Identifying Active Compounds from Chinese Medicinal Plants via Causal Variable Selection

Xuewei Wang xuewei@zju.edu.cn
Yiyu Cheng chengyy@zju.edu.cn
Pharmaceutical Informatics Institute, Zhejiang University, Hangzhou, 310027, P. R. China

Keywords: drug discovery, natural products, quantitative composition-activity relationship, Markov blanket

Abstract

Medicinal plants are growing to be a major source of drug discovery, and one challenging problem is to identify the drug candidates from numerous ingredients in medicinal plants. We present an approach to identify active compounds from Chinese medicinal plant based on causal variable selection techniques. We examined three methods, including stepwise regression, Incremental Association Markov Blanket (IAMB) and Grow-Shrink Markov Blanket (GS), in term of their ability to generate robust results. IAMB outperformed the other two methods and was applied to identify active compounds from one well-known Chinese medicinal prescription widely used to treat cardiovascular diseases. The biological relevance of identified compounds was confirmed by literature knowledge and experimental data, and thereby suggested our method be promising to facilitate the drug discovery from medicinal plants.

1. Introduction

The mainstream in pharmaceutical industries is moving away from single molecule or single target approaches to combinations or multiple target paradigm(Wermuth, 2004), and lead to a globally positive trend in favor of natural products for their great chemical diversity and potential to regulate multiple targets(Koehn & Carter, 2005). In particular, traditional Chinese medicine has been using the combination of medicinal plants to achieve synergistically therapeutic effects(Xue & Roy, 2003), and large amounts of prescriptions of combinatorial medicinal plants accumulated in thousands of years, are growing to be a promising source of new multiple-target therapeutics.

Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine learning for Health-Care Applications, Helsinki, Finland, 2008.

Bio-guided isolation has been a dominant strategy to screen active compounds from natural products. However, the large amount of compounds contained in the herbal medicine make the screening process greatly labor-intensive and time-consuming (Pieters & Vlietinck, 2005). Moreover, it is a single-target approach and can not account for the synergistic effects in Chinese medicinal plants, thus probably fail to identify active compounds from Chinese medicinal plants. Hence, alternative strategies should be established to take into count the characteristics of Chinese herbal medicine.

Different with bio-guided isolation, we proposed to investigate the relationship between the chemical composition (including the amounts and proportions of compounds) and biological activity of medicinal plants, and identify the compounds relevant to biological activity as candidates for drug development. Essentially, the identification of active compounds could be attributed to a causal variable selection problem, since these active compounds cause the biological activities of medicine plants. Meanwhile, numerous constituents of medicinal plants make chemical analysis data (characterizing the chemical composition) high-dimensional, whereas the high-costs or ethic issues of pharmacological studies lead to limited sample size, thus giving rise to the challenge of high-dimensional and small-sample. Recently developed Markov Blanket discovery algorithms which are feasible to causal discovery with high-dimensional and small sample might be particularly useful to address this problem (Tsamardinos I, Aliferis CF, & A, 2003).

In this paper, we examined two Markov blanket induction algorithms and compared to stepwise regression in term of their robustness in analyzing our real datasets, and applied the best one to identify the active compounds from one well-known Chinese medicinal prescription ("Xue-Fu-Zhu-Yu decoction") used for the prevention and treatment of atherosclerosis (Zhang & Cheng, 2006).

2. Material and Methods

2.1 Experimental data

"Xue-Fu-Zhu-Yu decoction" consists of six medicine plants including Paeonia lactiflora (PL), Ligusticum chuanxiong (LC), Citrus aurantium (CA), Carthamus tinctorius (CT), Prunus persica (PP) and Bupleurum falcatum (BF). The aqueous extract of the prescription was divided into six different components by porous resin chromatography. After that, 32 samples were prepared by mixing the six components based on a experimental design method (Fang, Shiu, & Pan, 1999).

Hyperlipidemic rat model (Gao et al., 2002) was used in pharmacological experiments, and the concentrations of total cholesterol (TC) in rat plasma were monitored to evaluate the biological activities of the 32 samples.

Liquid chromatographic technique (HPLC) was used to characterize the chemical composition of the 32 samples (Liu, Cheng, & Zhang, 2004), 33 chromatographic peaks were commonly present in the chromatograms of the 32 samples. See Fig.1 for an example, in which each chromatograph peak represents a compound, the absolute area of each peak was calculated to quantify the content of the compound.

