OPEN ACCESS Freely available online

PLoS one

Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities

Lei Chen[1], Wei-Ming Zeng[1], Yu-Dong Cai[2,5]*, Kai-Yan Feng[3,4], Kuo-Chen Chou[5]*

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

Abstract

The Anatomical Therapeutic Chemical (ATC) classification system, recommended by the World Health Organization, categories drugs into different classes according to their therapeutic and chemical characteristics. For a set of query compounds, how can we identify which ATC-class (or classes) they belong to? It is an important and challenging problem because the information thus obtained would be quite useful for drug development and utilization. By hybridizing the informations of chemical-chemical interactions and chemical-chemical similarities, a novel method was developed for such purpose. It was observed by the jackknife test on a benchmark dataset of 3,883 drug compounds that the overall success rate achieved by the prediction method was about 73% in identifying the drugs among the following 14 main ATC-classes: (1) alimentary tract and metabolism; (2) blood and blood forming organs; (3) cardiovascular system; (4) dermatologicals; (5) genitourinary system and sex hormones; (6) systemic hormonal preparations, excluding sex hormones and insulins; (7) anti-infectives for systemic use; (8) antineoplastic and immunomodulating agents; (9) musculoskeletal system; (10) nervous system; (11) antiparasitic products, insecticides and repellents; (12) respiratory system; (13) sensory organs; (14) various. Such a success rate is substantially higher than 7% by the random guess. It has not escaped our notice that the current method can be straightforwardly extended to identify the drugs for their 2^{nd}-level, 3^{rd}-level, 4^{th}-level, and 5^{th}-level ATC-classifications once the statistically significant benchmark data are available for these lower levels.

Citation: Chen L, Zeng W-M, Cai Y-D, Feng K-Y, Chou K-C (2012) Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities. PLoS ONE 7(4): e35254. doi:10.1371/journal.pone.0035254

Editor: Ozlem Keskin, Koç University, Turkey

Received November 8, 2011; **Accepted** March 14, 2012; **Published** April 13, 2012

Funding: This contribution is supported by National Basic Research Program of China (2011CB510102, 2011CB510101), National Natural Science Foundation of China (No. 31170952), Innovation Program of Shanghai Municipal Education Commission (No. 11ZZ143, No. 12YZ120, No. 12ZZ087) and Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: cai_yud@yahoo.com.cn (YDC); kcchou@gordonlifescience.org (KCC)

Introduction

Nowadays, the **A**natomical **T**herapeutic **C**hemical (ATC) classification system, recommended by the World Health Organization (WHO), is the most widely recognized classification system for drugs. This classification system divides drugs into different groups according to the organ or system on which they act and/or their therapeutic and chemical characteristics. Accordingly, the ATC classification is very helpful for studying utilization of drugs and categorizing them according to different purposes, therapeutic properties, chemical and pharmacological properties (see Report of the WHO Expert Committee, 2005; World Health Organ Tech Rep, Ser:1–119). In the ATC classification system, drugs are classified into 14 main classes (http://www.whocc.no/atc/structure_and_principles/). In order to understand this kind of complicated classification system, some efforts have been made [1,2]. In a pioneer study, Gurulingappa et al. [2] proposed a method to study the ATC-classification system by combining the information extraction and machine learning techniques. However, their method can be used to identify the drug compounds only within the class of "Cardiovascular System", one of the 14 main ATC classes.

During the past decade, many compound databases, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) [3,4], have been established. From these databases many compounds and their properties can be acquired. Such abundant informations provide an opportunity to analyze ATC classification system in greater detail. Encouraged by the successes of using machine learning and data mining methods to investigate complicated problems in a variety of biological areas [5,6,7,8,9], the present study was initiated in an attempt to develop a powerful method by which one can identify query drugs compound among all their 14 posible main classes.

According to a recent comprehensive review [10], to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the samples concerned with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-

validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, let us describe how to deal with these steps one by one.

Materials and Methods

Recently, the information of protein-protein interactions have been used for predicting various attributes of proteins (see, e.g., [11,12,13]), implying that interactive proteins are more likely to share common biological functions [11] than non-interactive ones [14]. Likewise, it is more likely that two interactive drug compounds may have the similar biological function. Actually, it is generally accepted that compounds with similar physicochemical properties often involve in similar biological activities [1]. Accordingly, it is reasonable to assume that the interactive drugs may likely belong to the same ATC-class, and so do those drugs with similar structures. Based on such rational, let us construct the following benchmark to develop a new method for identifying the ATC-classes of drugs.

