enzyme interaction in cellular networking [15]. The completion of the human genome project and the



(a) (b)

Figure 1: A schematic drawing to illustrate how to use Chou's distorted key theory to develop peptide drugs against HIV/AIDS. (a) shows a good fitting and binding of a peptide to the active site of HIV protease right before it is cleaved by the enzyme. (b) shows that the peptide has become a noncleavable one after its scissile bond is modified although it can still tightly bind to the active site. Such a modified peptide, or "distorted key", will automatically become an inhibitor candidate against HIV protease.

emergence of molecular medicine have provided excellent opportunity to discover unknown target enzymes for drugs. Many efforts were made in this regard by computationally analyzing drug-enzyme interactions. The most commonly used approaches are docking simulations (see, e.g., [16–19]) and protein cleavage site analysis (see, e.g., [8, 12, 13]) based on Chou's distorted key theory [14]. However, the latter approach is mainly used to find peptide drugs. Compared with the smaller organic compounds, although peptide drugs have the advantage of low toxicity to human body, they have the shortcoming of poor metabolic stability and low bioavailability due to their inability to readily crossing thru membrane barriers such as the intestinal and blood-brain barriers [20]. In contrast, the molecular docking is indeed a useful vehicle for investigating the interaction of an enzyme receptor with its organic inhibitor and revealing their binding mechanism as demonstrated by a series of studies [11, 19–23]. However, to conduct molecular docking, a necessary prerequisite is the availability of the 3D (three dimensional) structure of the targeted enzyme. Unfortunately, the 3D structures of many enzymes are still unknown. Although X-ray crystallography is a powerful tool in determining the 3D structures of enzymes, it is time-consuming and expensive. Particularly, not all enzymes can be successfully crystallized. For example, membrane enzymes are very difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane enzyme 3D structures have been determined. Although NMR 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., [24–30]), it is time-consuming and costly. To acquire the structural information in a timely manner, one has to resort to various structural bioinformatics tools (see, e.g., [18, 31, 32]). Unfortunately the number of templates for developing high quality 3D structures by structural bioinformatics is very limited.

Therefore, it would save us a lot of time and money if we could identify the interactions between enzymes and drugs before carrying out any intense study in this regard. In view of this, the present study was initiated in an attempt to develop a computational method based on the sequence-derived features that can be used to predict the drug-enzyme interactions in cellular networking.

As summarized in a comprehensive review [33] and demonstrated by a series of recent publications [34–37], to successfully develop the desired method, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) denote the drug-enzyme samples with an effective formulation that can truly reflect their intrinsic relation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) conduct a rigorous cross-validation to objectively evaluate its anticipated accuracy; (v) establish a user-friendly web-server for the predictor that is freely accessible to the public. Next, let us elaborate how to deal with these procedures one by one.

2. Materials and Methods

2.1. Benchmark Dataset. The data used in this study were collected from Kyoto Encyclopedia of Genes and Genomes (KEGG) [38] at http://www.kegg.jp/kegg/, which 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. For the current study, the benchmark dataset $\mathbb{S}$ can be formulated as

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^-, \quad (1)$$

where $\mathbb{S}^+$ is the positive subset that consists of the interactive enzyme-drug pairs only, while $\mathbb{S}^-$ is the negative subset that contains of the noninteractive enzyme-drug pairs only, and the symbol $\cup$ represents the union in the set theory. Here, the "interactive" pair means the pair whose two counterparts are interacted with each other in the drug-target networks as defined in the KEGG database [38], while the "noninteractive" pair means that its two counter parts are not interacted with each other in the drug-target networks. The positive dataset $\mathbb{S}^+$ contains 2,719 enzyme-drug pairs derived from Yamanishi et al. [39]. The negative dataset $\mathbb{S}^-$ contains 5,438 noninteractive enzyme-drug pairs, which were derived according to the following procedures: (i) separating each of the pairs in $\mathbb{S}^+$ into single drug and

enzyme; (ii) recoupling each of the single drugs with each of the single enzymes into pairs in a way that none of them occurred in $\mathbb{S}^+$; (iii) randomly picking the pairs, thus, formed until they reached the number two times as many as the pairs in $\mathbb{S}^+$. The 2,719 interactive enzyme-drug pairs and 5,438 noninteractive enzyme-drug pairs are given in Online Supporting Information S1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2013/701317) All the detailed information for the compounds or drugs listed there can be found in the KEGG database via their codes.

2.2. Sample Representation. Since each of the samples in the current network system contains an enzyme (protein) and a drug, a combination of the following two approaches was adopted to represent the enzyme-drug pair samples.

2.2.1. Drug

(a) 2D Molecular Fingerprints. Although the number of drugs is extremely large, most of them are small organic molecules and are composed of some fixed small structures [40]. The identification of small molecules or structures can be used to detect the drug-target interactions [41]. Molecular fingerprints are bit-string representations of molecular structure and properties [42]. It should be pointed out that there are many types of structural representations that have been suggested for the description of drug molecules, including physicochemical properties [43], chemical graphs [44], topological indices [45], 3D pharmacophore patterns, and molecular fields. In the current study, let us use the simple and generally adopted 2D molecular fingerprints to represent drug molecules, as described below.

First, for each of the drugs concerned, we can obtain a MOL file from the KEGG database [38] via its code that contains the detailed information of chemical structure. Second, we can convert the MOL file format into its 2D molecular fingerprint file format by using a chemical toolbox software called OpenBabel [46], which can be downloaded from the website at http://openbabel.org/. The current version of OpenBabel can generate four types of fingerprints: FP2, FP3, FP4, and MACCS. In the current study, we used the FP2 fingerprint format. It is a path-based fingerprint that identifies small molecule fragments based on all linear and ring substructures and maps them onto a bit-string using a hash function (somewhat similar to the daylight fingerprints [47, 48]). It is a length of 256-bit hexadecimal string obtained from the OpenBabel, and we can convert it to a 256-bit vector. Then, a molecular fingerprint can be formulated as a 256-D vector given by

$$\mathrm{MF} = \left[A_1 \ \cdots \ A_j \ \cdots \ A_{256}\right]^{\mathbf{T}}, \tag{2}$$

where A_j $(j = 1, 2, \ldots, 256)$ is an integer between 0 and 15, and $\mathbf{T}$ is the matrix transpose operator.

In order to capture as much useful information from a molecular fingerprint as possible, we can also convert the above 256-bit hexadecimal string into a 1024-bit binary vector, which is a digital sequence only including 0 and 1, and consider two different digital signal characteristics for the digital sequence as follows.

