Journal of Molecular Graphics and Modelling 29 (2010) 38–45

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Computational identification of bioactive natural products by structure activity relationship

Xi Zhou [a], Yongquan Li [b], Xin Chen [a,*]

[a] Department of Bioinformatics, Zhejiang University, Hangzhou 310058, PR China
[b] Institute of Biochemistry, Zhejiang University, Hangzhou 310058, PR China

ARTICLE INFO

Article history:
Received 9 December 2009
Received in revised form 12 April 2010
Accepted 18 April 2010
Available online 28 April 2010

Keywords:
Natural product
Structural activity relationship
Bioactive natural compound-likeness
Drug-likeness
Statistical learning

ABSTRACT

Natural products (NPs) have been widely used in traditional medicines and are a valuable source for new drug discovery. However, insufficient knowledge about their molecular mechanisms has limited the scope of their application and hindered the effort to design new drugs from their synergistic action strategies. Thus far, a systematic study of all NP ingredients in a traditional medicine recipe remains impractical. However encouraging results have begun to appear illustrating synergies between several principle active ingredients. In this work, we propose the use of structure activity relationship (SAR) to identify potential active ingredients in natural products, with the aim to facilitate experimental and computational characterizations of their therapeutic mechanisms and synergies. We call this approach the bioactive natural compound-likeness (BNC-likeness) approach, drawing a parallel to the concept of drug-likeness. In cross-validations and independent example tests, our approach displayed 90–92% sensitivity and 85–90% specificity, suggesting its practical usefulness. We also showed that BNC-like compounds were not just drug-like NP ingredients. BNC-like compounds and drug-like chemicals may share different structural characteristics. Therefore, BNC-likeness is a helpful novel conception inviting dedicated research.

1. Introduction

Natural products (NPs), especially in medicinal plants, are a valuable source of lead compounds in modern drug discovery. It was estimated that NP ingredients and their derivatives contributed to approximately one third of the top-selling drugs currently on market [1]. To date, a large number of NP ingredients have been extracted and purified from medicinal plants and other sources. These ingredients have being extensively screened for bioactivities with a broad number of therapeutic indications [2–6]. The therapeutic mechanisms discovered in NP studies were also routinely used to develop novel strategies of therapeutic intervention [7]. However, despite the great value of NPs, their basic and clinical pharmacology were only known to a very limited extent [2,6], which significantly restricted their scope of application in drug discovery.

Bioactive NPs are often mixtures in which no single ingredient can re-produce the same patient response [8,9]. This is arguably one of the biggest difficulties in their mechanism studies. However, it is the accumulating knowledge of synergies between ingredients that has attracted ever-increasing interest from both academic and commercial sectors.

Coordinated therapeutic intervention with multiple agents is not new in modern medicine. Well-known examples include the cocktail therapies for AIDS [10] and the combined chemotherapies for cancers [11]. Such combinations have been widely recognized as providing a helpful approach to treat complex and refractory diseases. Start-up pharmaceuticals also began to screen mixtures of compounds for therapeutic uses [12]. There had been a number of remarkable discoveries. For instance, the anti-psychotic drug chlorpromazine and the anti-protozoal drug pentamidine, neither of which showed any anti-tumor activity, in combination prevented tumor growth more effectively than paclitaxel [12]. This collective effect cannot be explained by what we know of these compounds and their primary therapeutic mechanisms.

In this direction, traditional medicines are arguably the most valuable repository of "clinically" proven mixtures of compounds, many of which have been found to work effectively and safely in cases where modern medicines have failed or proved insufficient to provide a palliative cure [13]. For example, clinical trials showed that sho-saiko-to, an extract of seven Chinese herbs, helped prevent liver cancer in patients with cirrhosis [14]. This was the first noted treatment across the spectrum of all healthcare systems that offered such benefits. Likewise Zemaphyte, a preparation of 10 herbs used for treating certain kinds of skin diseases in Traditional Chinese Medicine (TCM), had produced impressive responses in treating severe, widespread atopic eczema that was resistant to conventional steroidal therapies [15,16]. A most interesting

* Corresponding author. Tel.: +86 571 88208595; fax: +86 571 88208595.
E-mail address: xinchen@zju.edu.cn (X. Chen).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.04.007

example of NP synergy came from the use of Ginsen extract. The same extract was found to exert opposing activities, preventing or promote vascular formation, in tumor cells and wounded tissues respectively, both being beneficial [17]. In these cases and many others, available data suggested that mixtures of NP ingredients produced better clinical results than any single compound [8,9].

Experimental characterizations of the synergies in NP mixtures are difficult, but not impossible. In 2008, the first discussion on the mechanism of a TCM formula, realgar – *Indigo naturalis* formula (RIF), was published on PNAS [18]. The authors demonstrated the interactions among its three principal ingredients. One of them directly attacked a receptor oncoprotein in leukemia cells. The other two ingredients antagonized the toxicity, slowed leukemia cell growth, and enhanced the cellular uptake of the first ingredient by increasing the synthesis of its carrier proteins on cell membrane [18]. This novel mechanism of coordinated therapeutic actions surely offered new possibilities to develop anti-leukemia treatments.

On the other hand, the success story of RIF's mechanism study also revealed the bottle-neck of current NP mechanism studies. Usually, the large number of ingredients in an NP mixture renders it practically impossible to analyze all their interactions. A smaller set of "principal active ingredients", which may re-produce the same or similar patient responses, have to be identified *a priori*. Once these principal ingredients are in hand, demonstrating their synergistic therapeutic mechanisms becomes achievable and relatively simple with the state-of-the-art profiling technologies. Thus far, more than 150,000 compounds had been isolated from over 50,000 higher plant species, leaving the characterization of their bioactivities far behind [19]. The quick development of synthetic chemistry further contributed to the slow-down of NP study [20]. In well-studied medicinal plant extract, such as *Ginkgo biloba*, hundreds of compounds had been identified, but only a small fraction of constituents, e.g. certain terpene lactones and flavonoids, were considered as the main bioactive ingredients [21]. Therefore, a computational approach to focus our experimental study on compounds that are more likely to be bioactive could be very helpful.

