Regular Article

Novel Compound-Target Interactions Prediction for the Herbal Formula Hua-Yu-Qiang-Shen-Tong-Bi-Fang

Zihao Wang, a,b Hui-Heng Lin, a,c Kegang Linghu, a Run-Yue Huang, b,c Guangyao Li, a Huali Zuo, a Hua Yu, a Ging Chan, a,c and Yuanjia Hu a,c

© 2019 The Pharmaceutical Society of Japan

Herbal formulae have a long history in clinical medicine in Asia. While the complexity of the formulae leads to the compound-target interactions and the resultant multi-target therapeutic effects, it is difficult to elucidate the molecular/therapeutic mechanism of action for the many formulae. For example, the Hua-Yu-Qiang-Shen-Tong-Bi-Fang (TBF), an herbal formula of Chinese medicine, has been used for treating rheumatoid arthritis. However, the target information of a great number of compounds from the TBF formula is missing. In this study, we predicted the targets of the compounds from the TBF formula via network analysis and in silico computing. Initially, the information of the phytochemicals contained in the plants of the herbal formula was collected, and subsequently computed to their corresponding fingerprints for the sake of structural similarity calculation. Then a compound structural similarity network infused with available target information was constructed. Five local similarity indices were used and compared for their performance on predicting the potential new targets of the compounds. Finally, the Preferential Attachment Index was selected for it having an area under curve (AUC) of 0.886, which outperforms the other four algorithms in predicting the compound-target interactions. This method could provide a promising direction for identifying the compound-target interactions of herbal formulae in silico.

Key words  herbal formula; Hua-Yu-Qiang-Shen-Tong-Bi-Fang; natural product; in-silico target identification; network link prediction

Introduction

Utilization of medicinal plants for the benefits of human beings could be traced back to the dawn of the humanity. The extracts and the derivatives of the plants have been regarded as the building blocks of the healing systems such as Chinese herbal medicine,1 Japanese Kampo medicine,2 and Indian Ayurvedic medicine.3 Historically, plants have served as a rich source of compounds and new scaffolds for drug discovery. Until 2014, there have been 160 natural products inspired anticancer drugs approved for clinical use worldwide, compensating 77% of all the approved small molecule drugs for oncological treatment,4 yet most countries have not officially recognized the herb-based traditional medicine due to insufficient research data on the safety and efficacy of the medicine.5

While the therapeutic values of some of the herbal formulae have been established, there is still a great amount of research need to be done for many other formulae in terms of the efficacy and the safety of the treatment. To unveil the mask of confusion introduced by the combinatorial utilization of the active ingredients from various plants, the reductionist approach has been applied to study the complex system through isolating the active ingredients from the plants,5 and the pharmaceutical industry has witnessed its heyday from natural products since 19th century. However, the progress of the development of new drugs has been slowed down, and these newly introduced agents on trial have been observed with unexpected effects on other biochemical mechanisms which resulted in toxic reactions in clinical practice.6 The concept of “one gene, one drug, one disease” is being challenged by the increasingly accepted “polypharmacology,” and through the identification of molecular targets of the compounds, not only the mode of action of these compounds could be elucidated, but also the information of the so-called “off-targets” of the known compounds could emerge, which accounts for the undesired biological activity and toxicity on clinical trial.7 The comprehensive information of the interaction among the compounds and their respective targets may help researchers make an informed decision when selecting potential drugs leads. Thus, there is an increasing interest on detecting leads from a holistic perspective, hoping to enhance the successful rate of the new drugs with pleiotropic effects for polygenic diseases.8 Consequently, it became crucial for chemical biological researchers to identify the active ingredients of an herb, as well as their targeting proteins.