(b) Information Entropy. Shannon proposed that any information is redundant, and redundant size is related with the occurrence probability or uncertainty of each symbol such as numbers, letters, or words among the information. The information entropy for a system with a probability distribution $P(x_i)$ for two classes information entropy [49] is defined as

$$H_x = -\sum_x P\left(x_i\right) \log_2 P\left(x_i\right) \quad (i = 0, 1), \tag{3}$$

where $P(x_i)$ represents the occurrence probability of number i in the aforementioned 1024-bit binary vector and the information entropy H_x is a measure value of the information amount. For example, for the digital sequence 1001000011010010, the value of the information entropy H_x, thus, obtained is

$$\begin{aligned} P\left(x_0\right) &= \frac{9}{15} = 0.6, \\ P\left(x_1\right) &= \frac{6}{15} = 0.4, \\ H_x &= -\left(0.6 \times \log_2 0.6 + 0.4 \times \log_2 0.4\right) = 0.971. \end{aligned} \tag{4}$$

(c) Complexity Factor. The Lempel-Ziv (LZ) complexity [50] reflects the order that is retained in the sequence, and hence was adopted in this study. For each step only two operations were allowed in the process to get the LZ complexity: either copying the longest section from the part of a nonempty sequence or generating an additional symbol mark that ensures the uniqueness of per component $S(i_{k-1} \ \rightarrow \ i_k)$. Its substring is expressed by

$$S\left(i \longrightarrow j\right) = m_i m_{i+1} m_{i+2} \cdots m_j \quad \left(1 \le i \le j \le L\right), \tag{5}$$

where m_1 represents the 1st digital value, m_2 the 2nd value, and so forth. A nonempty digital sequence is synthesized according to the following formula:

$$\begin{aligned} \mathrm{Syn}\left(S\right) = {} & S\left(1 \longrightarrow i_1\right) \bullet S\left(i_1 + 1 \longrightarrow i_2\right) \\ & \bullet \cdots \bullet S\left(i_{m-1} + 1 \longrightarrow i_L\right). \end{aligned} \tag{6}$$

Suppose that $S = m_1 m_2 m_3 m_4 m_5 \cdots m_L$ has been reconstructed by the subsymbol m_r which is viewed as the newly inserted symbol. The substring up to m_r will be denoted by $S(1 \ \rightarrow \ r)\bullet$, where the bold dot $\bullet$ indicates that m_r is a newly inserted symbol for checking whether the rest of the substring $S(r+1 \ \rightarrow \ L)$ can be reconstructed by a simple process. At first suppose $S(q) = m_r + 1$, and see whether $S(q)$ is the substring for the subsequence $S(1 \ \rightarrow \ r)$, which means deleting the last symbol from the substring $S(1 \ \rightarrow \ r)S(q)$. If the answer is "no", we insert $S(q)$ into the sequence followed by a dot $\bullet$. Thus, it could not be obtained by the same operation. If the answer is "yes", no new symbol is needed, and we can go on to proceed with $S(q) = m_{r+1} m_{r+2}$ and repeat the same previous procedure. The LZ complexity is the number of dots (plus

one if the string is not terminated by a dot). For example, for the sequence 100100011010010, syn(P) and the corresponding complexity factor CF are described as

$$\begin{aligned} \text{Syn}\,(S) &= 1 \bullet 0 \bullet 01 \bullet 000 \bullet 11 \bullet 01001 \bullet 0 \\ \text{CF} &= 7. \end{aligned} \tag{7}$$

Thus, by adding the information entropy H_x (4) and complexity factor CF (7) into the molecular fingerprint MF (2), we obtained a total of $(256 + 1 + 1) = 258$ feature elements to represent a drug compound; that is, it can now be formulated as a 258-D vector given by

$$D = \left[A_1 \;\; A_2 \;\; \cdots \;\; A_{256} \;\; H_x \;\; \text{CF}\right]^{\mathbf{T}}, \tag{8}$$

where A_i has the same meaning as in (2), while H_x and CF are the information entropy and complexity factor, respectively, as described in the previous two sections.

2.2.2. Enzyme. The sequences of the enzymes involved in this study are given in Online Supporting Information S2. Now the problem is how to effectively represent these enzyme sequences for the current study. Generally speaking, there are two kinds of approaches to formulate enzyme sequences: the sequential model and the nonsequential or discrete model [51]. The most typical sequential representation for an enzyme sample $\mathbf{E}$ with L residues is its entire amino acid sequence; that is,

$$\mathbf{E} = R_1R_2R_3R_4R_5R_6R_7 \cdots R_L, \tag{9}$$

where R_1 represents the 1st residue, R_2 the 2nd residue, and so forth. An enzyme sample thus formulated can contain its most complete information. This is an obvious advantage of the sequential representation. To get the desired results, the sequence-similarity-search-based tools, such as BLAST [52, 53], are usually utilized to conduct the prediction. However, this kind of approach failed to work when the query enzyme did not have significant homology to enzyme of known characters. Thus, various nonsequential representation models were proposed. The simplest nonsequential model for an enzyme was based on its amino acid composition (AAC), as defined by

$$\mathbf{E} = \left[f_1 \;\; f_2 \;\; \cdots \;\; f_{20}\right]^{\mathbf{T}}, \tag{10}$$

where f_u $(u = 1, 2, \ldots, 20)$ are the normalized occurrence frequencies of the 20 native amino acids [54–56] in the enzyme $\mathbf{E}$, and $\mathbf{T}$ has the same meaning as in (2) and (8). The AAC-discrete model was widely used for identifying various attributes of proteins (see, e.g., [57–61]). However, as can be seen from (10), all the sequence order effects were lost by using the AAC-discrete model. This is its main shortcoming. To avoid completely losing the sequence-order information, the pseudo amino acid composition [62, 63] or Chou's PseAAC [3] was proposed to replace the simple AAC model. Since the concept of PseAAC was proposed in 2001 [62], it has penetrated into almost all the fields of protein attribute predictions and computational proteomics, such as predicting supersecondary structure [64], predicting metalloproteinase family [65], predicting membrane protein types [66, 67], predicting protein structural class [68], discriminating outer membrane proteins [69], identifying antibacterial peptides [70], identifying allergenic proteins [71], identifying bacterial virulent proteins [72], predicting protein subcellular location [73, 74], identifying GPCRs and their types [75], identifying protein quaternary structural attributes [76], predicting protein submitochondria locations [77], identifying risk type of human papillomaviruses [78], identifying cyclin proteins [79], predicting GABA(A) receptor proteins [80], and predicting cysteine S-nitrosylation sites in proteins [81], among many others (see a long list of papers cited in the References section of [33]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [36, 82], as well as other biological samples (see, e.g., [83, 84]). Because it has been widely and increasingly used, recently two powerful softwares called "PseAAC-Builder" [85] and "propy" [86] were established for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server PseAAC [87] built in 2008. According to a recent review [33], the general form of Chou's PseAAC for an enzyme sample can be formulated by

$$\mathbf{E} = \left[\psi_1 \;\; \psi_2 \;\; \cdots \;\; \psi_u \;\; \cdots \;\; \psi_\Omega\right]^{\mathbf{T}}, \tag{11}$$

where the subscript Ω is an integer, and its value as well as the components ψ_u $(u = 1, 2, \ldots, \Omega)$ will depend on how to extract the desired information from the amino acid sequence of $\mathbf{E}$ (cf. (10)). Next, let us describe how to extract useful information from the benchmark dataset $\mathbb{S}$ and Online Supporting Information S2 to define the enzyme samples concerned via (11).

To incorporate as much useful information as possible from an enzyme sample, we are to approach this problem from three different angles, followed by incorporating the feature elements thus obtained into the general form of PseAAC of (11).

(a) Amino Acid Composition. The components of amino acid composition have been widely used to predict various protein attributes [57–61]. In this study, they were also included as the first 20 elements in the general Chou's PseAAC (cf. (11)); that is,

$$\psi_u = f_u \quad (u = 1, 2, \ldots, 20), \tag{12}$$

where f_u has the same meaning as in (10).