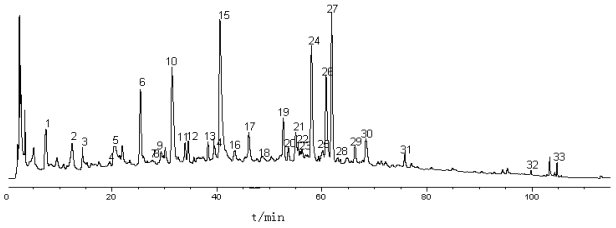

Figure 1 Chromatogram of "Xue-FU-Zhu-Yu decoction"

2.2 Algorithms

Two algorithms for Markov blanket induction, Grow-Shrink Markov Blanket (GS) (Margaritis & Thrun, 1999) and Incremental Association Markov Blanket (IAMB) (Tsamardinos I et al., 2003) were applied in our study. GS algorithm is the first algorithm developed for discovering Markov blanket, its framework includes growing phase and shrinking phase. IAMB follows the same two-phase structure with GS algorithm and adopts one dynamic heuristic in the growing phase to improve the static and potentially inefficient heuristic of GS. In detail, IAMB iteratively reorders the variables after a new variable enters the blanket, and the reordering operation is implemented using mutual information heuristic.

Fisher's z-test was used in these two algorithms, since the variables in our datasets are continuous. Causal Explorer toolbox (Aliferis CF, Tsamardinos I, Statnikov A, & LE, 2003) was used to implement Markov blanket discovery algorithms. We implemented stepwise regression in Matlab7.0 and then compared to GS and IAMB. All the three algorithms need to set up the p-value cutoff for statistical test, in our analysis, p-value cutoffs (α) were set to be 0.01, 0.05, 0.1 and 0.15, respectively.

2.3 Measure of algorithm robustness

Robustness is an important issue for variable selection algorithms in the case of high-dimensional and small sample. We evaluated the robustness of algorithms by by an instability index calculated from the result of leave-one-out cross validation (LOO), which was defined as following:

$$Instability\_index = -\sum_{i=1}^{n} \frac{Nx_i}{N} log(\frac{Nx_i}{N})$$

N represents the number of instances; N_{xi} is the times selected in LOO procedure for ith variable x_i.

The metric can reflect the total variation of all variable subsets during the LOO process. Theoretically speaking, the lower the entropy metric, the more robust the algorithm, and this metric would be set to be zero when no variable was chosen in LOO procedure.

3. Results

3.1 Robustness of algorithms

The robustness of algorithms was investigated at four significant levels and the instability index of stepwise regression, GS and IAMB were depicted in Figure2.

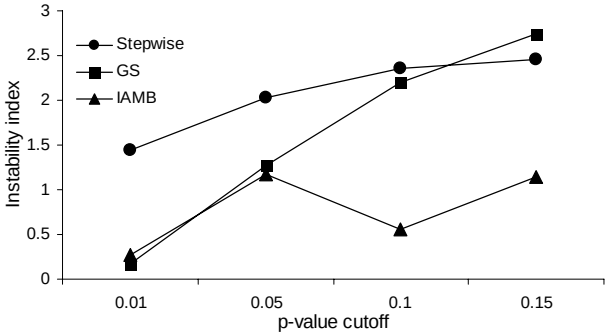

Figure 2 Robustness of the three algorithms

From figure2, GS and IAMB yielded nearly equivalent results when p-value cutoff was set to be 0.01 and 0.05. The potential reason was that GS would induce less false-positive variables into Markov blanket under more strict independence test condition. In the meantime, more variables would be selected and the influence of sample variations on the algorithms would be amplified, when the confidence level of statistical test increased, thereby, GS tended to become more sensitive along with increasing p-value cutoff. In addition, the instability index of IAMB was always lowest, which indicated that this algorithm might be more stable than the other two algorithms.

To further look into the stability of the three methods in LOO procedure, the variables selected by IAMB, GS and stepwise regression were given in table1, table2 and table3, respectively. From the three tables, we can see that when α increased, stepwise tend to introduce more new variables, but not increase the frequencies of introduced variables. GS behaved differently, tending to increase the frequencies of selected variables and introduce new variables. In contrast, IAMB tended to increase the frequencies of selected variables and introduced few new variables, and thus be more convenient to make decisions which chromatographic peaks should be selected for further validation.