Decades before, researchers in academia and pharmaceutical industry heavily relied on biochemical or molecular experimental methods (so-called wet lab or simply wet) for the purpose of target identifications.9,10 Nowadays, the rapidly developing and state-of-the-art computational methods, sometimes termed dry methods, have played an equally important role by complementing wet labs’ weaknesses in large scale and systematic data analyses for both targets and off-targets of bioactive compounds. The polypharmacological characteristic

1 These authors contributed equally to this work.

* To whom correspondence should be addressed. e-mail: gchan@umac.mo; yuanjiah@umac.mo

© 2019 The Pharmaceutical Society of Japan
of natural products inspired efforts to predict compound-target interactions. Keiser et al. developed techniques to predict targets for ligands based on the chemical similarity among the ligands of the corresponding receptors. Gfeller et al. developed a method of calculating the chemical and shape similarity among molecules to predict their potential targets. Several network-based methods have been developed to study the drug target interactions. Yildirim et al. constructed and analyzed the bipartite drug-target network, and revealed the topological features of the drug-target network. Yamanishi et al. integrated chemical structure data, genomic sequence data as well as known drug-target network information into a supervised learning method to predict drug-target interactions. Cheng et al. developed and compared the performance of three supervised inference methods including drug-based similarity inference, target-based similarity inference and network-based inference for drug-target interaction prediction, and they concluded that network-based inference method outperforms the other two. Although machine learning-based methods tended to perform well, they required extremely large amount of data regarding the characteristics of the chemicals and the protein targets, of which natural products often lack. We therefore proposed a network-based method to predict the unknown compound-target interactions in the context of an herbal formula. Our method relied on the topological features of the compounds’ structural similarity and the known compound-target interactions network.

In this study, we used Hua-Yu-Qiang-Shen-Tong-Bi-Fang (TBF) formula as an example to showcase our method for predicting compound-target interactions of natural products. TBF is an herbal formula developed by the Department of Rheumatology and Immunology at the Guangdong Provincial Hospital of Chinese Medicine, and it has been frequently prescribed to patients suffered from rheumatoid arthritis in clinical practice for more than 20 years. TBF formula is composed of organ parts of the 10 herbs, including Salvia miltiorrhiza (SM), Dioscorea nipponica MAKINO (DN), Astragalus membranaceus (AM), Paeonia lactiflora (PL), Saussurea involucrata (SI), Eucommia ulmoides (EU), Drynaria fortunei (DF), Dipsacus asperoides (DA), Rehmannia glutinosa (RG), and Glycyrrhiza uralensis (GU). Although the clinical efficacy of the TBF formula has been observed in clinic, the mechanism of action for the treatment has been elusive largely due to the scarce chemical and biological information of several herbs contained in the formula. Just like many other herbal formulae, TBF formula contains herbs that lack chemical information as well as target information. By choosing TBF formula as an example, we hope to propose a method applicable to other herbal formulae in terms of compound-target interaction prediction.

We predicted the targets of the phytochemicals in TBF formula by computing the topological structural similarity indices of the network. As illustrated in the Fig. 1, the 10 herbs of the formula were dissected into molecular level, after that, all of the phytochemicals were computed into their respective fingerprints before calculating the Tanimoto coefficient (Tc) among each other. A threshold value of 0.85 was applied.
to identify the structural similarity among each compound, as 85% of the compounds having a Tc value of not less than 0.85 compared to a known active compound were also found to be active.\textsuperscript{18} The experimentally validated compound-target interactions were collected from ChEMBL\textsuperscript{19} and DrugBank\textsuperscript{20} databases. After constructing the compound-target interaction network enriched with compound structural similarity information, the likelihood of each nonexistent compound-target interaction was deduced using link prediction algorithms based on the five local similarity indices subsequently. These indices include Common Neighbors Index (CN), Adamic-Adar Index (AA), Jaccard Index (JC), Preferential Attachment Index (PA) and Resource Allocation Index (RA). The receiver operating characteristic (ROC) curves for different algorithms were calculated on the data sets using 10-fold cross validation, and the area under the ROC curve (AUC) was calculated to evaluate the performance of each algorithm on predicting the compound-target interactions for the phytochemicals. Finally, we selected PA algorithm, which showed the best performance among the five algorithms, to predict the potential targets of the phytochemicals in TBF.