In this work, we proposed the use of structure activity relationship (SAR) to predict the potentially bioactive ingredients within an NP mixture. This strategy draws a close parallel to the concept of drug-likeness, which says that, all orally active drugs, regardless of their different therapeutic uses, go through the same absorption, distribution, metabolism and excretion system, and therefore shall display shared structural characteristics distinguishing them from other chemicals [22,23]. The most well-know example of "drug-likeness" was probably the "rule of five" [24]. With more sophisticated machine learning algorithms, later scientists developed more accurate statistical models to predict drug-like compounds [25,26]. Most bioactive NPs are taken orally. Therefore, the same philosophy may apply.

Many studies have reported that NPs and synthetic molecules occupied different chemical spaces, displaying a number of dissimilarities in their molecular properties and structural features [27]. For example, Henkel et al. reported that NPs and synthetic molecules differed in the number of bridgehead atoms and the frequencies of various functional groups [28]. Feher and Schmidt compared NPs, drugs, and molecules originating from combinatorial chemistry. Several average differences were reported in the number of chiral centers, the frequency of aromatic rings, the degree of saturation, and the presence of various heteroatoms [29]. Singh et al. also compared combinatorial libraries, drugs, NPs, and general small molecules in the small molecule repository of NIH (MLSMR). By analyzing scaffolds and fingerprints, clear differences between NPs and other libraries were reported [30]. Lee and Schneider summarized the scaffold architectures of NPs, and found many distinctive ones that may offer new chemical diversity for combinatorial chemistry [31]. Koch et al. analyzed a large database of NPs and organized NP scaffolds in the form of a tree. This information was used to navigate within the scaffold space to identify interesting NP-specific regions [27]. Based on the above differences, many SAR models had been built to distinguish NPs from synthetic molecules or other types of small molecules with high accuracy. For instance, Stahura et al. used a set of descriptors and constructed a Shannon entropy-based model to classify NPs and synthetic molecules [32]. Ertl et al. discussed various recent approaches to analyze and chart the chemical space covered by NPs. A natural products-likeness score was calculated with a Bayesian model [33].

Thus far, with many studies focused on characterizing NPs as a unique group of compounds, efforts to charter the chemical space occupied by bioactive NP ingredients have not yet been reported. On the other hand, drug-likeness studies were usually conducted with all known compounds or with only synthetic chemicals [25]. Considering the unique characteristics of NP ingredients, the SAR model describing "drug-like" molecules may not be precisely the one describing "likely bioactive" NP ingredients. Therefore dedicated efforts to define the "bioactive natural compound-likeness" (BNC-likeness) of an NP ingredient might be necessary.

Below we show evidences supporting the above propositions. We constructed a BNC-likeness model, which was expected to predict ~20% of all NP ingredients as potentially bioactive, in which >90% of the known bioactive NP ingredients were included. This model was not just a general drug-likeness model. The BNC-likeness model and the drug-likeness model probably relied on different structural characteristics to identify potential bioactive NPs or drug-like compounds. It is therefore our hope that this proof-of-concept work would be able to invite further researches charting the space of bioactive NP ingredients, and consequently facilitates the studies of NP synergies, both experimentally and computationally.

2. Methods

2.1. Example dataset

Known examples of bioactive NP ingredients (positive examples) and non-bioactive NP ingredients (negative examples) are required to train an SAR model classifying them. In the Dr. Duke's Phytochemical and Ethnobotanical Database [34], there were 7549 NP ingredients extracted from approximately 2000 species of higher plants. Among them, 790 bioactive ones were also annotated with 3D structures in the NP ingredient database developed by the Developmental Therapeutics Program (DTP) of National Cancer Institute (NCI) [35]. These 790 high-quality examples of bioactive NP ingredients were used as our positive examples.

Unlike bioactive NP ingredients, there was no data source for validated non-bioactive NP ingredients. Therefore, similar to most "drug-likeness" studies [25,26], we used a random set of NP ingredients without known bioactivity as negative examples. This set of negative examples also contained 790 compounds, which did not overlap with our positive examples, had no "DTP_names" (no activity in all DTP screenings and NCI tumor cell screenings) [36], and had no activity record in the Dr. Duke's Phytochemical and Ethnobotanical Database [34]. Therefore, the entire training dataset was consisted of 1580 NP ingredients, half positive, half negative.

2.2. Molecular descriptor

Training structural activity models requires each compound to be represented as a vector of "molecular descriptors", which are real numbers describing structural characteristics of this compound. In this work, we used DRAGON (version 5.4) [37] to compute a total

Table 1
Descriptors generated by the DRAGON software.

No.	Descriptor category	Population
1	Constitutional descriptors	48
2	Topological descriptors	119
3	Walk and path counts	47
4	Connectivity indices	33
5	Information indices	47
6	2D autocorrelations	96
7	Edge adjacency indices	107
8	Burden eigenvalues	64
9	Topological charge indices	21
10	Eigenvalue-based indices	44
11	Randic molecular profiles	41
12	Geometrical descriptors	74
13	RDF descriptors	150
14	3D-MoRSE descriptors	160
15	WHIM descriptors	99
16	GETAWAY descriptors	197
17	Functional group counts	154
18	Atom-centered fragments	120
19	Charge descriptors	14
20	Molecular properties	31

Table 2
The confusion matrix of our SVM model predicting bioactive natural compounds. TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative. All accuracy indicators were given as (mean ± SD).

		Actual	
		Positive	Negative
Predicted	Positive	TP: 712.86 ± 1.38	FP: 77.43 ± 1.18
	Negative	FN: 78.14 ± 1.38	TN: 712.56 ± 1.18

of 1666 descriptors for each compound. These descriptors could be grouped into 20 categories (Table 1). Among the 1666 descriptors, 19 uninformative ones (taking on the same value for all training examples) were removed, resulting in 1647 descriptors to represent each compound.