(b) Dipeptide Composition. Dipeptide composition has been used to predict the protein secondary structural contents [88, 89] as well as various protein attributes (see, e.g., [90–93]). The number of different dipeptides is $20 \times 20 = 400$. Suppose that the normalized occurrence frequencies of the 400 dipeptides in an enzyme sample are given by

$$f_u^{(2)} \quad (u = 1, 2, \ldots, 400). \tag{13}$$

Incorporating the above 400 dipeptide components into (11), we have

$$\psi_{u+20} = f_u^{(2)} \quad (u = 1, 2, \ldots, 400). \tag{14}$$

(c) Sequential Evolution Information. Biology is a natural science with a historic dimension. All biological species have developed starting out from a very limited number of ancestral species. Their evolution involves changes of single residues, insertions and deletions of several residues [94], gene doubling, and gene fusion. With these changes accumulated for a long period of time, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share many common attributes [18], such as having basically the same biological function and residing at a same subcellular location. To extract the sequential evolution information and use it to define the components of (11), the PSSM (Position Specific Scoring Matrix) was used as described next.

According to Schäffer et al. [95], the sequence evolution information of enzyme **E** with L amino acid residues can be expressed by an $L \times 20$ matrix, as given by

$$\mathbf{P}_{\text{PSSM}}^{(0)} = \begin{bmatrix} E_{1\to1}^0 & E_{1\to2}^0 & \cdots & E_{1\to20}^0 \\ E_{2\to1}^0 & E_{2\to2}^0 & \cdots & E_{2\to20}^0 \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to1}^0 & E_{L\to2}^0 & \cdots & E_{L\to20}^0 \end{bmatrix}, \tag{15}$$

where $E_{i\to j}^0$ represents the original score of the ith amino acid residue ($i = 1, 2, \ldots, L$) in the enzyme sequence changed to amino acid type j ($j = 1, 2, \ldots, 20$) in the process of evolution. Here, the numerical codes $1, 2, \ldots, 20$ are used to represent the 20 native amino acid types denoted by A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. The $L \times 20$ scores in (15) were generated by using PSI-BLAST [96] to search the UniProtKB/Swiss-Prot database (Release 2013-05) through three iterations with 0.001 as the E value cutoff for multiple sequence alignment against the sequence of the enzyme **E**. In order to make every element in (15) be scaled from their original score ranges into region of $[0, 1]$, we performed a conversion through the standard sigmoid function to make it become

$$\mathbf{P}_{\text{PSSM}}^{(1)} = \begin{bmatrix} E_{1\to1}^1 & E_{1\to2}^1 & \cdots & E_{1\to20}^1 \\ E_{2\to1}^1 & E_{2\to2}^1 & \cdots & E_{2\to20}^1 \\ \vdots & \vdots & \vdots & \vdots \\ E_{L\to1}^1 & E_{L\to2}^1 & \cdots & E_{L\to20}^1 \end{bmatrix}, \tag{16}$$

where

$$E_{i\to j}^1 = \frac{1}{1 + e^{-E_{i\to j}^0}} \quad (1 \le i \le L, 1 \le j \le 20). \tag{17}$$

Now, we extract the useful information from (16) to define the components of (11) via the following approach:

$$\psi_{u+420} = \ell_u \quad (u = 1, 2, \ldots, 20), \tag{18}$$

where

$$\ell_j = \frac{1}{L} \times \sum_{k=1}^{L} E_{k\to j}^1 \quad (j = 1, 2, \ldots, 20). \tag{19}$$

(d) Grey System Model Approach. The grey system theory [97] is quite useful in dealing with complicated systems that lack sufficient information, or need to process uncertain information. According to the grey system theory, we can extract the following information from the jth column of (16); that is,

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

where

$$\mathbf{B}_j = \begin{bmatrix} -E_{2\to j}^1 & -E_{1\to j}^1 - 0.5E_{2\to j}^1 & 1 \\ -E_{3\to j}^1 & -\sum_{i=1}^{2} E_{i\to j}^1 - 0.5E_{3\to j}^1 & 1 \\ \vdots & \vdots & \vdots \\ -E_{L\to j}^1 & -\sum_{i=1}^{L-1} E_{i\to j}^1 - 0.5E_{L\to j}^1 & 1 \end{bmatrix},$$
$$\mathbf{U}_j = \begin{bmatrix} E_{2\to j}^1 - E_{1\to j}^1 \\ E_{3\to j}^1 - E_{2\to j}^1 \\ \vdots \\ E_{L\to j}^1 - E_{L-1\to j}^1 \end{bmatrix}. \tag{21}$$

Therefore, based on the grey system theory and (20), we can extract another $20 \times 2 = 40$ quantities from (16) to define the components of (11); that is,

$$\varphi_j = \begin{cases} w_1 a_1^j & \text{when } j \text{ is an odd number} \\ w_2 a_2^j & \text{when } j \text{ is an even number} \end{cases} \quad 1 \le j \le 20, \tag{22}$$

where a_1^j and a_2^j are given by (20); w_1 and w_2 are weight factors, which were all set to 1 in the current study.

Substituting the elements in (12), (14), (18), and (22), we finally obtain a total of $\Omega = 20 + 400 + 20 + 40 = 480$ components for the PseAAC of (11), where

$$\psi_u = \begin{cases} f_u & \text{when } 1 \le u \le 20 \\ f_u^{(2)} & \text{when } 21 \le u \le 420 \\ \ell_u & \text{when } 421 \le u \le 440 \\ \varphi_u & \text{when } 440 \le u \le 480. \end{cases} \tag{23}$$

In other words, in this study (11) or Chou's PseAAC is a 480-D vector, whose 480 components are given by (23) derived from the amino acid composition, dipeptide composition, sequential evolution information, and grey system theory.

(e) Representing Enzyme-Drug Pairs. Now the pair between an enzyme molecule **E** and a drug compound **D** can be formulated by combing (8) and (11), as given by

$$\begin{aligned} \mathbf{G} &= \mathbf{D} \oplus \mathbf{E} \\ &= \left[A_1 \ \cdots \ A_{256} \ H_x \ \text{CF} \ \psi_1 \ \cdots \ \psi_{480}\right]^{\mathbf{T}}, \end{aligned} \tag{24}$$

where $\mathbf{G}$ represents the enzyme-drug pair, $\oplus$ the orthogonal sum [51], and each of the $(258 + 480) = 738$ feature elements is given in (8) and (23).

For the convenience of the later formulation, let us use x_i $(i = 1, 2, \ldots, 738)$ to represent the 738 components of (24); that is,

$$\mathbf{G} = \left[x_1 \;\; x_2 \;\; \cdots \;\; x_i \;\; \cdots \;\; x_{738}\right]^{\mathbf{T}}. \tag{25}$$

To optimize the prediction results, different weights were usually tested for each of the elements in (25). However, since it would consume a lot of computational time for a total of 738 weight factors, here let us adopt the normalization approach to deal with this problem as done in [98, 99]; that is, convert x_i in (25) to y_i according to the following equation:

$$y_i = \frac{2\tan^{-1}\left(x_i\right)}{\pi} \quad (i = 1, 2, \ldots, 738), \tag{26}$$

where $\tan^{-1}$ means arctangent. By means of (26), every component in (25) will be converted into the range of $[-1, 1]$; that is, we have $-1 \leq y_i \leq 1$. As demonstrated in [98, 99], the normalization approach via (26) was quite effective in enhancing the quality of prediction operated in a high dimension space. Therefore, in this study, we would not to take the procedure of optimizing the weight factors, significantly reducing the computational times.

2.3. Fuzzy K-Nearest Neighbour Algorithm. The K-NN (K-Nearest Neighbor) classifier is quite popular in pattern recognition community owing to its good performance and simple-to-use feature. According to the K-NN rule [100], named also as the "voting K-NN rule," the query sample should be assigned to the subset represented by a majority of its K nearest neighbors, as illustrated in Figure 5 of [33].