**Experimental**

**Data Preparation** The chemical information of the TBF formula was manually curated from the TCM Database @ Taiwan (last updated on March 25th, 2014),\textsuperscript{21} which is the world’s largest database containing ingredient information for traditional Chinese medicine with free of charge. Besides, the information regarding the chemical constituents of the 10 herbs was also supplemented from the Shanghai Institute of Organic Chemistry of Chinese Academy of Sciences. Chemistry Database (last updated on June 17th, 2016)\textsuperscript{22} as well as SciFinder database to render the chemical information as comprehensive as possible. The compounds’ information obtained from the 10 herbs were mapped to their corresponding compound-target interactions using information from ChEMBL database (version 23)\textsuperscript{19} and DrugBank database (version 5.0).\textsuperscript{20}

**Network Construction** The TBF compound-target interaction network enriched with compound structural similarity information was visualized using Cytoscape (version 3.6.1)\textsuperscript{23} and NetworkX (version 2.1).\textsuperscript{24} For each phytochemical contained in the formula, we computed their Molecular ACCess System (MACCS) keys\textsuperscript{25} using settings from RDKit.\textsuperscript{26} The similarity between each two compounds was then calculated by using Tc as a metric. If the Tc values between the two compounds were not less than 0.85, then the two would be considered structurally similar, and they were expected to show about the same activity against the same set of proteins.\textsuperscript{18} The structurally similar compounds were visualized in a network before introducing their known compound-target interactions to the network. A compound-target interaction network containing structural similarity information of compounds could be represented as a network graph $G (N, E)$, where compounds and targets are the nodes and the compound-target interactions and the structural similarity information are the edges. The information of all the nodes and edges were stored in a list respectively before added to the graph $G$ and visualized using NetworkX (version 2.1)\textsuperscript{24} for the subsequent analysis

**Link Prediction Algorithms** To predict the potential compound-target interactions, similarity-based algorithms were used to predict the possible links. For a pair of nodes $x$ and $y$, where $x$ represented a random compound and $y$ represented a random protein target, they were assigned with a score $S_{xy}$ which was defined as an index to measure the similarity between $x$ and $y$. All of the non-observed links were ranked regarding their scores, and the links with higher scores, i.e., the links connecting the more similar nodes, were considered having a greater credibility to be the true links. The similarity between the two nodes could be evaluated by comparing a variety of their features,\textsuperscript{27} and we focused on the network topological structural similarity of the compounds and targets nodes. To further clarify our methodology, the modeling process of the prediction method was illustrated in Fig. 2. In this simple graph, common neighbors index was recruited for calculating the scores of each potential compound-target pairs in the whole graph (a). The compound-target pairs were ranked according to their scores calculated, and the highly ranked compound-target pairs “a-B” and “a-C” were selected as potential compound-target interactions and visualized in the predicted graph (b). In this study, we specifically implemented 5 local similarity indices for similarity evaluation between each pair of nodes.

**Common Neighbors (CN)**

The main idea of using this method was that two nodes, $x$ and $y$, were more likely to be connected if they shared more common neighbors, namely determining the similarity of the two nodes by measuring the overlaps among the neighbors of the two nodes.\textsuperscript{28} For a pair of nodes in a network, $I(x)$ denotes the sets of neighbors of the node $x$, and $I(y)$ denotes the sets of neighbors of the node $y$. The predicted score for each compound-target interaction is:

$$S_{CN}^{xy} = |I(x) \cap I(y)|$$

**Jaccard Index (JC)**

Paul Jaccard originally proposed JC in 1901,\textsuperscript{29} and this index designed for comparing the similarity of the two objects had been commonly used in information retrieval.\textsuperscript{29} In our case, JC measured the probability of the node $x$ and the node $y$ having the same neighbors. The prediction score based on JC is defined as:

$$S_{JC}^{xy} = \frac{|I(x) \cap I(y)|}{|I(x) \cup I(y)|}$$

**Preferential Attachment Index (PA)**

The basic idea of using this method was based on the phenomenon that nodes with higher degree tended to generate more new links. Let $k_x$ be the degree of the node $x$, and $k_y$ be the degree of the node $y$, then the likelihood of the existence of the link between $x$ and $y$ was proportional to $k_x \times k_y$.\textsuperscript{30} Since this index only depends on the degrees of the nodes, rather than considering the complete information of the nodes’ neighbors, PA requires the least computational complexity of all the similarity indices. The link between the node $x$ and the node $y$ can be determined by the following score:

$$S_{PA}^{xy} = k_x \times k_y$$

**Adamic–Adar Index (AA)**

This index was originally proposed by Lada A. Adamic and Eytan Adar to predict the potential friendship between the two strangers based on their shared characteristics,\textsuperscript{30} and it had been applied to predict the missing links in biological

$$S_{AA}^{xy} = \sum_{z \in I(x) \cap I(y)} \frac{1}{\log(k_z)}$$
networks. In this study, AA directly counted the common neighbors of the two nodes \( x \) and \( y \) by assigning the less connected neighbors more weights, taking the degree of each neighbor (\( k_z \)) into consideration when predicting the missing links. AA is defined as:

\[
S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}
\]