2.3. Prediction model

A Support Vector Machine (SVM) model was trained to learn the structural activity relationship of bioactive NP ingredients. The SVM algorithm has been well documented elsewhere [38]. For implementation, we used the software package Libsvm (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). After performance comparisons, we chose the radial basis function kernel for its best performance (data not shown). In this setup, two parameters, the regularization parameter C and the kernel width parameter gamma were optimized by a grid search in an empirical range ([0.001, 1000] for C and [3E-5, 4e4] for gamma). The model performance with each parameter set was estimated with a 10-fold cross-validation.

The harmonic mean of sensitivity and specificity, defined here as the F_2-measure, was used to measure the performance of a model. Based on whether a real positive or negative example in the testing set was predicted as positive or negative, the prediction result for each compound can be divided into four categories: True Positive (*TP*), when a real positive example was predicted correctly; True Negative (*TN*), when a real negative example was predicted correctly; False Positive (*FP*), when a real negative example was predicted as positive; and False Negative (*FN*), when a real positive example was predicted as negative. These four indicators are collectively known as the confusion matrix (Table 2). Many accuracy measurements were derived from these indicators, e.g. sensitivity ($TP/(TP+FN)$) and specificity ($TN/(TN+FP)$). Sensitivity reflects the accuracy on positive examples and specificity reflects the accuracy on negative examples. Giving equal weights to sensitivity (*sen*) and specificity (*spe*), their harmonic mean ($F_2 = 2 \times sen \times spe/(sen + spe)$) may provide a balanced view of model accuracy on both types of examples. Unlike the overall accuracy ($(TP+TN)/(TP+TN+FP+FN)$), the F_2-measure was not sensitive to the positive-to-negative ratios in cross-validations (training) and independent test. This was desirable in our application, as we do not have a reliable estimation of the actual bioactive to non-bioactive ratio in all NP ingredients.

2.4. Feature selection

In order to evaluate the importance of different molecular descriptors in bioactive ingredient prediction, we used a derivative of the Decision Tree (DT) algorithm, the RuleSet algorithm, to pick out important descriptors. In DT and its derivative RuleSet, a prune level (parameter "Pruning CF") needs to be adjusted in order to avoid over-fitting. In this work, this parameter was iteratively optimized with our training dataset, so that in a 10-fold cross-validation, the average accuracy observed with the fraction of training data (9/10 dataset in each iteration) would match the average accuracy observed with the fraction of testing data (1/10 dataset in each iteration). For implementation, we used the well-known decision tree toolbox See5 [39] (the commercial version of the C4.5 algorithm). See5 assigned an importance score to each selected descriptor based on its native attribute winnowing process [39]. In all our distribution analyses, the descriptor frequencies were weighted by this importance score. To obtain statistically robust results, we embedded the above descriptor selection process in a bootstrap framework. If the same descriptor was picked out in multiple bootstrap iterations, its importance scores in all iterations were summed up to make its final importance score.

3. Results and discussions

3.1. Example dataset

Known examples of bioactive NP ingredients and non-bioactive NP ingredients are required to train an SAR model recognizing bioactive NP ingredients. As detailed in Section 2, we assembled an example set consisting of 790 bioactive (positive) and 790 non-bioactive (negative) examples.

As an intuitive way to check whether the positive examples are likely separable from the negative examples, we compared several of their structural properties (Table 3). The properties for all NP ingredients in the DTP database [35] and 1965 random drug-like compounds from the World Drug Index (WDI) database [40] were also compared as references. Mean and median values were computed to characterize the distribution of each property in each compound group. Results showed that, generally, our set of negative examples was extremely similar to the set of all NP ingredients, with most property medians identical and others very close. This reflected the fact that our negative examples were randomly chosen from all NP ingredients. Since the fraction of bioactive NP ingredients was small, the randomly chosen negative examples nicely reflected the background properties of all NP ingredients. On the other hand, the properties of our positive examples were clearly different from those of our negative examples and those of all NP ingredients, suggesting the possibility to recognize bioactive NP ingredients by their structural features. In addition, it was observed that the differences between our positive examples and the drug-like molecules were often bigger than the differences between drug-like molecules and NP ingredients, justifying the proposition that bioactive natural compounds may not be just drug-like NP ingredients. Also, it was observed, interestingly, that our examples of bioactive NP ingredients showed an increased number of H-bond donors and acceptors, which was consistent with the

Table 3
Several structural properties of bioactive NP ingredients, non-bioactive NP ingredients, drug-like compounds, and all NP ingredients in the DTP database. The mean/median values are given.

Property	Bioactive NPs (n = 790)	Non-bioactive NPs (n = 790)	Drug-like (n = 1965)	NPs in DTP (n = 13 827)
Molecular weight	362.32/342.4	267.3/247.9	405.7/362.5	248.3/228.6
Number of rings	3.39/3	1.63/1	3.42/3	1.64/1
Number of rotatable bonds	4.15/3	4.99/4	5.21/4	4.17/3
Number of carbon atoms	19.76/19	13.55/12	21.3/20	12.65/12
Number of nitrogen atoms	0.71/0	1.42/1	1.85/1	1.25/1
Number of oxygen atoms	5.34/5	2.32/2	5.21/4	2.37/2
Number of sulfur atoms	0.031/0	0.22/0	0.13/0	0.2/0
Ratio of carbon to all heavy atoms	0.70/0.70	0.72/0.70	0.68/0.71	0.72/0.75
Ratio of nitrogen to all heavy atoms	0.03/0	0.08/0.06	0.06/0.03	0.07/0.06
Ratio of oxygen to all heavy atoms	0.19/0.19	0.12/0.12	0.16/0.14	0.13/0.12
Number of H-bond donor atoms (N and O)	2.20/2	0.99/1	2.64/2	1.13/1
Number of H-bond acceptor atoms (N, O, F)	5.96/5	3.59/3	6.93/6	3.50/3

general belief that bioactive natural compounds had high binding affinities for their specific receptor systems and that their biological actions were often highly selective [29]. The drug-likeness criteria, "rule-of-five", say that orally active drug-like molecules usually have less than five H-bond donors and less than 10 H-bond acceptors. Although bioactive NP ingredients showed an increased number of H-bond donors and acceptors, most bioactive NPs were still compliant with the rule-of-five. The level of compliancy was not significantly different from that of the known drugs. This observation seem to suggest that bioactive NP ingredients were often both orally active (rule-of-five compliant [22]) and highly specific in target binding (more hydrogen bonds [29]). These properties are greatly valued in new drug discovery and development.