Fuzzy K-NN classification method [101] is a special variation of the K-NN classification family. Instead of roughly assigning the label based on a voting from the K nearest neighbors, it attempts to estimate the membership values that indicate how much degree the query sample belongs to the classes concerned, Obviously, it is impossible for any characteristic description to contain complete information, which would make the classification ambiguous. In view of this, the fuzzy principle is very reasonable and particularly useful in dealing with complicated biological systems, such as identifying nuclear receptor subfamilies [102], characterizing the structure of fast-folding proteins [103], classifying G protein-coupled receptors [104], predicting protein quaternary structural attributes [105], predicting protein structural classes [106, 107], and so forth.

Next, let us give a brief introduction how to use the fuzzy$K$-NN approach to identify the interactions between the enzymes and the drug compounds in the network concerned.

Supposing that $\mathbb{S}(N) = \{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_N\}$ is a set of vectors representing N enzyme-drug pairs in a training set classified into two classes $\{C^+, C^-\}$, where C^+ denotes the interactive pair class, while C^- the noninteractive pair class; $\mathbb{S}^*(\mathbf{G}) = \{\mathbf{G}_1^*, \mathbf{G}_2^*, \ldots, \mathbf{G}_K^*\} \subset \mathbb{S}(N)$ is the subset of the K nearest neighbor pairs to the query pair $\mathbf{G}$. Thus, the fuzzy membership value for the query pair $\mathbf{G}$ in the two classes of $\mathbb{S}(N)$ is given by

$$\begin{aligned}
\mu^+(\mathbf{G}) &= \frac{\sum_{j=1}^{K} \mu^+\left(\mathbf{G}_j^*\right) d\left(\mathbf{G}, \mathbf{G}_j^*\right)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}}, \\
\mu^-(\mathbf{G}) &= \frac{\sum_{j=1}^{K} \mu^-\left(\mathbf{G}_j^*\right) d\left(\mathbf{G}, \mathbf{G}_j^*\right)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} d\left(\mathbf{G}, \mathbf{G}_j^*\right)^{-2/(\varphi-1)}},
\end{aligned} \tag{27}$$

where K is the number of the nearest neighbors counted for the query pair $\mathbf{G}$; $\mu^+(\mathbf{G}_j^*)$ and $\mu^-(\mathbf{G}_j^*)$, the fuzzy membership values of the training sample $\mathbf{G}_j^*$ to the class C^+ and C^-, respectively, as will be further defined next; $d(\mathbf{G}, \mathbf{G}_j^*)$, the cosine distance between $\mathbf{G}$ and its jth nearest pair $\mathbf{G}_j^*$ in the training dataset $\mathbb{S}(N)$; $\varphi(> 1)$, the fuzzy coefficient for determining how heavily the distance is weighted when calculating each nearest neighbor's contribution to the membership value. Note that the parameters K and φ will affect the computation result of (27), and they will be optimized by a grid-search as will be described later. Also, various other metrics can be chosen for $d(\mathbf{G}, \mathbf{G}_j^*)$, such as Euclidean distance, Hamming distance [108], and Mahalanobis distance [55, 109].

The quantitative definitions for the aforementioned $\mu^+(\mathbf{G}_j^*)$ and $\mu^-(\mathbf{G}_j^*)$ in (27) are given by

$$\begin{aligned}
\mu^+\left(\mathbf{G}_j^*\right) &= \begin{cases} 1, & \text{if } \mathbf{G}_j^* \in C^+ \\ 0, & \text{otherwise,} \end{cases} \\
\mu^-\left(\mathbf{G}_j^*\right) &= \begin{cases} 1, & \text{if } \mathbf{G}_j^* \in C^- \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{28}$$

Substituting the results obtained by (27) into (28), it follows that if $\mu^+(\mathbf{G}) > \mu^-(\mathbf{G})$ then the query pair $\mathbf{G}$ is an interactive couple; otherwise, noninteractive. In other words, the outcome can be formulated as

$$\mathbf{G} \in \begin{cases} C^+, & \text{if } \mu^+(\mathbf{G}) > \mu^-(\mathbf{G}) \\ C^-, & \text{otherwise.} \end{cases} \tag{29}$$

If there is a tie between $\mu^+(\mathbf{G})$ and $\mu^-(\mathbf{G})$, the query pair $\mathbf{G}$ will be randomly assigned to one of the two classes. However, case like that is quite rare and in this study never happened.

The predictor, thus, established is called iEzy-Drug, where "i" means identify, and "Ezy-Drug" means the interaction between enzyme and drug. To provide an intuitive overall picture, a flowchart is provided in Figure 2 to show the process of how the classifier works in identifying enzyme-drug interactions.

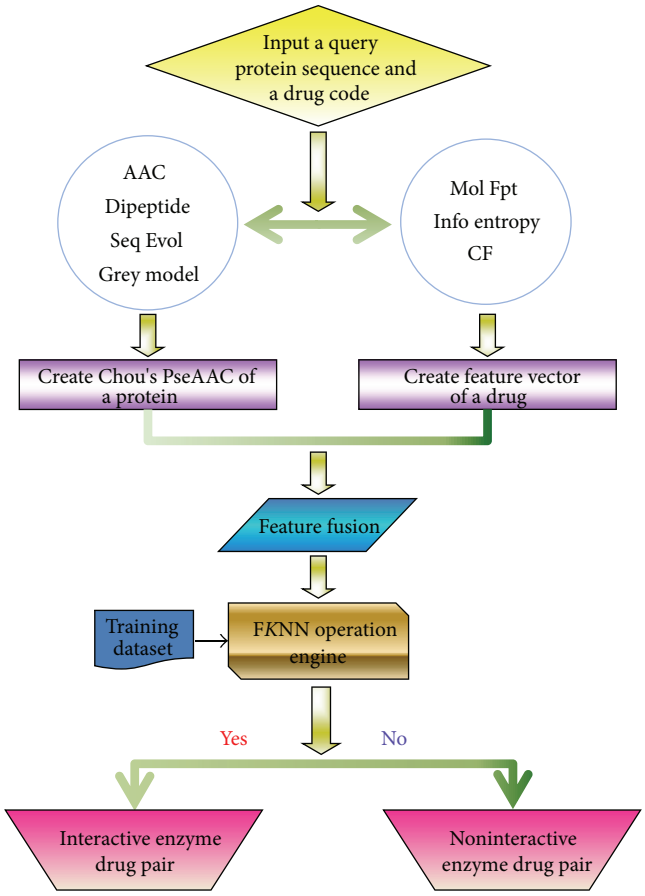


FIGURE 2: A flowchart to show the operation process of the iEzy-Drug predictor. See the text for further explanation.

2.4. Criteria for Performance Evaluation. In the literature, the following equation set is often used for examining the performance quality of a predictor:

$$\begin{aligned} \mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{Acc} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{MCC} &= \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \end{aligned} \tag{30}$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient.