Benchmark Dataset

The dataset for drugs was obtained from the public available database KEGG [3,4] at ftp://ftp.genome.jp/pub/kegg/medicus/drug/drug (June, 2011). There are totally 9,758 drugs. After excluding those without the information of ATC-codes, the remaining are 4,376 drug samples, from which further screening was performed to remove those without the information of both chemical-chemical interactions and chemical-chemical similarities. After the above winnowing procedures, we finally obtained the benchmark dataset $\mathbb{S}$ containing 3,883 drugs classified into 14 main ATC-classes, as can be formulated by

$$\mathbb{S}=\mathbb{S}_1\cup\mathbb{S}_2\cup\mathbb{S}_3\cup\mathbb{S}_4\cup\mathbb{S}_5\cup\mathbb{S}_6\cup\cdots\cup\mathbb{S}_{14} \qquad (1)$$

where $\mathbb{S}_1$ represents the subset for the 1st main ATC class called "Alimentary tract and metabolism", $\mathbb{S}_2$ the subset for the 2nd main ATC class "Blood and blood forming organs", $\mathbb{S}_3$ the subset for the 3rd main ATC class "Cardiovascular system", and so forth (cf. **Table 1**); while $\cup$ represents the symbol for "union" in the set theory. For convenience, hereafter let us just use C_1, C_2, C_3, …, C_{14} as the tags of the 14 classes. A breakdown of the 3,883 drugs into the 14 main ATC-classes is given in **Table 1**. For the codes of these drugs in each of the 14 classes, see Supporting Information S1. During the course of constructing the benchmark dataset, the information from http://www.genome.jp/kegg-bin/get_htext?br08303.keg was used that collected the drug compounds and their ATC classification information from http://www.whocc.no/atc_ddd_index/ and provided the ATC code for each drug.

Because some drugs may belong to more than one main ATC-class, like the case in dealing with proteins with multiple location sites [15,16,17], it is instructive to introduce the concept of the "virtual drugs" as illustrated as follows. A drug compound belonging to two different ATC-classes will be counted as 2 virtual samples even though they have an identical chemical structure; if belonging to three different classes, 3 virtual samples; and so forth. Accordingly, the total number of the different virtual drug samples is generally greater than that of the total different structural drug samples. Their relationship can be formulated as follows [18]

$$N(\text{vir})=N(\text{struct})+\sum_{\varphi=1}^{M}(\varphi-1)N(\varphi) \qquad (2)$$

where $N(\text{vir})$ is the number of total different virtual drug samples in $\mathbb{S}$, $N(\text{struct})$ the number of total different structural drugs, $N(1)$ the number of drugs belonging to one ACT-class, $N(2)$ the number of drugs belonging to two ATC-classes, and so forth; while M is the number of total main ACT-classes (for the current case, $M=14$ (cf. **Table 1**).

For the current 3,883 drugs in $\mathbb{S}$, 3,295 occur in one class, 370 in two classes, 110 in three classes, 37 in four classes, 27 in five classes, 44 in six classes, and none in seven or more classes (**Figure 1**). Substituting these data into **Eq.1**, we have

$$\begin{aligned}N(\text{vir}) &= N(\text{struct})+(1-1)\times 3295+(2-1)\times 370\\ &\quad+(3-1)\times 110+(4-1)\times 37+(5-1)\times 27\\ &\quad+(6\text{-}1)\times 44+\sum_{L=7}^{14}(L-1)\times 0\\ &=3883+370+220+111+108+220=4912\end{aligned} \qquad (3)$$

which is fully consistent with the figures in **Table 1** and the data in Supporting Information S1.

Prediction Based on Chemical-Chemical Interactions

Based on the fact that the interactive compounds often involve in similar biological activities [11], it is feasible to predict the ATC-class of a query drug using the information of chemical-chemical interactions, as described below.

STITCH (Search tool for interactions of chemicals) [19] is a large database containing known and predicted interactions between chemicals and between proteins derived from experiments, literature and other databases. We downloaded the information of chemical-chemical interactions from http://stitch.embl.de:8080/download/chemical_chemical.links.v2.0.tsv.gz. Each of these interactions was evaluated by a confidence score, ranging from 1 to 1000, to reflect the likelihood of its occurrence. For any two drugs d_1 and d_2, their interaction confidence score was denoted by $Q_i(d_1,d_2)$. Particularly, if the interaction between d_1 and d_2 does not exist in STITCH, their interaction confidence score was set as zero, *i.e.*, $Q_i(d_1,d_2)=0$.