3.2. Prediction accuracy

A Support Vector Machine (SVM) model was trained with our example dataset to predict bioactive natural compounds. Its parameters, C and σ, were optimized by a grid search. Model performances during parameter optimization were estimated by 10-fold cross-validations with the F_2-measure (detailed in Section 2). Accuracy of the resulting model was estimated by a more precise 100-iteration bootstrap experiment. In each iteration, 1/10 of the examples were randomly left out and only these 1/10 examples were used to test the resulting model and compute the confusion matrix. After 100 iterations, the mean and standard deviation (SD) of the confusion matrix was calculated (Table 2). The final SVM model was able to correctly recognize $90.12 \pm 1.3\%$ bioactive NP ingredients (sensitivity) and correctly reject $90.21 \pm 1.4\%$ non-bioactive NP ingredients (specificity). Its F_2-measure was estimated to be $90.16 \pm 1.3\%$, and its overall accuracy was estimated to be $90.17 \pm 1.3\%$.

In addition, a receiver operating characteristic (ROC) curve analysis was performed to verify whether the SVM approach was effective. In an ROC curve, sensitivity was plotted against 1-specificity for models trained with all range of parameters. Therefore, the ROC curve analysis is a global and unbiased evaluation of a prediction algorithm itself, as it does not depend on any specific parameter used to train a prediction model [41]. The larger the area under the curve (AUC), the more effective a prediction algorithm will be. As shown in Fig. 1, we had a large AUC of 91.48% for the SVM approach, indicating its strong predictive capability [42].

It was not the intention of this work to find the most accurate model to identify potentially bioactive NP ingredients, but rather to demonstrate that the likelihood of an NP ingredient being bioactive would be reflected in its structural characteristics, and this relationship could be detected by statistical learning algorithms. Therefore, a reasonable null-hypothesis would be that there are no shared structural characteristics in bioactive NP ingredients and our SVM model just randomly classified compounds with respect to the positive to negative ratio in the test data. We simulated this random classifier 10,000 times and obtained the distribution of its F_2-measure. With these data, the null-hypothesis was safely rejected with $p < 1 \times 10^{-13}$.

3.3. Rediscovery of bioactive NP ingredients

The usefulness of our model predicting potentially bioactive NP ingredients was further evaluated with an independent evaluation dataset, which did not overlap with the examples used in model training. When preparing this independent evaluation dataset, we focused on nine widely used medicinal herb families. 81 new positive examples and 81 new negative examples were selected (Table 4). Manual literature check was performed to ensure the quality of this independent evaluation dataset (Supplemental Table S1).

As shown in Table 4 and detailed in Supplemental Table S1, 75 bioactive NP ingredients (92.59%) were successfully recognized and 69 non-bioactive NP ingredients (85.18%) were correctly identified. This sensitivity (92.59%) and specificity (85.18%) were comparable to those observed during model construction (90.12% and 90.21%). The F_2-measure estimated with the independent evaluate dataset (89.01%) was also close to that observed with the training dataset (90.16%). This similarity in accuracy measurements indicated that our model was free from significant over-fitting problems and should be useful in real-world applications.

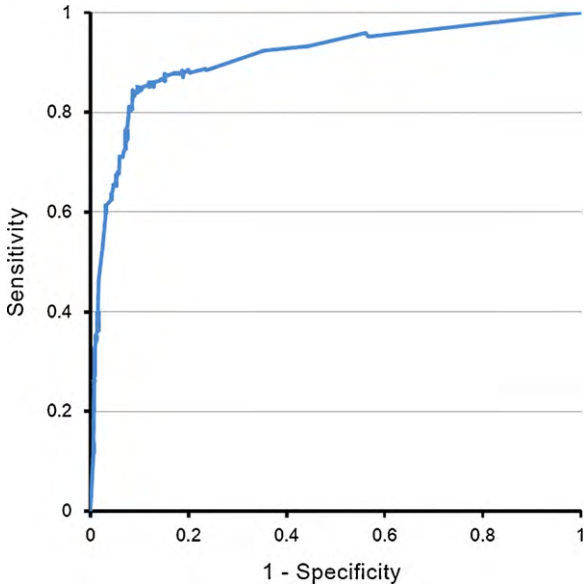

Fig. 1. The receiver operating characteristic curve of SVM models trained with all range of parameters. The area under curve (AUC) was 91.48%.

Table 4
The composition of our independent evaluation examples and the incorrectly predicted ones. Detailed prediction results were given in Supplemental Table S1.

Herb family	Number of positive examples (false negative predictions)	Number of negative examples (false positive predictions)
Apiaceae(Umbelliferae)	10(1)	8(1)
Araliaceae	9(0)	24(2)
Caprifoliaceae	1(0)	6(1)
Compositae	3(1)	10(3)
Ginkgoaceae	25(3)	0
Iridaceae	1(0)	0
Labiatae	23(1)	2(1)
Leguminosae	8(0)	18(3)
Magnoliaceae	1(0)	13(1)

Furthermore, our model was used to identify potentially bioactive compounds in the entire DTP natural product database. As expected, the majority (11156/13824, 80.7%) of the compounds were predicted as non-bioactive. Taken together, our model was expected to predict ~20% of all NP ingredients as potentially bioactive, in which >90% of the known bioactive NP ingredients were included. This accuracy was likely to be useful to focus related biomedical researches on compounds with higher potential.