To most biologists, however, the four metrics as formulated in (30) are not quite intuitive and easier-to-understand, particularly for the Mathew's correlation coefficient. Here, let us adopt the Chou's symbols to formulate the previous four metrics. By means of Chou's symbols [111, 112], the rates of correct predictions for the interactive enzyme-drug pairs in dataset $\mathbb{S}^+$ and the noninteractive enzyme-drug pairs in dataset $\mathbb{S}^-$ are, respectively, defined by (cf. (1))

$$\begin{aligned} &\Lambda^+ \\ &= \frac{N^+ - N_-^+}{N^+}, \quad \text{for the interactive enzyme-drug pairs,} \\ &\Lambda^- \\ &= \frac{N^- - N_+^-}{N^-}, \quad \text{for the noninteractive enzyme-drug pairs,} \end{aligned} \tag{31}$$

where N^+ is the total number of the interactive enzyme-drug pairs investigated, while N_-^+ is the number of the interactive enzyme-drug pairs incorrectly predicted as the noninteractive enzyme-drug pairs; N^- is the total number of the noninteractive enzyme-drug pairs investigated, while N_+^- is the number of the noninteractive enzyme-drug pairs incorrectly predicted as the interactive enzyme-drug pairs. The overall success prediction rate is given by [113] as follows:

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}. \tag{32}$$

It is obvious from (31)-(32) that if and only if none of the interactive enzyme-drug pairs and the noninteractive enzyme-drug pairs are mispredicted; that is, $N_-^+ = N_+^- = 0$ and $\Lambda^+ = \Lambda^- = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be smaller than 1.

The relations between the symbols in (32) and those in (30) are given by

$$\begin{aligned} \mathrm{TP} &= N^+ - N_-^+, \\ \mathrm{TN} &= N^- - N_+^-, \\ \mathrm{FP} &= N_+^-, \\ \mathrm{FN} &= N_-^+. \end{aligned} \tag{33}$$

Substituting (33) into (30) and also noting (31)-(32), we obtain

$$\begin{aligned} \mathrm{Sn} &= \Lambda^+ = 1 - \frac{N_-^+}{N^+}, \\ \mathrm{Sp} &= \Lambda^- = 1 - \frac{N_+^-}{N^-}, \\ \mathrm{Acc} &= \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}, \\ \mathrm{MCC} &= \frac{1 - \left(\left(N_-^+/N^+\right) + \left(N_+^-/N^-\right)\right)}{\sqrt{\left(1 + \left(N_+^- - N_-^+\right)/N^+\right)\left(1 + \left(N_-^+ - N_+^-\right)/N^-\right)}}. \end{aligned} \tag{34}$$

Now it is obvious to see from (34): when $N_-^+ = 0$ meaning none of the interactive enzyme-drug pairs was

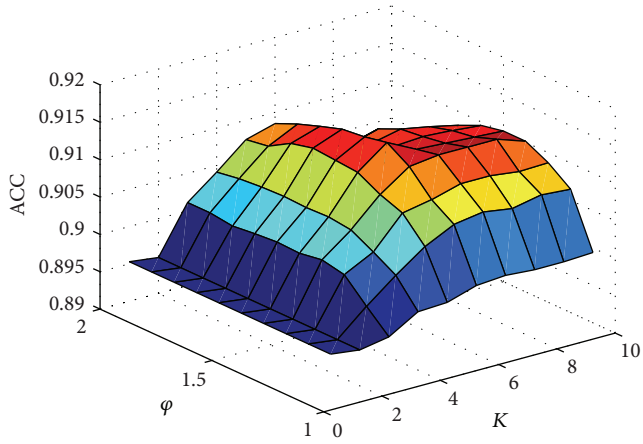


Figure 3: A 3D plot to show how the parameter in (27) was optimized for the iEzy-Drug predictor.

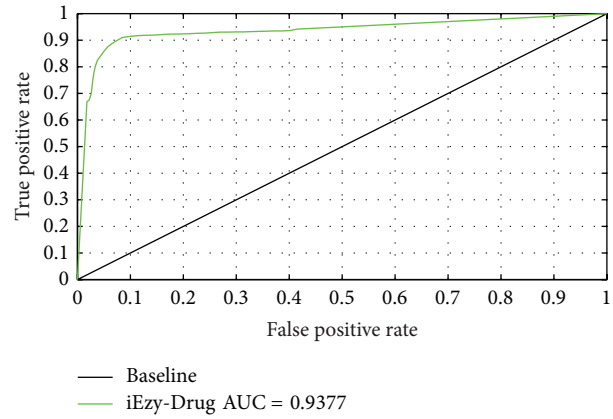


Figure 4: A plot for the ROC curve to quantitatively show the performance of the iEzy-Drug predictor.

mispredicted to be a noninteractive enzyme-drug pair, we have the sensitivity Sn = 1; while $N_-^+ = N^+$ meaning that all the interactive enzyme-drug pairs were mispredicted to be the noninteractive enzyme-drug pairs, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning none of the noninteractive enzyme-drug pairs was mispredicted, we have the specificity Sp = 1; while $N_+^- = N^-$ meaning all the noninteractive enzyme-drug pairs were incorrectly predicted as interactive enzyme-drug pairs, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of the interactive enzyme-drug pairs in the dataset $\mathbb{S}^+$ and none of the noninteractive enzyme-drug pairs in $\mathbb{S}^-$ was incorrectly predicted, we have the overall accuracy Acc = Λ = 1; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the interactive enzyme-drug pairs in the dataset $\mathbb{S}^+$ and all the noninteractive enzyme-drug pairs in $\mathbb{S}^-$ were mispredicted, we have the overall accuracy Acc = Λ = 0. The MCC correlation coefficient is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$ meaning that none of the interactive enzyme-drug pairs in the dataset $\mathbb{S}^+$ and none of the noninteractive enzyme-drug pairs in $\mathbb{S}^-$ were mispredicted, we have MCC = 1; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have MCC = 0 meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$, we have MCC = −1 meaning total disagreement between prediction and observation. As we can see from the previous discussion, it is much more intuitive and easier-to-understand when using (34) to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient. It is instructive to point out that the metrics as defined in (30) and (34) are valid for single label systems; for multilabel systems, a set of more complicated metrics should be used as given in [114].

3. Results and Discussion

3.1. Cross-Validation. How to properly examine the prediction quality is a key for developing a new predictor and estimating its potential application value. Generally speaking, the following three cross-validation methods are often used to examine a predictor of its effectiveness in practical application: independent dataset test, subsampling or K-fold (such as 5-fold, 7-fold, or 10-fold) test, and jackknife test [108]. However, as elaborated by a penetrating analysis in [115], considerable arbitrariness exists in the independent dataset test. Also, as demonstrated by (27)–(29) in [33], the subsampling test (or K-fold cross-validation) cannot avoid arbitrariness either. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been widely recognized and increasingly utilized by investigators to examine the quality of various predictors (see, e.g., [66, 71, 74, 80]). Accordingly, the success rate by the jackknife test was also used to optimize the two uncertain parameters K and φ in (27). The result, thus, obtained is shown in Figure 3, from which we obtain when $K = 6$ and $\varphi = 1.5$ the iEzy-Drug predictor reaches its optimized status.

The success rates thus obtained by the jackknife test in identifying interactive Enzyme-drug pairs or noninteractive enzyme-drug pairs on the benchmark dataset $\mathbb{S}$ (cf. Online Supporting Information S1) are given in Table 1, where for facilitating comparison, the corresponding result by He et al. [110] is also given. As we can see from the table, the overall accuracy Acc achieved by iEzy-Drug was 91.03%, remarkably higher than 85.48%, the corresponding rate obtained by He et al. [110] on the same benchmark. Furthermore, listed in Table 1 are also the values obtained by iEzy-Drug for the other three metrics; that is, Sn = 90.81%, Sp = 91.14%, and MCC = 80.39%, indicating that the accuracy of iEzy-Drug is not only very high but also quite stale.