Suppose that a training dataset $\mathbb{S}^{\text{train}}$ consists of n drugs $d_k(k=1,2,\cdots,n)$, and that the 14 main ATC-classes are denoted by $\mathbb{C}=[C_1,C_2,\cdots,C_{14}]$, where C_1 represents "Alimentary tract and metabolism", C_2 "Blood and blood forming organs", and so forth (see **Table 1**). The ATC-classes of any drug d_i can be formulated as

$$C(d_i)=\left\{c_{i,1},\ c_{i,2},\ \cdots,\ c_{i,14}\right\} \qquad (i=1,\ 2,\ \cdots,\ n) \qquad (4)$$

where

$$c_{i,j}=\begin{cases}1, & \text{if } d_i \text{ belongs to } C_j\\ 0, & \text{otherwise}\end{cases} \quad (i=1,2,\cdots,n;\ j=1,2,\cdots,14) \qquad (5)$$

According to the chemical-chemical interaction approach, the likelihood for a query drug d belonging to C_j, denoted as $\Pi^{(i)}\left(d\mapsto C_j\right)$, can be calculated by

$$\Pi^{(i)}\left(d\mapsto C_j\right)=\max_{d_k\in\mathbb{S}^{\text{train}}}Q_i(d,d_k)\cdot c_{k,j} \quad (j=1,2,\cdots,14) \qquad (6)$$

where $d_k\in\mathbb{S}^{\text{train}}$ means that d_k is an element of the training dataset

Table 1. Breakdown of the benchmark dataset $\mathbb{S}$ according to the 14 main ATC classes.

Tag	The 1st-level ATC class	Number of drugs
C_1	Alimentary tract and metabolism	540
C_2	Blood and blood forming organs	133
C_3	Cardiovascular system	591
C_4	Dermatologicals	421
C_5	Genito-urinary system and sex hormones	248
C_6	Systemic hormonal preparations, excluding sex hormones and insulins	126
C_7	Antiinfectives for systemic use	521
C_8	Antineoplastic and immunomodulating agents	232
C_9	Musculo-skeletal system	208
C_{10}	Nervous system	737
C_{11}	Antiparasitic products, insecticides and repellents	127
C_{12}	Respiratory system	427
C_{13}	Sensory organs	390
C_{14}	Various	211
Number of total virtual drugs $N(\mathrm{vir})$		4,912[a]
Number of total structural different drugs $N(\mathrm{struct})$		3,883[b]

[a]See Eqs.2–3 for the definition about the number of virtual drugs, and its relation with the number of structural different drugs.
[b]Of the 3,883 structural different drugs, 3,295 belong to one class, 370 to two classes, 110 to three classes, 37 to four classes, 27 to five classes, and 44 to six classes. See Supporting Information S1 for the detailed drug codes listed in each of 14 ATC-classes.
doi:10.1371/journal.pone.0035254.t001

$\mathbb{S}^{\mathrm{train}}$. According **Eq.6**, the likelihood that d belongs to C_j can be formulated as the maximum of the interaction confidence scores between d and those drugs that belong to C_j in the training dataset $\mathbb{S}^{\mathrm{train}}$. Obviously, the larger the score is, the more likely that d belongs to C_j. When $\Pi^{(i)}\left(d \mapsto C_j\right)=0$, it means that the probability for the drug d belonging to the class C_j is zero. Given a query drug compound d, suppose the outcome derived from **Eq.6** is

$$\begin{aligned} &\Pi^{(i)}(d \mapsto C_8) > \Pi^{(i)}(d \mapsto C_1) > \Pi^{(i)}(d \mapsto C_2) > \cdots \\ &> \Pi^{(i)}(d \mapsto C_{14}) > 0 \end{aligned} \tag{7}$$