Resource Allocation Index (RA)

The mechanism of RA can be used to predict the link between a pair of nodes, \( x \) and \( y \), which are not directly connected. The common neighbors of the pair could act as transmitters to distribute the resource between the two nodes. That is, the similarity between the node \( x \) and the node \( y \) is determined by the amount of resource the node \( y \) receives from the node \( x \). This index was proposed under the assumption that each transmitter could give away its unit of resource equally to its neighbors. And the scoring metric based on RA can be described as following:

\[
S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}
\]

Evaluation of the Link Prediction Performance

To evaluate the performance of each algorithm, known compound-target interactions of the formula were split into two sets randomly: the training set and the probe set. As exemplified in the training graph of Fig. 2, 2 out of the 8 known compound-target interactions were randomly chosen to be the probe, and these two links were masked as missing links during the link prediction process for the whole graph. The 10-fold cross validation was introduced to the evaluation method so that each known compound-target interaction could serve in the probe set. In other words, the entire known compound-target interactions were divided into 10 equal-sized parts by chance, and each part was assigned into the probe set alternately, while the others were used as training set. The ROC curve of each prediction algorithm was plotted based on the ROC curves averaged from the 10 folds, and the \( AUC \) was calculated using the composite trapezoidal rule subsequently. The higher the \( AUC \) value, the better the performance of the link prediction algorithm.

Results and Discussion

Compound Distribution of the Herbs in TBF

Using information from the databases described in Experimental, we retrieved the chemical information of the 10 herbs constituting the TBF formula. A total of 800 unique phytochemicals were identified, and the phytochemical distribution for the herbs was illustrated in the Fig. 3. Among the 10 herbs, Glycyrrhiza uralensis (GU) was found to have the most phytochemical information, with 208 phytochemicals identified by fellow researchers. While the information of the phytochemicals contained in Drynaria fortunei (DF) was relatively poor, with only 30 phytochemicals elucidated so far. GU is one of the most commonly used herbs, appearing in about 60% of all the traditional Chinese medicine (TCM) prescriptions, and the bioactive compounds in GU had been extensively studied. On the other hand, research works on DF had been relatively limited. Currently found phytochemicals from DF mainly belonged to the groups of flavonoids and phenylpropanoids.
The experimentally validated compound-target interaction information of the herbs was obtained from the ChEMBL\(^{19}\) and DrugBank\(^{20}\) databases, and was shown in the bar chart of Fig. 3.

The molecular formulae of the phytochemicals contained in the TBF formula were computed to their corresponding 2D fingerprints (MACCS keys), and the all-by-all Tanimoto similarity matrices were calculated for the 800 phytochemicals to determine their structural similarity. By applying a threshold value of \(T_c \geq 0.85\), 4217 pairs of structurally similar phytochemicals were identified, involving 657 unique compounds.

Performance Evaluation of the Algorithms in Compound-Target Interactions Prediction

A total of 143 phytochemicals among the 657 structurally similar compounds were identified for their known human targets either through ChEMBL\(^{19}\) or DrugBank\(^{20}\) databases. 2500 pairs of experimentally validated compound-target interactions (Table S1) of the TBF formula were used to generate a compound-target bipartite network (Fig. 4), where a compound and a target were connected if the target was proven to be responsible for the compound.