3.4. Comparison of drug-likeness and bioactive natural compound-likeness

The above results demonstrated that dedicated SAR models, which we call the Bioactive Natural Compound-likeness (BNC-likeness) models, were useful to narrow down the range of potentially bioactive NP ingredients in an NP mixture. As mentioned in Section 1, the concepts of drug-likeness and BNC-likeness are similar. It is therefore interesting to examine whether they are actually the same, or in other words, whether bioactive NP ingredients are just drug-like NP ingredients, which can be recognized accurately by a drug-likeness model. For this purpose, we constructed a typical drug-likeness model and compared it with our BNC-likeness model.

The dataset we used to train the drug-likeness model was prepared as described in [43]. The World Drug Index (WDI) [40] database listed 59,000 drugs and pharmacologically active compounds. Examples of drug-like compounds (positive examples) were randomly selected from the WDI records containing valid structure files and no obvious errors. The Available Chemicals Directory (ACD) [44] is a widely used database of commercially available compounds. Examples of non-drug-like compounds (negative examples) were randomly chosen from the ACD records that were not included in the WDI database and contained no obvious errors. This resulted in 3930 training examples, 1965 positive and 1965 negative. For each compound, we computed the same 1647 molecular descriptors as for our BNC-likeness model.

An SVM drug-likeness model was trained with the above dataset, displaying $88.84 \pm 2.1\%$ sensitivity, $87.69 \pm 2.3\%$ specificity and $88.26 \pm 2.1\%$ F_2-measure. Its overall accuracy was estimated to be $88.14 \pm 2.2\%$, which was comparable to those reported in previous drug-likeness studies (83.22~92.73%) [25].

As shown in Table 5, if the drug-likeness model was used to classify the natural compound dataset, over 10% drop in sensitivity, specificity and F_2-measure would result. Similarly, if the BNC-likeness model was used to classify the drug-likeness dataset, around 10% drop in sensitivity, specificity and F_2-measure would be observed. These noticeable accuracy drops suggested that our drug-likeness model and BNC-likeness model were likely different. They might rely on different characteristics to tell drug-like chemicals from ordinary compounds and to tell BNC-like NP ingredients from ordinary NP ingredients.

3.5. Important descriptors for drug-likeness prediction and BNC-likeness prediction

In order to develop a deeper insight into the differences between drug-likeness models and BNC-likeness models, we evaluated the importance of each descriptor for drug-likeness prediction and for BNC-likeness prediction.

The RuleSet algorithm, a derivative of the decision tree algorithm, was used to reconstruct drug-likeness models and BNC-likeness models from the same training datasets. A RuleSet model is a collection of decision tree-based classification rules, which usually uses only a small number of descriptors (i.e. 8–25 descriptors in this study). Therefore, the descriptors selected in these rules are regarded more important for the classification task. In a bootstrap experiment, descriptors selected more often are considered more important.

In this work, our BNC-likeness dataset was re-sampled with replacement to generate 50 bootstrap datasets. One RuleSet model was trained with each bootstrap dataset, as detailed in Section 2. These RuleSet BNC-likeness models showed an average sensitivity of $88.98 \pm 0.7\%$, specificity of $86.71 \pm 0.8\%$ and F_2-measure of $87.83 \pm 0.7\%$. These accuracy indicators were inferior to those observed with our SVM BNC-likeness model, yet still satisfactory. With the same procedures, we generated 50 bootstrap drug-likeness datasets, trained 50 RuleSet drug-likeness models, which displayed an average sensitivity of $84.96 \pm 1.5\%$, specificity of $82.70 \pm 1.6\%$ and F_2-measure of $83.81 \pm 1.5\%$. These performance indicators were also inferior to those observed with our SVM drug-likeness model. Yet, they were comparable to those of the RuleSet BNC-likeness models, which allowed a fair comparison between the important descriptors selected for BNC-likeness prediction and those selected for drug-likeness prediction. In the RuleSet BNC-likeness models, 180 descriptors were used, while in the RuleSet drug-likeness models, 328 descriptors were used (Supplemental Table 2).

The software tool we used to calculate our descriptors, DRAGON v5.4, categorized its molecular descriptors into 20 groups based on their calculation methods. According to this categorizing sys-

Table 5
Comparison of our SVM BNC-likeness model and our SVM drug-likeness model.

	Number of examples	Sensitivity	Specificity	F_2-measure	Overall accuracy
Drug-likeness model	3930	88.84%	87.69%	88.26%	88.14%
BNC-likeness model	1580	90.21%	90.12%	90.16%	90.17%
BNC-likeness model predicting drug-like molecules	3930	76.71%	79.87%	78.26%	78.29%
Drug-likeness model predicting BNC-like molecules	1580	81.62%	76.41%	78.93%	78.07%

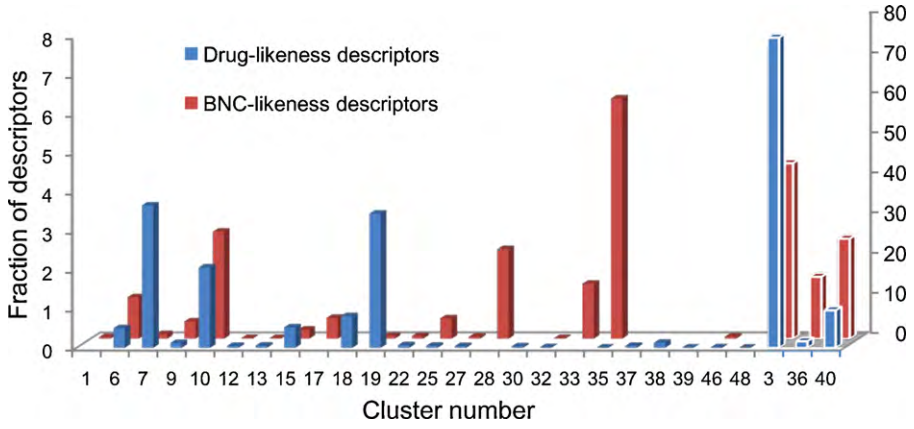

Fig. 2. Distribution of descriptors important for BNC-likeness prediction and drug-likeness prediction. 50 descriptor clusters were generated by the k-means approach. Only categories including important descriptors were shown. Blue columns represented the fraction of descriptors important for BNC-likeness prediction. Red columns represented the fraction of descriptors important for drug-likeness prediction. Because the fractions for Cluster 3, 36 and 40 were much bigger than the rest, they were plotted with a different scale as shown to the right.

tem, we plotted the distribution of the 180 descriptors important for BNC-likeness prediction and the distribution of 328 descriptors important for drug-likeness prediction (Supplemental Fig. 1). Although the contrast was not sharp, there were observable differences between these two distributions.