To provide a graphical illustration to show the performance of the current binary classifier iEzy-Drug as its discrimination threshold is varied, a 2D plot, called Receiver Operating Characteristic (ROC) curve [116, 117], was also given (Figure 4). In the ROC curve, the vertical coordinate Y is for the true positive rate or Sn (cf. (34)), while horizontal coordinate X for the false positive rate

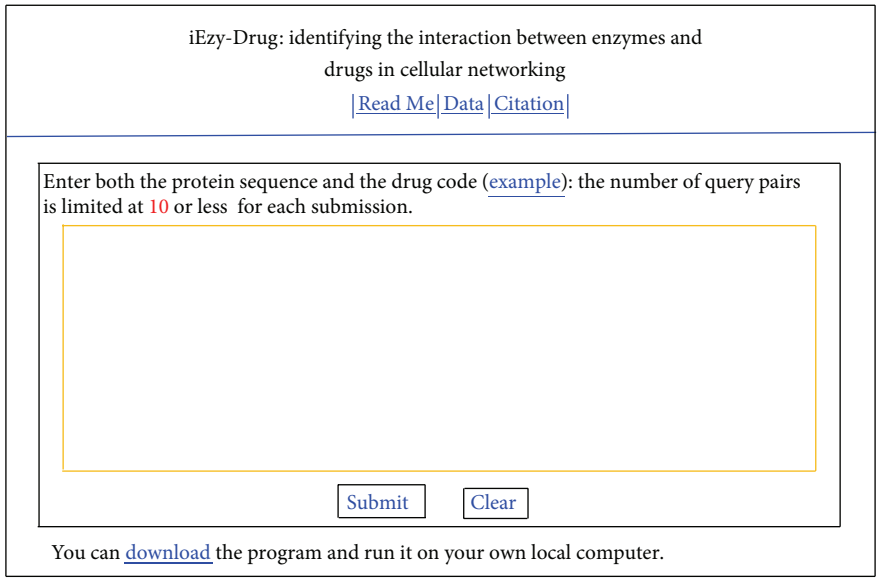


Figure 5: A semiscreenshot to show the top page of the iEzy-Drug web-server. Its web-site address is at http://www.jci-bioinfo.cn/iEzy-Drug/.

Table 1: The jackknife success rates obtained with iEzy-Drug in identifying interactive enzyme-drug pairs and noninteractive enzyme-drug pairs for the benchmark dataset $\mathbb{S}$ (cf. Online Supporting Information S1).

Method	Acc	Sn	Sp	MCC
iEzy-Drug[a]	7425/8157 = 91.03%	2469/2719 = 90.81%	4956/5438 = 91.14%	80.39%
NN predictor[b]	85.48%	N/A	N/A	N/A

[a]See (27) where the parameters $K = 6$ and $\varphi = 1.5$.
[b]See [110].

or 1-Sp. The best possible prediction method would yield a point with the coordinate (0, 1) representing 100% true positive rate (sensitivity Sn) and 0 false positive rate or 100% specificity. Therefore, the (0, 1) point is also called a perfect classification. A completely random guess would give a point along a diagonal from the point (0, 0) to (1, 1). The area under the ROC curve, also called Area Under the ROC (AUROC), is often used to indicate the performance quality of a binary classifier; the value 0.5 of AUROC is equivalent to random prediction, while 1 of AUROC represents a perfect one. As we can see from Figure 4, the AUROC value obtained by iEzy-Drug is 0.9377.

The reason why iEzy-Drug can remarkably improve the prediction quality is that it has introduced the 2D molecular fingerprints to represent drug samples see Online Supporting Information S3 for the detailed fingerprint expressions for the drugs listed in Online Supporting Information S1 and that it has successfully used PseAAC to incorporate the features derived from the sequences of enzymes that are essential for identifying the interaction of enzymes with drugs in the cellular networking.

To enhance the value of its practical applications, the web server for iEzy-Drug has been established that can be freely accessible at http://www.jci-bioinfo.cn/iEzy-Drug/. It is anticipated that the web server will become a useful high throughput tool for both basic research and drug development in the relevant areas, or at the very least play a complementary role to the existing method [39, 110, 118] for which so far no web-server whatsoever has been provided yet.

3.2. The Protocol or User Guide. 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 means of the 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/iEzy-Drug/ and you will see the top page of the predictor on your computer screen, as shown in Figure 5. Click on the Read Me button to see a brief introduction about iEzy-Drug predictor and the caveat when using it.

Step 2. Either type or copy/paste the query pairs into the input box at the center of Figure 5. Each query pair consists of two parts: one is for the protein sequence and the other for the drug. The enzyme sequence should be in FASTA format, while the drug in the KEGG code. Examples for the query pairs input can be seen by clicking on the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the four query pairs in the Example

window as the input, after clicking the Submit button, you will see on your screen that the "hsa: 10056" enzyme and the "D0021" drug are an interactive pair, and that the "hsa: 100" enzyme and the "D0037" drug are also an interactive pair, but that the "hsa: 3295" enzyme and the "D00889" drug are not an interactive pair, and that the "hsa: 7366" enzyme and the "D03601" drug are not an interactive pair either. All these results are fully consistent with the experimental observations. It takes about 3 minutes before the results are shown on the screen.

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

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

Step 6. The program code is also available by clicking the button download on the lower panel of Figure 5.

Acknowledgments

The authors would like to thank 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 Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (no. 20120BDH80023), Natural Science Foundation of Jiangxi Province, China (no. 2010GZS0122, 20122BAB201020), the Department of Education of Jiangxi Province (GJJ12490), the LuoDi plan of the Department of Education of Jiangxi Province (KJLD12083), 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 paper.

References

[1] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 304–305, 2000.

[2] S. H. Koenig and R. D. Brown, "H_2CO_3 as substrate for carbonic anhydrase in the dehydration of $H_2CO_3(-)$," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 69, no. 9, pp. 2422–2425, 1972.

[3] S. X. Lin and J. Lapointe, "Theoretical and experimental biology in one," *Journal of Biomedical Science and Engineering*, vol. 6, pp. 435–442, 2013.

[4] K. C. Chou and S. P. Jiang, "Studies on the rate of diffusion-controlled reactions of enzymes. Spatial factor and force field factor," *Scientia Sinica*, vol. 17, no. 5, pp. 664–680, 1974.

[5] K. C. Chou, "The kinetics of the combination reaction between enzyme and substrate," *Scientia Sinica*, vol. 19, no. 4, pp. 505–528, 1976.

[6] K. C. Chou and G. P. Zhou, "Role of the protein outside active site on the diffusion-controlled reaction of enzyme," *Journal of the American Chemical Society*, vol. 104, no. 5, pp. 1409–1413, 1982.

[7] R. A. Poorman, A. G. Tomasselli, R. L. Heinrikson, and F. J. Kezdy, "A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base," *The Journal of Biological Chemistry*, vol. 266, no. 22, pp. 14554–14561, 1991.

[8] K.-C. Chou, "A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins," *The Journal of Biological Chemistry*, vol. 268, no. 23, pp. 16938–16948, 1993.

[9] G. Z. Liang and S. Z. Li, "A new sequence representation as applied in better specificity elucidation for human immunodeficiency virus type 1 protease," *Biopolymers*, vol. 88, no. 3, pp. 401–412, 2007.

[10] J. J. Chou, "Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach," *Journal of Protein Chemistry*, vol. 12, no. 3, pp. 291–302, 1993.