In search of a suitable index for predicting potential compound-target interactions, we chose 5 commonly used local similarity indices for computation, and we compared their prediction accuracy by calculating the area under the ROC curve (AUC) of the prediction results (Table 1). Preferential Attachment Index (PA) has not only exhibited an AUC as large as 0.900, but also achieved the largest average AUC among the 5 algorithms with 0.886 ± 0.010. As illustrated in the Fig. 5,
the ROC curves were plotted according to the average prediction results calculated by the 5 algorithms using 10-fold cross validation. Clearly, all of the methods performed better than chance (the gray dashed line in Fig. 5 represents the chance diagonal for prediction, and it has an \( AUC \) of 0.5). Particularly, PA outperformed the other four indices in this study by always presenting the best true positive rate (TPR) at any false positive rate (FPR) (Fig. 5). While Jaccard Index (JC) performed poorly in this scenario with an \( AUC \) of 0.748 ± 0.013.

As previously explained in Experimental, the scoring system based on PA was contingent on the degrees of the nodes, specifically, a protein target interacting with various compounds tended to accommodate compounds sharing multiple structural similarities with others. And since each node had its own degree, thus each compound-target pair had a PA score of more than zero in the network. In this way, PA was good for discriminating compound-target interactions having similar network topological features. However, the other four indices, CN, JC, AA, and RA, were dependent on the common neighbors between the two nodes. In this way, more than 90% of the entire compound-target pairs were assigned with zero scores due to their lack of common neighbors, making it difficult for ranking these interactions, and resulting in a great error when predicting compound-target interactions, thus not suitable for prediction. Interestingly, seeing the results (Table 1 and Fig. 5) of different methods, the performance

Table 1. The Area Under the ROC Curve (\( AUC \)) of the 5 Local Similarity Indices

| Local similarity index | \( AUC_{\text{max}} \) | \( AUC_{\text{min}} \) | \( AUC_{\text{ave.}} \) |
|------------------------|----------------------|----------------------|----------------------|
| PA                     | 0.900                | 0.866                | 0.886 ± 0.010        |
| AA                     | 0.771                | 0.723                | 0.755 ± 0.013        |
| CN                     | 0.771                | 0.723                | 0.755 ± 0.013        |
| RA                     | 0.771                | 0.722                | 0.754 ± 0.013        |
| JC                     | 0.763                | 0.717                | 0.748 ± 0.013        |

PA: Preferential Attachment Index; AA: Adamic–Adar Index; CN: Common Neighbors Index; RA: Resource Allocation Index; JC: Jaccard Index. Amongst, PA outperformed other algorithms by scoring the highest values of \( AUC \).

Fig. 5. The Average ROC Curves Using 10-Fold Cross Validation for the Five Link Prediction Algorithms

The \( AUC \) of each algorithm was shown in the legend at the lower right corner of the graph with PA having the highest \( AUC \) while JC having the lowest \( AUC \). (Color figure can be accessed in the online version.)

Fig. 6. A Compound-Target Interaction Network Based on the High-Potential Interactions (Top 20) Predicted by PA Method

The pink edges were drawn to connect the compounds with their potential targets, while the gray edges represent experimentally validated compound-target interactions. The node representing quercetin was located at the center of the nine nodes denoting the compounds involved with the top 20 ranked potential compound-target interactions. (Color figure can be accessed in the online version.)
of the four types of common neighbors-based algorithms had close performance to each other while the only degree-based PA index outperformed others. For the four types of common neighbors-based methods, it is possible that the common neighbors principle is playing the dominant role in these methods, and thus the slight changing the index calculation way may not generate large differences among their performances. Also, this interesting outcome may be suggesting that, similar to other real-world networks, the dynamics of the constructed compound-target interactions network in this study is closer to the preferential attachment model.30)

Prediction of Compound-Target Interactions in TBF Formula Using Preferential Attachment Index (PA) Method We chose PA as an optimal similarity index to predict the potential compound-target interactions for the total phytochemicals contained in the TBF formula, due to its good performance in this study. A total of 345550 pairs of compound-target interactions were assigned with a likelihood score based on PA method. A ROC curve for the compound-target interactions prediction in the whole graph was plotted (Fig. S1), and a threshold value of 0.2 was selected as a cutoff point for the highly confident compound-target interactions according to the Youden’s index38) (Table S2). The potential compound-target interactions were shown in Table S3 and visualized in the network shown in the Fig. S2. Compared with the network in Fig. 3, among the 466 compounds not having known compound-target interactions in the network, 329 of them were assigned with potential human targets by using PA method.