In order to observe the differences more clearly, we used the k-means clustering approach (50 clusters and 500 runs) [45] to regroup our 1647 descriptors into 50 *de facto* clusters based on their pair-wise Pearson's correlation coefficients, so that descriptors in each cluster would contain similar information. The compositions of these descriptor clusters were given in Supplemental Table 3. According to this categorizing system, the descriptors important for both classification tasks were scattered in 27 clusters. Fig. 2 plotted their distributions, in which clear distinctions were shown. Notably, the descriptors in Cluster 35, 33, 28, and 36 were extensively used in BNC-likeness prediction but were rarely used for drug-likeness prediction. The opposite was found for the descriptions in Cluster 19, 7, and 18. They played important roles in drug-likeness prediction but did not contribute much to BNC-likeness prediction.

Taken together with the decreased accuracies observed when our BNC-likeness model was used to make drug-likeness predictions and vice versa, these results suggested that BNC-likeness prediction and drug-likeness prediction probably relied on different structure characteristics.

4. Limitations and future perspectives

It is important to restate here that, just as "drug-like" molecules do not necessarily result in useful drugs, "BNC-like" compounds do not necessarily indicate bioactive NP ingredients. The "BNC-likeness" concept is "activity-independent". It is used to abstract the shared structural characteristics of bioactive NP ingredients, regardless of their actual activities, reflecting their shared requirements such as good target access, fine stability, and harmless metabolites. However, the honor of a true bioactive NP ingredient belongs only to those natural compounds that also bind targets important in disease pathologies. Therefore, "BNC-likeness" is intended to direct scientists toward natural compounds with potentially higher pharmaceutical value, and thereby focus our searches for principal ingredients to re-produce the combined effects of an NP mixture.

Besides, bioactive NP ingredients may not all contribute to the synergy of a recipe. One possible way to examine the role of each ingredient is to identify its molecular targets and analyze their interactions with targets of other active ingredients. Many patterns of beneficial interactions have been well understood and categorized [46]. In the past years, we have been actively developing computational tools to facilitate this line of research. We had proposed a molecular docking based method to predict the potential molecular targets of an NP ingredient [47]. In nine case studies, approximately ~50% of the predicted targets were supported by literature and ~75% diseases/conditions associated with the predicted targets were found to be alleviated by these ingredients [48]. We have also created a series of databases for pharmacologically important proteins, linking predicted targets to possible physiological responses, such as therapeutic effects, adverse reactions and pharmacokinetic behaviors [49–51]. These tools have been demonstrated useful in analyzing the molecular mechanisms of single NP ingredients, e.g. the cytotoxicity mechanism of ganoderic acid D [52]. However, their application to analyze the synergy in a full recipe of traditional medicine has not yet been attempted. This is because that the number of ingredients in a recipe is usually too big, and the amount of computation required to predict their targets turns out to be prohibitive. The BNC-likeness approach may therefore fill in this gap, allowing us to focus on a much smaller group of potentially bioactive compounds and thereby reduce the computation requirements, enabling our analysis tools to work on full recipes.

In this work, BNC-likeness relied entirely on structure characteristics. However, there are also many non-structural characteristics shared by bioactive natural compounds. For example, the likelihood of an NP ingredient being active may correlate to its content in herb, its binding affinities to its molecular targets, and the potencies of the *in-vivo* natural ligands competing for the same targets. If these data are available, corresponding non-structural descriptors can be easily incorporated into the BNC-likeness models. Together with other groups, we are actively developing relevant databases and software tools, to collect or compute these non-structural descriptors.

The accuracy and usefulness of BNC-likeness models may be further improved, if many limitations of this work could be properly addressed in later studies. First, the example dataset we used was not yet perfect. Rapid accumulating NP bioactivity screening data would surely provide more diversified positive examples and more accurate negative examples. The studies on structure classifications of NP ingredients [27] may also help to enhance the representativeness of our example dataset.

Another factor to consider is, that NP ingredients may act as "pro-drugs", which are not bioactive themselves but their metabolites are [53,54]. It is possible to develop computational tools to predict the likely *in-vivo* derivatives of a natural compound based

on our knowledge of the secondary metabolite biosynthesis pathways in its source [55] and the chemical modification pathways in humans [56]. BNC-likeness models would then be able to examine whether these derivatives were potentially bioactive.

In addition, it would be desirable if BNC-likeness could be described with simple interpretable rules, instead of black-box SAR models, like the famous "rule-of-five" to describe drug-likeness [24]. This probably requires further efforts to develop novel chemical descriptors that are more relevant to the bioactivities of NPs, which may likely represent certain common molecular mechanisms of NP actions. In this work, although rule-based statistical learning algorithms were also used to build BNC-likeness models, the rule sets generated were relatively complicated and not easily interpretable in a biological or chemical sense. Progresses in statistical learning algorithms [26] and descriptor calculations [57,58] may offer new opportunities to describe BNC-likes in the form of simple rules. On the other hand, it needs to be noted that "rule-of-five" is a very simple criteria [59]. Currently we had less than 4800 small molecule drugs and drug leads [60], but it was reported that a huge number (>17 million) of compounds were compliant with the "rule-of-five" criteria in PubChem [22]. In other words, simple rules like "rule-of-five" usually had high sensitivity but lack of specificity. More complicated statistical model like our SVM model and many other black-box models [25] may achieve balanced sensitivity and specificity. The "rule-of-five" criteria can be useful to alert medicinal chemist when venturing too far from a reasonable property range, while more accurate statistical models shall be used to practically guide a screening process [22].