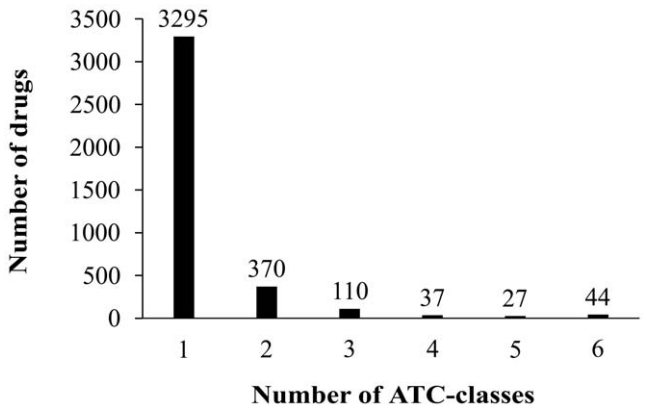

Figure 1. An illustration to show the distribution about the numbers of ATC-classes a same drug may belong to. For the 3,883 drugs in $\mathbb{S}$, 3,295 belong to one class, 370 to two classes, 110 to three classes, 37 to four classes, 27 to five classes, 44 to six classes, and none to seven or more classes.
doi:10.1371/journal.pone.0035254.g001

which means that the highest probability for the drug d belonging to the ATC-class is C_8 (''Antineoplastic and immunomodulating agents''), followed by C_1 (''Alimentary tract and metabolism''), and so forth (cf. **Table 1**). If there is a tie between two terms in **Eq.7**, then the probabilities for the drug belonging to the two corresponding classes are the same. But this kind of tie case rarely happened.

Note that the outcome of **Eq.6** might turn out to be trivial, i.e.,

$$\Pi^{(i)}\left(d \mapsto C_j\right)=0 \quad (j=1,2,\cdots,14) \tag{8}$$

indicating that no chemical-chemical interaction exists for the query drug d in the training dataset $\mathbb{S}^{\mathrm{train}}$; i.e.,

$$Q_i(d,d_k)=0 \quad (\text{for } d_k \in \mathbb{S}^{\mathrm{train}} \text{ or } k=1,2,\ \cdots,\ n) \tag{9}$$

Under such a circumstance, no meaningful result would be obtained by the ''interaction-based'' method, and we should instead use the ''similarity-based method as described in the next section.

Prediction Based on Chemical-Chemical Similarities

Likewise, based on the fact that the compounds with similar physicochemical properties often have the same biological activities [1], we can also use the information of chemical-chemical similarities as another feasible avenue to predict the ATC-class for a query drug. To realize this, let us first introduce how to use graphical representation to measure the similarity between two drug compounds.

Graphical approaches can provide intuitive pictures and useful insights for studying and analyzing complicated biological systems, as demonstrated by many studies on a series of important biological topics (see, e.g., [20,21,22,23,24,25,26,27,28,29,30]). Here, a special graphic approach was utilized to estimate the

similarity of two compounds. Hattori *et al.* [31] first proposed a means to measure the similarity of two compounds via their graph representations. Since each chemical structure can be easily represented by a 2D (two-dimensional) graph where vertices stand for atoms and edges for bonds between them, the similarity of two compounds can be estimated by the Jaccard coefficient [32,33] based on their maximum common subgraph. The similarity scores between compounds by this method can be obtained from the website at http://www.genome.jp/ligand-bin/search_compound. According to the graphical method by Hattori *et al.* [31], given two drug compounds d_1 and d_2, their similarity score was denoted by $Q_s(d_1,d_2)$. When the similarity score between d_1 and d_2 does not exist in http://www.genome.jp/ligand-bin/search_compound, their similarity was set as zero; *i.e.*, $Q_s(d_1,d_2)=0$.

Thus, the prediction method based on the chemical-chemical similarities can be formulated in a way almost completely parallel to that of the chemical-chemical interactions as done in the preceding section.

Now, instead of Eq.6, we have

$$\Pi^{(s)}\left(d\mapsto C_j\right)=\max_{d_k\in\mathbb{S}^{\text{train}}} Q_s(d,d_k)\cdot c_{k,j}\quad (j=1,2,\cdots,14) \qquad (10)$$

where the superscript and subscript "s" stands for the 1st letter of "similarity", implying that the calculation is now based on "chemical-chemical similarity" instead of "chemical-chemical interaction" as done in **Eq.6**.

Prediction by Integrating the Interaction-Based and Similarity-Based Methods

Given a query drug compound d, when the integrated method was used to identify its ATC-class, the prediction involved the following two steps.

Step 1. The interaction-based method (cf. **Eq.6**) was first applied to identify its ATC-class.

Step 2. If the probabilities thus obtained for the drug belonging to all the 14 ATC-classes were zero as indicated in **Eq.8**, meaning no meaningful results were obtained at all, then the prediction would continue using the similarity-based method (cf. **Eq.10**).