Among the top potential compound-target interactions predicted by our method (Fig. 6), quercetin was expected to interact with solute carrier organic anion transporting polypeptide 1B1 (OATP1B1) (UniProt accession number: Q9Y6L6; ChEMBL ID: 1697668), which belongs to the same organic anion transporting polypeptides family as solute carrier organic anion transporting polypeptide 1B3 (OATP1B3) (UniProt accession number: Q9NPD5; ChEMBL ID: 1743121). Quercetin was a naturally occurring flavonoid widely existed in plants,39) and researchers discovered that it could target OATP1B1.39) Both OATP1B1 and OATP1B3 were considered to mediate Na⁺-independent uptake of organic anions including bile acids, and they were involved in the clearance of bile acids from the liver.40) Wu et al. had found out that quercetin could inhibit OATP1B1-mediated estrone-3-sulphate uptake in vitro.41) These findings in turn supported our prediction that quercetin could potentially target OATP1B1.

Although the network-based compound-target interaction prediction method was useful in predicting target information for structurally similar phytochemicals, its feasibility was restricted to the compounds connected to the neighbors with validated compound-target interactions, thus this method was not applicable to predict the compound-target interactions for small sub-networks as in the bottom part of the Fig. 4, where target information was absent. One way to overcome this obstacle would be to encompass the information of all the natural products in nature as comprehensive as possible in a network. There were about 250000 natural products currently discovered from nature,42) and through analyzing the topological features of the desired compounds in a more comprehensive compound-target interactions network, more natural products could be qualified for network-based compound-target interaction prediction.

Conclusion We have proposed and evaluated a method to predict the potential compound-target interactions for phytochemicals contained in an herbal formula in silico. This method utilized compounds’ structural similarity information and link prediction algorithms to predict potential compound-target interactions for the phytochemicals. Among the five link prediction algorithms analyzed in this study, PA has exhibited a relatively good performance for the TBF formula, with an AUC value of 0.886 ± 0.010. Finally, we used PA as an optimal index to predict the potential compound-target interactions, and we selected the top 20% compound-target interactions as potential compound-target interactions. One of the top predicted compound-target interactions for quercetin, OATP1B1, has been studied and validated in existing literature. The good performance of using PA method for compound-target interaction prediction of the phytochemicals in the TBF formula suggested that the network-based compound-target interaction prediction method could be extended into other herbal formulas, especially when the chemical and biological information of the herbs contained in the formulae is limited. PA was considered superior to CN-based link prediction algorithms due to its characteristic of only using the degree information of each node in a network for prediction. And this characteristic complemented the lack of compound-target interactions and structural similarity information for many herbal formulae. This study not only offered an alternative when dealing with the lack of target information for an herbal formula, but also provided new perspectives for studying the target as well as the off-target effects of drug leads in network pharmacology.

Acknowledgments We thank the Science and Technology Development Fund of Macao SAR and Guangdong-Macao Traditional Chinese Medicine Technology Industrial Park for financial support for projects 013-2015-A1 and D-Pro-0274-2017 for this research. We also thank Yuehao Jia for helpful discussions.

Conflict of Interest The authors declare no conflict of interest.

Supplementary Materials The online version of this article contains supplementary materials.

References
1) Li S., Zhang B., Zhang N., BMC Syst. Biol., 5 (Suppl. 1), S10 (2011).
2) Yakubo S., Ito M., Ueda Y., Okamoto H., Kimura Y., Amano Y., Togo T., Adachi H., Mitsuma T., Watanabe K., Evid. Based Complement. Alternat. Med., 2014, 535146–535155 (2014).
3) Mukherjee P. K., Nema K. N., Venkatesh P., Debnath P. K., J. Ethnopharmacol., 143, 424–434 (2012).
4) Newman D. J., Cragg G. M., J. Nat. Prod., 79, 629–661 (2016).
5) World Health Organization, “General Guidelines for Methodologies on Research and Evaluation of Traditional Medicine.” <http://apps.who.int/iris/bitstream/handle/10665/66783/WHO?fsequence=1>, accessed Sep. 23, 2018.
6) Cragg G. M., Newman D. J., Biochim. Biophys. Acta, 1830, 3670–3695 (2013).
7) van der Greef J., McBurney R. N., Nat. Rev. Drug Discov., 4, 961–967 (2005).