5. Conclusions

In this work, we proposed the use of structure activity relationship to narrow down the range of potentially bioactive ingredients in a natural product mixture. We called this approach the Bioactive Natural Compound-likeness (BNC-likeness) approach, drawing a parallel to the concept of drug-likeness. In cross-validations and independent example tests, our BNC-likeness approach displayed 90–92% sensitivity and 85–90% specificity. This level of accuracy was similar to those of the state-of-the-art drug-likeness classification models, suggesting its practical usefulness. We also showed that BNC-likeness and drug-likeness were different. There would be a ~10% drop in accuracy if a BNC-likeness model were used to predict drug-like molecules and vice versa. Additional results indicated that probably different structural features were important in bioactive natural compound prediction and drug-like molecule prediction. It is our hope that this preliminarily work would attract further efforts to charter the space of bioactive NP ingredients, thereby providing a computational solution to focus us on potentially more important compounds when analyzing the coordinated therapeutic mechanisms of natural products.

Acknowledgements

This work is support by the National Natural Science Foundation of China (NSFC) Grant 30970690, and the Zhejiang Provincial Natural Science Foundation of China Grant No. R207609. We thank Mr. Chris Wood at Zhejiang University for editing this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2010.04.007.

References

[1] W.R. Strohl, The role of natural products in a modern drug discovery program, Drug Discov. Today 5 (2000) 39–41.
[2] M.C. Sutter, Y.X. Wang, Recent cardiovascular drugs from Chinese medicinal plants, Cardiovasc. Res. 27 (1993) 1891–1901.
[3] D.Y.B.D.L. Zhu, X.C. Tang, Recent studies on traditional Chinese medicinal plants, Drug Dev. Res. 39 (1996) 147–157.
[4] F. Li, S. Sun, J. Wang, D. Wang, Chromatography of medicinal plants and Chinese traditional medicines, Biomed. Chromatogr. 12 (1998) 78–85.
[5] X. Gong, N.J. Sucher, Stroke therapy in traditional Chinese medicine (TCM): prospects for drug discovery and development, Trends Pharmacol. Sci. 20 (1999) 191–196.
[6] K.H. Lee, Novel antitumor agents from higher plants, Med. Res. Rev. 19 (1999) 569–596.
[7] F.J. Evans, Natural products as probes for new drug target identification, J. Ethnopharmacol. 32 (1991) 91–101.
[8] R. Chaudhury, Herbal Medicine for Human Health, WHO Regional Office for South-East Asia, 1992.
[9] K. Chan, Progress in traditional Chinese medicine, Trends Pharmacol. Sci. 16 (1995) 182–187.
[10] J. Henkel, Attacking AIDS with a 'cocktail' therapy? FDA Consum. 33 (1999) 12–17.
[11] J. Feliu, M. Sereno, J.D. Castro, C. Belda, E. Casado, M. Gonzalez-Baron, Chemotherapy for colorectal cancer in the elderly: whom to treat and what to use, Cancer Treat. Rev. 35 (2009) 246–254.
[12] A.A. Borisy, P.J. Elliott, N.W. Hurst, M.S. Lee, J. Lehar, E.R. Price, et al., Systematic discovery of multicomponent therapeutics, Proc. Natl. Acad. Sci. USA 100 (2003) 7977–7982.
[13] T. Xue, R. Roy, Studying traditional Chinese medicine, Science 300 (2003) 740–741.
[14] H. Oka, S. Yamamoto, T. Kuroki, S. Harihara, T. Marumo, S.R. Kim, et al., Prospective study of chemoprevention of hepatocellular carcinoma with Sho-saiko-to (TJ-9), Cancer 76 (1995) 743–749.
[15] R. Sheehan-Dare, J. Cotterill, Experience with the Hexascan in argon laser treatment of vascular skin lesions, Br. J. Dermatol. 127 (1992) 33–34.
[16] M.P. Sheehan, M.H. Rustin, D.J. Atherton, C. Buckley, D.W. Harris, J. Brostoff, et al., Efficacy of traditional Chinese herbal therapy in adult atopic dermatitis, Lancet 340 (1992) 13–17.
[17] S. Sengupta, S.A. Toh, L.A. Sellers, J.N. Skepper, P. Koolwijk, H.W. Leung, et al., Modulating angiogenesis: the yin and the yang in ginseng, Circulation 110 (2004) 1219–1225.
[18] L. Wang, G.B. Zhou, P. Liu, J.H. Song, Y. Liang, X.J. Yan, et al., Dissection of mechanisms of Chinese medicinal formula Realgar–Indigo naturalis as an effective treatment for promyelocytic leukemia, Proc. Natl. Acad. Sci. USA 105 (2008) 4826–4831.
[19] M.F.K.A.D. Balandrin, N.R. Farnsworth, In Human Medicinal Agents from Plants, American Chemical Society, Washington, DC, 1993.
[20] M.S. Butler, The role of natural product chemistry in drug discovery, J. Nat. Prod. 67 (2004) 2141–2153.
[21] B. Singh, P. Kaur, Gopichand, R.D. Singh, P.S. Ahuja, Biology and chemistry of *Ginkgo biloba*, Fitoterapia 79 (2008) 401–418.
[22] G. Vistoli, A. Pedretti, B. Testa, Assessing drug-likeness – what are we missing? Drug Discov Today 13 (2008) 285–294.
[23] R.D. Brown, M. Hassan, M. Waldman, Combinatorial library design for diversity, cost efficiency, and drug-like character, J. Mol. Graph. Model 18 (2000), 427–437, 537.
[24] C.A. Lipinski, Drug-like properties and the causes of poor solubility and poor permeability, J. Pharmacol. Toxicol. Methods 44 (2000) 235–249.
[25] Q. Li, A. Bender, J. Pei, L. Lai, A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification, J. Chem. Inf. Model 47 (2007) 1776–1786.
[26] N. Schneider, C. Jackels, C. Andres, M.C. Hutter, Gradual in silico filtering for druglike substances, J. Chem. Inf. Model 48 (2008) 613–628.
[27] M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, et al., Charting biologically relevant chemical space: a structural classification of natural products (SCONP), Proc. Natl. Acad. Sci. USA 102 (2005) 17272–17277.
[28] T. Henkel, R. Brunne, H. Müller, F. Reichel, Statistical investigation into the structural complementarity of natural products and synthetic compounds, Angew. Chem. Int. Ed. 38 (1999) 643–647.
[29] M. Feher, J.M. Schmidt, Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry, J. Chem. Inf. Comput. Sci. 43 (2003) 218–227.
[30] N. Singh, R. Guha, M.A. Giulianotti, C. Pinilla, R.A. Houghten, J.L. Medina-Franco, Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository, J. Chem. Inf. Model 49 (2009) 1010–1024.
[31] M.L. Lee, G. Schneider, Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries, J. Comb. Chem. 3 (2001) 284–289.
[32] F.L. Stahura, J.W. Godden, L. Xue, J. Bajorath, Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations, J. Chem. Inf. Comput. Sci. 40 (2000) 1245–1252.
[33] P. Ertl, S. Roggo, A. Schuffenhauer, Natural product-likeness score and its application for prioritization of compound libraries, J. Chem. Inf. Model 48 (2008) 68–74.
[34] J. Duke, Dr. Duke's Phytochemical and Ethnobotanical Databases, 2006, http://www.ars-grin.gov/duke/ (accessed October 1, 2009).
[35] G.M. Cragg, M.R. Boyd, Y.F. Hallock, D.J. Newman, E.A. Sausville, M.K. Wolpert, Natural products drug discovery at the national cancer institute. Past achieve-