Jackknife Cross-Validation

In statistical prediction, the following three cross-validation methods are often used to examine the quality of a predictor: independent dataset test, subsampling (or k-fold crossover) test, and jackknife test [34]. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [35]. The reasons are as follows. (i) For the independent dataset test, although all the samples used to test a predictor are outside the training dataset used to train the prediction engine so as to exclude the "memory" effect or bias, the way of how to select the independent samples for testing the predictor could be quite arbitrary unless the number of independent samples is sufficiently large. This kind of arbitrariness might lead to completely different conclusions. For instance, a predictor achieving a higher success rate than the other for a given independent testing dataset might not able to keep so when tested by another independent testing dataset [34]. (ii) For the subsampling (or k-fold crossover) test, the concrete procedure usually used in literatures was the 5-fold, 7-fold or 10-fold cross-validation. The problem with this kind of subsampling test was that the number of possible selections in dividing a benchmark dataset would be an astronomical figure even for a very simple dataset, as elucidated in [35] and demonstrated by Eqs.28–30 in [10]. Therefore, in any practical subsampling cross-validation tests, only an extremely small fraction of the possible selections were taken into account. Since different selections would always yield different results even for a same benchmark dataset and a same predictor, the subsampling test could not avoid the arbitrariness either. A test method unable to generate a unique outcome should not be deemed as a good one. (iii) In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling (or k-fold crossover) test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. Accordingly, the jackknife test has been widely recognized and increasingly adopted by many investigators to examine the quality of various predictors (see, e.g., [36,37,38,39,40,41,42,43,44,45,46,47]). Accordingly, in this study we are to use the jackknife test to examine the prediction quality as well.

Accuracy Measurement

For any given set of query drugs, we can obtain a series of candidate ATC-classes using the aforementioned prediction methods. Ranked by the likelihood according to their descending order, the prediction accuracy can be defined as

$$AC_j=\frac{CP_j}{N}\qquad (j=1,2,\ \cdots,\ 14) \qquad (11)$$

where CP_j denotes the number of drugs whose j-th order predicted ATC-class is one of the true ATC-class, and N denotes the total number of query drugs whose ATC-classes are to be identified. According to such a definition, the result of higher AC_j with smaller j or lower AC_j with larger j indicates that the predicted hits are more concentrated meaning a better prediction. Obviously, the result with high 1st-order prediction accuracy AC_1 always represents a good quality of prediction.

The average number of ATC-classes for the N query drugs is defined as

$$AN=\frac{\sum_{i=1}^{N} T_i}{N} \qquad (12)$$

where T_i is the number of ATC-classes for the i-th query drug. Thus, another parameter for measuring the proportion of the true classes successfully identified by the first m-order prediction hits can be calculated as [13]

$$L_m=\frac{\sum_{i=1}^{N} P_{i,m}}{\sum_{i=1}^{N} T_i} \qquad (13)$$

where $P_{i,m}$ denotes the number of the first m predicted candidate ATC-classes that are the true ATC-classes for the i-th drug in the dataset. Usually, m could take the smallest integer that is equal to or greater than AN; *i.e.*,

$$m=\begin{cases} AN, \text{ if } AN \text{ is an integer} \\ 1+\text{Int}[AN], \text{ otherwise} \end{cases} \quad (14)$$

where the operator Int means taking the integer part of the quantity right after it. Again, the result of larger L_m with smaller m implies a better prediction with less uncertainty.

Results and Discussion

For clarity, the original benchmark dataset $\mathbb{S}$ of 3,883 drugs (cf. Supporting Information S1) can be separated into two subsets; i.e.,

$$\mathbb{S}=\mathbb{S}^{(i)}\bigcup\mathbb{S}^{(s)} \quad (15)$$

where $\mathbb{S}^{(i)}$ contains 2,144 drugs that had the chemical-chemical interaction information, while $\mathbb{S}^{(s)}$ contains $(3{,}883-2{,}144)=1{,}739$ drugs that had no chemical-chemical interaction information. Listed in **Table 2** are the results obtained by the aforementioned three different prediction methods in identifying the 14 main ATC classes for the drugs investigated. By examining the table, we can observe the following.