[11] Q. S. Du, S. Wang, D. Q. Wei, S. Sirois, and K. C. Chou, "Molecular modeling and chemical modification for finding peptide inhibitor against severe acute respiratory syndrome coronavirus main proteinase," *Analytical Biochemistry*, vol. 337, no. 2, pp. 262–270, 2005.

[12] Y. Gan, H. Huang, Y. Huang et al., "Synthesis and activity of an octapeptide inhibitor designed for SARS coronavirus main proteinase," *Peptides*, vol. 27, no. 4, pp. 622–625, 2006.

[13] Q. S. Du, H. Sun, and K. C. Chou, "Inhibitor design for SARS coronavirus main protease based on 'distorted key theory,'" *Medicinal Chemistry*, vol. 3, no. 1, pp. 1–6, 2007.

[14] K. C. Chou, "Prediction of human immunodeficiency virus protease cleavage sites in proteins," *Analytical Biochemistry*, vol. 233, no. 1, pp. 1–14, 1996.

[15] J. Knowles and G. Gromo, "Target selection in drug discovery," *Nature Reviews Drug Discovery*, vol. 2, no. 1, pp. 63–69, 2003.

[16] A. C. Cheng, R. G. Coleman, K. T. Smyth et al., "Structure-based maximal affinity model predicts small-molecule druggability," *Nature Biotechnology*, vol. 25, no. 1, pp. 71–75, 2007.

[17] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," *Journal of Molecular Biology*, vol. 261, no. 3, pp. 470–489, 1996.

[18] K. C. Chou, "Review: structural bioinformatics and its impact to biomedical science," *Current Medicinal Chemistry*, vol. 11, no. 16, pp. 2105–2134, 2004.

[19] K. C. Chou, D. Q. Wei, and W. Z. Zhong, "Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS," *Biochemical and Biophysical Research Communications*, vol. 308, pp. 148–151, 2003, (Erratum in: *Biochemical and Biophysical Research Communications*, vol. 310, p. 675, 2003).

[20] K. C. Chou, D. Q. Wei, Q. S. Du, S. Sirois, and W. Z. Zhong, "Progress in computational approach to drug development againt SARS," *Current Medicinal Chemistry*, vol. 13, no. 27, pp. 3263–3270, 2006.

[21] G. P. Zhou and F. A. Troy, "NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure," *Current Protein and Peptide Science*, vol. 6, no. 5, pp. 399–411, 2005.

[22] R. B. Huang, Q. S. Du, C. H. Wang, and K. C. Chou, "An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus," *Biochemical and Biophysical Research Communications*, vol. 377, no. 4, pp. 1243–1247, 2008.

[23] Q. S. Du, R. B. Huang, C. H. Wang, X. M. Li, and K. C. Chou, "Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus," *Journal of Theoretical Biology*, vol. 259, no. 1, pp. 159–164, 2009.

[24] M. J. Berardi, W. M. Shih, S. C. Harrison, and J. J. Chou, "Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching," *Nature*, vol. 476, no. 7358, pp. 109–114, 2011.

[25] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature*, vol. 451, no. 7178, pp. 591–595, 2008.

[26] B. OuYang, S. Xie, M. J. Berardi et al., "Unusual architecture of the p7 channel from hepatitis C virus," *Nature*, vol. 498, pp. 521–525, 2013.

[27] K. Oxenoid and J. J. Chou, "The structure of phospholamban pentamer reveals a channel-like architecture in membranes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 10870–10875, 2005.

[28] M. E. Call, K. W. Wucherpfennig, and J. J. Chou, "The structural basis for intramembrane assembly of an activating immunoreceptor complex," *Nature Immunology*, vol. 11, no. 11, pp. 1023–1029, 2010.

[29] R. M. Pielak, J. R. Schnell, and J. J. Chou, "Mechanism of drug inhibition and drug resistance of influenza A M2 channel," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 18, pp. 7379–7384, 2009.

[30] J. Wang, R. M. Pielak, M. A. McClintock, and J. J. Chou, "Solution structure and functional analysis of the influenza B proton channel," *Nature Structural and Molecular Biology*, vol. 16, no. 12, pp. 1267–1271, 2009.

[31] K. C. Chou, "Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein," *Journal of Proteome Research*, vol. 4, no. 5, pp. 1681–1686, 2005.

[32] K. C. Chou, "Insights from modeling three-dimensional structures of the human potassium and sodium channels," *Journal of Proteome Research*, vol. 3, no. 4, pp. 856–861, 2004.

[33] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[34] W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iLoc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, pp. 634–644, 2013.

[35] X. Xiao, P. Wang, W. Z. Lin, J. H. Jia, and K. C. Chou, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, pp. 168–177, 2013.

[36] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, article e69, 2013.

[37] K. C. Chou, Z. C. Wu, and X. Xiao, "ILoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular BioSystems*, vol. 8, no. 2, pp. 629–641, 2012.

[38] M. Kotera, M. Hirakawa, T. Tokimatsu, S. Goto, and M. Kanehisa, "The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals," *Methods in Molecular Biology*, vol. 802, pp. 19–39, 2012.

[39] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[40] P. Finn, S. Muggleton, D. Page, and A. Srinivasan, "Pharmacophore discovery using the Inductive Logic Programming system PROGOL," *Machine Learning*, vol. 30, no. 2-3, pp. 241–270, 1998.

[41] I. Vogt, D. Stumpfe, H. E. Ahmed, and J. Bajorath, "Methods for computer-aided chemical biology. Part 2: evaluation of compound selectivity using 2D molecular fingerprints," *Chemical Biology and Drug Design*, vol. 70, pp. 195–205, 2007.

[42] H. Eckert and J. Bajorath, "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches," *Drug Discovery Today*, vol. 12, no. 5-6, pp. 225–233, 2007.

[43] S. Laurent, L. V. Elst, and R. N. Muller, "Comparative study of the physicochemical properties of six clinical low molecular weight gadolinium contrast agents," *Contrast Media & Molecular Imaging*, vol. 1, no. 3, pp. 128–137, 2006.

[44] E. Gregori-Puigjané, R. Garriga-Sust, and J. Mestres, "Indexing molecules with chemical graph identifiers," *Journal of Computational Chemistry*, vol. 32, no. 12, pp. 2638–2646, 2011.

[45] B. Ren, "Application of novel atom-type AI topological indices to QSPR studies of alkanes," *Computers and Chemistry*, vol. 26, no. 4, pp. 357–369, 2002.

[46] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: an open chemical toolbox," *Journal of Cheminformatics*, vol. 3, p. 33, 2011.

[47] V. J. Gillet, P. Willett, and J. Bradshaw, "Similarity searching using reduced graphs," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 338–345, 2003.

[48] D. Butina, "Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 4, pp. 747–750, 1999.

[49] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, pp. 3–55, 2001.

[50] V. D. Gusev, L. A. Nemytikova, and N. A. Chuzhanova, "On the complexity measures of genetic sequences," *Bioinformatics*, vol. 15, no. 12, pp. 994–999, 1999.

[51] K. C. Chou and H. B. Shen, "Review: recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[52] S. F. Altschul, "Evaluating the statistical significance of multiple distinct local alignments," in *Theoretical and Computational Methods in Genome Research*, S. Suhai, Ed., pp. 1–14, Plenum, New York, NY, USA, 1997.

[53] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers and Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[54] H. Nakashima, K. Nishikawa, and T. Ooi, "The folding type of a protein is relevant to the amino acid composition," *Journal of Biochemistry*, vol. 99, no. 1, pp. 153–162, 1986.

[55] K. C. Chou and C. T. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions," *The Journal of Biological Chemistry*, vol. 269, no. 35, pp. 22014–22020, 1994.