ments and new defections for the new millennium, in: K.W. Stephen, A.H. Martin, T. Robert, J.T.C. Ewan, N. Neville (Eds.), Biodiversity: New Leads for the Pharmaceutical and Agrochemical Industries, RSC Publishing, Cambridge UK, 2000, pp. 22–45.
[36] M.C. Nicklaus, Technical Notes of Structure Files of NCI Open Database Compounds, September 2003 SD File of Combined DTP Releases, 2007, http://129.43.27.140/ncidb2/download-notes_2003-09.htm (accessed September 23, 2009).
[37] DRAGON, version 5.4, Talete srl, Milano, Italy, 2007.
[38] S. Winters-Hilt, A. Yelundur, C. McChesney, M. Landry, Support vector machine implementations for classification & clustering, BMC Bioinformatics 7 (Suppl. 2) (2006) S4.
[39] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 2003.
[40] World Drug Index (WDI), Derwent Information, London, 2007.
[41] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.
[42] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159.
[43] J. Sadowski, H. Kubinyi, A scoring scheme for discriminating between drugs and nondrugs, J. Med. Chem. 41 (1998) 3325–3329.
[44] Available Chemicals Directory (ACD), Molecular Design Limited, Calif, 2007.
[45] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (1982) 129–137.
[46] J. Jia, F. Zhu, X. Ma, Z. Cao, Y. Li, Y.Z. Chen, Mechanisms of drug combinations: interaction and network perspectives, Nat. Rev. Drug Discov. 8 (2009) 111–128.
[47] Y.Z. Chen, D.G. Zhi, Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule, Proteins 43 (2001) 217–226.
[48] X. Chen, C.Y. Ung, Y. Chen, Can an in silico drug-target search method be used to probe potential mechanisms of medicinal plant ingredients? Nat. Prod. Rep. 20 (2003) 432–444.
[49] X. Chen, Z.L. Ji, Y.Z. Chen, TTD: therapeutic target database, Nucleic Acids Res. 30 (2002) 412–415.
[50] L.Z. Sun, Z.L. Ji, X. Chen, J.F. Wang, Y.Z. Chen, ADME-AP: a database of ADME associated proteins, Bioinformatics 18 (2002) 1699–1700.
[51] Z.L. Ji, L.Y. Han, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Drug adverse reaction target database (DART): proteins related to adverse drug reactions, Drug Saf. 26 (2003) 685–690.
[52] Q.X. Yue, Z.W. Cao, S.H. Guan, X.H. Liu, L. Tao, W.Y. Wu, et al., Proteomics characterization of the cytotoxicity mechanism of ganoderic acid D and computer-automated estimation of the possible drug target network, Mol. Cell. Proteomics 7 (2008) 949–961.
[53] D. Classen-Houben, D. Schuster, T. Da Cunha, A. Odermatt, G. Wolber, U. Jordis, et al., Selective inhibition of 11beta-hydroxysteroid dehydrogenase 1 by 18alpha-glycyrrhetinic acid but not 18beta-glycyrrhetinic acid, J. Steroid Biochem. Mol. Biol. 113 (2009) 248–252.
[54] A. Kavitha, P. Prabhakar, M. Narasimhulu, M. Vijayalakshmi, Y. Venkateswarlu, K. Venkateswarlu Rao, et al., Isolation, characterization and biological evaluation of bioactive metabolites from Nocardia levis MK-VL_113, Microbiol. Res. 165 (2010) 199–210.
[55] M.K. Julsing, A. Koulman, H.J. Woerdenbag, W.J. Quax, O. Kayser, Combinatorial biosynthesis of medicinal plant secondary metabolites, Biomol. Eng. 23 (2006) 265–279.
[56] F.P. Guengerich, Cytochrome P450s and other enzymes in drug metabolism and toxicity, AAPS J. 8 (2006) E101–E111.
[57] R.V. Todeschini, Consonni (Eds.), Handbook of Molecular Descriptors, Wiley–VCH, Weinheim, 2000.
[58] C.W. Yap, H. Li, Z.L. Ji, Y.Z. Chen, Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties, Mini Rev. Med. Chem. 7 (2007) 1097–1107.
[59] M. Vieth, M.G. Siegel, R.E. Higgs, I.A. Watson, D.H. Robertson, K.A. Savin, et al., Characteristic physical properties and structural fragments of marketed oral drugs, J. Med. Chem. 47 (2004) 224–232.
[60] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, et al., DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.