Performance of the Interaction-Based Method

For the 2,144 drugs in $\mathbb{S}^{(i)}$ we could use **Eq.6** to conduct the prediction. The results thus obtained are listed in column 2 of **Table 2**, from which we can see that the 1st-order prediction by the jackknife test on the 2,114 drugs was 67.72%. The success rates generally followed a descending trend with increasing of the order number, indicating that the predicted ATC-classes were well sorted for each of the samples investigated. The average number of the ATC-classes in $\mathbb{S}^{(i)}$ was $AN=2664/2144=1.24$ (see **Eq.12**).

Table 2. The jackknife success rates by three different methods in identifying the drugs among the 14 main ATC-classes.

Prediction order	Interaction-based[a]	Similarity-based[b]	Integrated[c]
1	67.72%	78.49%	72.55%
2	21.13%	18.86%	20.11%
3	13.43%	8.63%	11.28%
4	7.18%	5.23%	6.31%
5	4.76%	2.88%	3.91%
6	3.54%	1.73%	2.73%
7	1.63%	0.12%	0.95%
8	0.75%	0.35%	0.57%
9	0.75%	0.12%	0.46%
10	0.56%	0.06%	0.33%
11	0.09%	0.00%	0.05%
12	0.28%	0.00%	0.15%
13	0.09%	0.00%	0.05%
14	0.05%	0.00%	0.03%

[a]Using **Eq.6** on the 2,144 drugs in the benchmark dataset $\mathbb{S}$ that had the chemical-chemical interaction information.
[b]Using **Eq.10** on the $3{,}883-2{,}144=1{,}739$ drugs in the benchmark dataset $\mathbb{S}$ that had no chemical-chemical interaction information.
[c]Using the integrated method by hybridizing **Eq.6** and **Eq.10** on the 3,883 drugs in the benchmark dataset $\mathbb{S}$ as given in Supporting Information S1.
doi:10.1371/journal.pone.0035254.t002

Thus, it follows according to **Eq.14** that $m=2$, meaning that the first 2-order predictions should be taken into consideration. Substituting these data into **Eq.13**, we obtained the overall success rate by the predictions of the first two orders for the 2,144 drugs in $\mathbb{S}^{(i)}$ was $L_m=71.51\%$, indicating that the interaction-based method is quite promising in identifying the ATC-classed of drugs. However, this method could only be used to deal with those drugs that had the chemical-chemical interaction information.

Performance of Similarity-Based Method

For the remaining 1,739 drugs in the dataset $\mathbb{S}^{(s)}$ (cf. **Eq.15**) that did not have the chemical-chemical information, the similarity-based method (cf. **Eq.10**) was used as a backup, and the results thus obtained are shown in column 3 of **Table 2**. It can be seen from there that the 1st-order prediction by the jackknife test on the 1,739 drugs was 78.49%. The average number of ATC-classes for the drugs in $\mathbb{S}^{(s)}$ was $AN=2248/1739=1.29$ (see **Eq.12**), and hence we have $m=2$ (**Eq.14**), meaning that the first 2-order predictions should be taken into account. Substituting these data into **Eq.13**, we obtained the overall success rate by the first two orders predictions for the 1,739 drugs without the chemical-chemical interaction information was 75.31%, indicating that the similarity-based method was quite promising as well.

At a first glance at **Table 2**, it looks like that the success rates by the similarity-based method (**Eq.10**) are higher than those by the interaction-based method (**Eq.6**). However, since the success rates by the two methods as reported in **Table 2** were derived from two different datasets, $\mathbb{S}^{(i)}$ and $\mathbb{S}^{(s)}$ (cf. **Eq.15**) respectively, they might not able to reflect the true superiority between the two methods. To make a comparison between them in a more fair manner, let us construct a new dataset, denoted as $\mathbb{S}^{(i+s)}$. It consists of 2,138 drugs with each containing both chemical-chemical interaction and chemical-chemical similarity informations. The details of such a dataset is given in Supporting Information S2.

Listed in **Table 3** are the results obtained by the methods in identifying the 14 main ATC classes for the 2,138 drugs in the $\mathbb{S}^{(i+s)}$dataset. As we can see from the table, the 1st-order prediction accuracy by the interaction-based method was 67.40%, while that by the similarity-based method was 40.36%.