[56] K.-C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," *Proteins*, vol. 21, no. 4, pp. 319–344, 1995.

[57] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[58] G. P. Zhou, "An intriguing controversy over protein structural class prediction," *Journal of Protein Chemistry*, vol. 17, no. 8, pp. 729–738, 1998.

[59] I. Bahar, A. R. Atilgan, R. L. Jernigan, and B. Erman, "Understanding the recognition of protein structural classes by amino acid composition," *Proteins*, vol. 29, pp. 172–185, 1997.

[60] J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *Journal of Molecular Biology*, vol. 266, no. 3, pp. 594–600, 1997.

[61] G. P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins*, vol. 50, no. 1, pp. 44–48, 2003.

[62] K. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, pp. 246–255, 2001, (Erratum in: *Proteins*, vol. 44, p. 60, 2001).

[63] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[64] D. Zou, Z. He, J. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," *Journal of Computational Chemistry*, vol. 32, no. 2, pp. 271–278, 2011.

[65] M. M. Beigi, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach," *Journal of Structural and Functional Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[66] Y. K. Chen and K. B. Li, "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," *Journal of Theoretical Biology*, vol. 318, pp. 1–12, 2013.

[67] C. Huang and J. Q. Yuan, "A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types," *The Journal of Membrane Biology*, vol. 246, pp. 327–334, 2013.

[68] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5-6, pp. 320–327, 2010.

[69] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[70] M. Khosravian, F. K. Faramarzi, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods," *Protein & Peptide Letters*, vol. 20, pp. 180–186, 2013.

[71] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, pp. 133–137, 2013.

[72] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "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 Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[73] T. H. Chang, L. C. Wu, T. Y. Lee, S. P. Chen, H. D. Huang, and J. T. Horng, "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," *Journal of Computer-Aided Molecular Design*, vol. 27, pp. 91–103, 2013.

[74] S. Zhang, Y. Zhang, H. Yang, C. Zhao, and Q. Pan, "Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies," *Amino Acids*, vol. 34, no. 4, pp. 565–572, 2008.

[75] R. Zia Ur and A. Khan, "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 & Peptide Letters*, vol. 19, pp. 890–903, 2012.

[76] X. Y. Sun, S. P. Shi, J. D. Qiu, S. B. Suo, S. Y. Huang, and R. P. Liang, "Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform," *Molecular BioSystems*, vol. 8, pp. 3178–3184, 2012.

[77] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[78] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[79] H. Mohabatkar, "Prediction of cyclin proteins using Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[80] H. Mohabatkar, M. M. Beigi, and A. Esmaeili, "Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[81] Y. Xu, J. Ding, L. Y. Wu, and K. C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, Article ID e55844, 2013.

[82] W. Chen, H. Lin, P. M. Feng, C. Ding, Y. C. Zuo, and K. C. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, Article ID e47843, 2012.

[83] B. Li, T. Huang, L. Liu, Y. Cai, and K. C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[84] Y. Jiang, T. Huang, C. Lei, Y. F. Gao, Y. D. Cai, and K. C. Chou, "Signal propagation in protein interaction network during colorectal cancer progression," *BioMed Research International*, vol. 2013, Article ID 287019, 9 pages, 2013.

[85] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[86] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, pp. 960–962, 2013.

[87] H. Shen and K. C. Chou, "PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition," *Analytical Biochemistry*, vol. 373, no. 2, pp. 386–388, 2008.

[88] K. C. Chou, "Using pair-coupled amino acid composition to predict protein secondary structure content," *Journal of Protein Chemistry*, vol. 18, no. 4, pp. 473–480, 1999.

[89] W. Liu and K. C. Chou, "Prediction of protein secondary structure content," *Protein Engineering*, vol. 12, no. 12, pp. 1041–1050, 1999.

[90] M. Bhasin and G. P. S. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *The Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262–23266, 2004.

[91] H. Lin and H. Ding, "Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 269, no. 1, pp. 64–69, 2011.

[92] H. Lin and Q. Li, "Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components," *Journal of Computational Chemistry*, vol. 28, no. 9, pp. 1463–1466, 2007.

[93] L. Nanni and A. Lumini, "Combing ontologies and dipeptide composition for predicting DNA-binding proteins," *Amino Acids*, vol. 34, no. 4, pp. 635–641, 2008.

[94] K.-C. Chou, "The convergence-divergence duality in lectin domains of selectin family and its implications," *FEBS Letters*, vol. 363, no. 1-2, pp. 123–126, 1995.

[95] A. A. Schäffer, L. Aravind, T. L. Madden et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[96] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST: a tool for discovery in protein databases," *Trends in Biochemical Sciences*, vol. 23, no. 11, pp. 444–447, 1998.

[97] J. Deng, "Grey entropy and grey target decision making," *Journal of Grey System*, vol. 22, no. 1, pp. 1–24, 2010.

[98] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[99] J.-Y. Shi, S.-W. Zhang, Q. Pan, Y.-M. Cheng, and J. Xie, "Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition," *Amino Acids*, vol. 33, no. 1, pp. 69–74, 2007.

[100] T. Denoeux, "A κ-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[101] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbours algorithm," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, no. 4, pp. 580–585, 1985.

[102] X. Xiao, P. Wang, and K. Chou, "iNR-physchem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix," *PLoS ONE*, vol. 7, no. 2, Article ID e30869, 2012.

[103] I. Roterman, L. Konieczny, W. Jurkowski, K. Prymula, and M. Banach, "Two-intermediate model to characterize the structure of fast-folding proteins," *Journal of Theoretical Biology*, vol. 283, no. 1, pp. 60–70, 2011.

[104] X. Xiao, P. Wang, and K. C. Chou, "GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions," *Molecular BioSystems*, vol. 7, no. 3, pp. 911–919, 2011.

[105] X. Xiao, P. Wang, and K. C. Chou, "Quat-2L: a web-server for predicting protein quaternary structural attributes," *Molecular Diversity*, vol. 15, no. 1, pp. 149–155, 2011.

[106] X. Zheng, C. Li, and J. Wang, "An information-theoretic approach to the prediction of protein structural class," *Journal of Computational Chemistry*, vol. 31, no. 6, pp. 1201–1206, 2010.

[107] H. Shen, J. Yang, X. Liu, and K. C. Chou, "Using supervised fuzzy clustering to predict protein structural classes," *Biochemical and Biophysical Research Communications*, vol. 334, no. 2, pp. 577–581, 2005.

[108] K.-C. Chou and C.-T. Zhang, "Review: prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[109] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, pp. 49–55, 1936.

[110] R. M. Centor, "Signal detectability: the use of ROC curves and their analyses," *Medical Decision Making*, vol. 11, no. 2, pp. 102–106, 1991.

[111] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering*, vol. 14, no. 2, pp. 75–79, 2001.

[112] K. C. Chou, "Prediction of protein signal sequences and their cleavage sites," *Proteins*, vol. 42, pp. 136–139, 2001.

[113] K. C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[114] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, pp. 1092–1100, 2013.

[115] K. C. Chou and H. B. Shen, "Cell-PLoc 2. 0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Science*, vol. 2, pp. 1090–1103, 2010.

[116] M. Gribskov and N. L. Robinson, "Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching," *Computers and Chemistry*, vol. 20, no. 1, pp. 25–33, 1996.

[117] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.

[118] Z. He, J. Zhang, X. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS ONE*, vol. 5, no. 3, Article ID e9603, 2010.