The average number of ATC-classes for the drugs in $\mathbb{S}^{(i+s)}$ was 1.24 (see **Eq.12**), and hence we have $m=2$ (**Eq.14**), meaning that the first 2-order predictions should be taken into account. Substituting these data into **Eq.13**, we obtained the overall success rate by the 1st two orders predictions for the 2,138 drugs in $\mathbb{S}^{(i+s)}$ by the interaction-based method (**Eq.6**) was 71.26%, while that by the similarity-based method (**Eq.10**) was only 43.69%, indicating that the interaction-based method is superior to the similarity-based method in identifying the ATC-classes of drugs. That is why in the integrated method the first step was to use the interaction method (**Eq.6**) to identify the ATC-classes for any query drugs. When, and only when no meaningful result was obtained by the interaction-based method, was the similarity-based method (**Eq.10**) used as a backup to continue the prediction (see the Section of "Prediction by Integrating the Interaction-Based and Similarity-Based Methods").

Performance of Integrated Prediction Method

Shown in the 4th column of **Table 2** are the results obtained by the integrated method in identifying the 14 main ATC classes for the 3,883 drugs in the benchmark dataset $\mathbb{S}$. As we can see there, the 1st-order prediction accuracy was 72.55%. The average numbers of ATC-classes for the drugs in $\mathbb{S}$ was $AN=4912/3883=1.27$ (see **Eq.12**). Thus, it follows according to **Eq.14** that $m=2$, meaning that the first 2-order predictions

Table 3. A comparison between the similarity-based method (Eq.10) and the interaction-based method (Eq.6) in identifying the 2,138 drugs in the $\mathbb{S}^{(i+s)}$ dataset (cf. Supporting Information S2).

Prediction order	Similarity-based	Interaction-based	Difference
1	40.36%	67.40%	27.04%
2	13.89%	21.09%	7.20%
3	9.17%	13.47%	4.30%
4	5.99%	7.16%	1.17%
5	3.32%	4.91%	1.59%
6	2.76%	3.46%	0.70%
7	0.65%	1.54%	0.89%
8	0.23%	0.75%	0.52%
9	0.09%	0.75%	0.66%
10	0.05%	0.56%	0.51%
11	0.05%	0.09%	0.04%
12	0.00%	0.33%	0.33%
13	0.09%	0.09%	0.00%
14	0.05%	0.05%	0.05%

doi:10.1371/journal.pone.0035254.t003

should be taken into consideration. Substituting these data into **Eq.13**, we obtained the overall success rate by the first two orders predictions for the drugs in $\mathbb{S}$ was 73.25%.

These results indicate that the integrated method performed quite well in identifying drugs among their 14 main ATC-classes, and that more attention should be paid to the results hit by the first two order predictions because they covered more than 70% of the true ATC-classes.

Finally, it is instructive to point out that although the above demonstrations were given for identifying query drug compounds among their main (or 1st level) classification, the method developed here can be straightforwardly extended to cover the 2nd, 3rd, 4th, 5th or any lower-level classification as long as the corresponding statistically significant datasets for training the predictor are available.

Supporting Information

Supporting Information S1 List of the 4,376 drugs in the ATC classification system extracted from KEGG.
(PDF)

Supporting Information S2 This dataset $\mathbb{S}^{(i+s)}$ contains 2,138 drugs classified into 14 main ATC classes. Each of the drugs listed here contains both chemical-chemical interaction and chemical-chemical similarity informations. Among the 2,138 different drugs (2,655 virtual drugs), 1,838 belong to one class; 190 to two classes; 57 to three classes, 19 to four classes, 14 to five classes, and 20 to six classes. None of the drugs listed here belongs to seven and more classes.
(PDF)

Acknowledgments

The authors are very much indebted to the Academic Editor for taking time from her busy schedule to edit our paper. Many thanks are also due to the two anonymous experts for their constructive comments, which were very helpful for strengthening the presentation of this paper.

Author Contributions

Conceived and designed the experiments: LC WMZ YDC KCC. Performed the experiments: LC WMZ. Analyzed the data: LC WMZ KYF KCC. Contributed reagents/materials/analysis tools: LC YDC. Wrote the paper: LC KYF KCC.

References

1. Dunkel M, Günther S, Ahmed J, Wittig B, Preissner R (2008) SuperPred: drug classification and target prediction. Nucleic acids research 36: W55–W59.
2. Gurulingappa H, Kolářik C, Hofmann-Apitius M, Fluck J (2009) Concept-based semi-automatic classification of drugs. Journal of chemical information and modeling 49: 1986–1992.
3. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27–30.
4. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic acids research 38: D355–D360.
5. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol44, 60) 43: 246–255.
6. Cai YD, Lu L, Chen L, He JF (2010) Predicting subcellular location of proteins using integrated-algorithm method. Molecular Diversity 14: 551–558.
7. Chou KC, Shen HB (2008) ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Comm 376: 321–325.
8. Cai YD, Liu XJ, Xu X, Zhou GP (2001) Support vector machines for predicting protein structural class. BMC bioinformatics 2: 3.
9. Chou KC, Shen HB (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357: 633–640.
10. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273: 236–247.
11. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Molecular systems biology 3: 88.
12. Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks PLoS ONE 5: e10972.
13. Hu L, Huang T, Shi X, Lu WC, Cai YD, et al. (2011) Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS ONE 6: e14556.
14. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, et al. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci U S A 101: 2888–2893.
15. Chou KC, Shen HB (2010) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0 PLoS ONE 5: e9931.
16. Wu ZC, Xiao X, Chou KC (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Molecular BioSystems 7: 3287–3297.
17. Chou KC, Wu ZC, Xiao X (2012) iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Molecular Biosystems 8: 629–641.
18. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370: 1–16.
19. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P (2008) STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 36: D684–688.
20. Chou KC, Forsen S (1980) Graphical rules for enzyme-catalyzed rate laws. Biochemical Journal 187: 829–835.
21. Zhou GP, Deng MH (1984) An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochemical Journal 222: 169–176.
22. Chou KC (1989) Graphic rules in steady and non-steady enzyme kinetics. Journal of Biological Chemistry 264: 12074–12079.
23. Chou KC (1990) Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry 35: 1–24.
24. Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Chou KC, et al. (1993) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. Journal of Biological Chemistry 268: 14875–14880.
25. Chou KC, Kezdy FJ, Reusser F (1994) Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Analytical Biochemistry 221: 217–230.

26. Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Canadian Journal of Chemistry 86: 342–357.
27. Chou KC (2010) Graphic rule for drug metabolism systems. Current Drug Metabolism 11: 369–378.
28. Zhou GP (2011) The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. Journal of Theoretical Biology 284: 142–148.
29. Chou KC, Lin WZ, Xiao X (2011) Wenxiang: a web-server for drawing wenxiang diagrams. Natural Science 3: 862–865.
30. Zhou GP (2011) The Structural Determinations of the Leucine Zipper Coiled-Coil Domains of the cGMP-Dependent Protein Kinase I alpha and its Interaction with the Myosin Binding Subunit of the Myosin Light Chains Phosphase. Proteins & Peptide Letters 18: 966–978.
31. Hattori M, Okuno Y, Goto S, Kanehisa M (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. Journal of the American Chemical Society 125: 11853–11865.
32. Jaccard P (1912) THE The Distribution of the Flora in the Alpine Zone. 1. New Phytologist 11: 37–50.
33. Watson GA (1983) An algorithm for the single facility location problem using the Jaccard metric. SIAM Journal on Scientific and Statistical Computing 4: 748–756.
34. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30: 275–349.
35. Chou KC, Shen HB (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science, 2010, 2, 1090–1103). Nature Protocols 3: 153–162.
36. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263: 203–209.
37. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. Journal of Theoretical Biology 257: 17–26.
38. Chou KC, Wu ZC, Xiao X (2011) iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. PLoS One 6: e18258.
39. Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology 281: 18–23.
40. Chou KC, Shen HB (2010) Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. PLoS ONE 5: e11335.
41. Wu ZC, Xiao X, Chou KC (2012) iLoc-Gpos: A Multi-Layer Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Gram-Positive Bacterial Proteins. Protein & Peptide Letters 19: 4–14.
42. Gu Q, Ding YS, Zhang TL (2010) Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. Protein & Peptide Letters 17: 559–567.
43. Lin J, Wang Y (2011) Using a novel AdaBoost algorithm and Chou's pseudo amino acid composition for predicting protein subcellular localization. Protein & Peptide Letters 18: 1219–1225.
44. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214.
45. Xiao X, Wu ZC, Chou KC (2011) iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. Journal of Theoretical Biology 284: 42–51.
46. Lin WZ, Fang JA, Xiao X, Chou KC (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE 6: e24756.
47. Wang P, Xiao X, Chou KC (2011) NR-2L: A Two-Level Predictor for Identifying Nuclear Receptor Subfamilies Based on Sequence-Derived Features. PLoS ONE 6: e23505.