Table 1 computational results of IAMB

α	12#	16#	21#	24#	25#	29#	30#	31#	32#
0.01						1		1	31
0.05	1		13			19		19	31
0.1	1		15		1	30		30	31
0.15	1	1	30	1	4	29	4	30	31

Table 2 computational results of GS

α	3#	12#	14#	15#	19#	20#	21#	22#	23#
0.01									
0.05		1					4		1
0.1		7		1	1		17	1	1
0.15	1	16	2			1	10	2	18
α	24#	25#	26#	27#	28#	29#	30#	31#	32#
0.01						1			31
0.05		15		2	1	1			31
0.1		15		12		1		1	31
0.15	1	4	1	6	1		1	1	31

Table 3 computational results of stepwise regression

α	1#	2#	4#	12#	16#	21#
0.01	4	2	1			
0.05	4	2	1		1	1
0.1	4	2	1	1	1	1
0.15	4	2	1	1	1	1
α	25#	29#	30#	31#	32#	33#
0.01		1		1	14	17
0.05		1	1	10	14	17
0.1		16	1	17	14	17
0.15	1	16	1	17	14	17

3.2 Active compounds identified

21#, 29#, 31# and 32#chromatographic peaks could be regarded as active compounds. A chemical analysis method (HPLC-DAD-ESI-MS analysis) was performed to determine the compounds identified (Liu et al., 2004), and the chromatograms of the prescription were compared with that of individual herbs to determine the original herb of these compounds, the result see table 4.

Table.4. HPLC-DAD-ESI-MS Identification

Peak	$[M+H]^+$	Other ions (ESI^+)	$[M-1]^-$	Other ions (ESI^-)	Identify	Plant material
21	1043	–	–	–	Safflor yellow	CT
29	595	617	593	629,653	Poncirin	CA
31	273		271		Naringeninb	CA
32	585	607	583	643,1167	Benzoylpaeoniflorin	PL

The chemical analysis results showed that peak 21# was safflor yellow from in the flower petals of *Carthamus tinctorius*, peak 29#、31# were respectively poncirin and naringenin from *Citrus aurantium*, peak 32# was Benzoylpaeoniflorin from *Paeonia lactiflora*. The former three compounds were all flavonoids and the latter compound was monoterpene glucosides, these types compound potentially contribute to the therapeutic effects on cardiovascular diseases.

These identified active compounds have been validated to contribute to the therapeutic effects of decreasing total cholesterol. In detail, safflor yellow (21#) and poncirin (29#) were reported to be beneficial for decreasing the total cholesterol (Monforte et al., 1995). Naringenin (31#) and Benzoylpaeoniflorin (32#) were found to have the effects of anti-oxidation and scavenge hydroxyl radicals which can contribute to reduce the level of cholesterol (Liu et al., 2004).

4. Conclusion

In our work, the results indicated that Markov blanket discovery algorithm was feasible to discover active compounds from chemical analysis and pharmacological data of Chinese medicine plants. Chromatographic fingerprint to obtain chemical analysis data, could holistically measure the chemical compositions of herbal medicine, which provided the opportunity for investigating the relationships between multiple compounds and pharmacological activities, and thus alleviate the drawbacks of single-target screening utilized by bio-guided isolation. In the meantime, the paradigm focused on computational analysis of the dataset from chemical analysis and pharmacological experiment, thus provided a promising framework to develop efficient computer-aided screening methods for botanical drug

discovery.Many Chinese medicine formulas have been showing their therapeutic effects in the last thousands years, based on which the hit rate of drug discovery will be potentially much higher than “blind” screening.

However, some compounds with very low contents might be neglected by currently chromatographic analysis techniques. With the improvement of chemical analysis techniques, much more compounds could be quantified in future, and more potential active compounds could be identified by our method. For now, we just applied the causal variable selection to one Chinese medicinal prescription, and the applications to more prescriptions will help to convince the utility of the methods in this area. Furthermore, it’s also expected to extend our method to take into account the cooperation of different compounds, which is deemed as important for the therapeutic effects of Chinese medicine prescriptions.

Acknowledgments

This project was financially supported by the Chinese National Basic Research Priorities Program (No.2005CB523402) and a key grant from the National Natural Science Foundation of China (No. 90209005). We are highly appreciated Dr. C. Aliferis provide the causal explorer toolbox.

References

Aliferis CF, Tsamardinos I, Statnikov A, & LE, B. (2003). Causal Explorer: a causal probabilistic network learning toolkit for biomedical discovery, *Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*.

Fang, K. T., Shiu, W. C., & Pan, J. X. (1999). Uniform designs based on Latin squares. *Statistica Sinica, 9*(3), 905-912.

Gao, Y., Yu, W. G., Han, F., Lu, X. Z., Gong, Q. H., Hu, X. K., et al. (2002). [Effect of propylene glycol mannate sulfate on blood lipids and lipoprotein lipase in hyperlipidemic rat]. *Yao Xue Xue Bao, 37*(9), 687-690.

Koehn, F. E., & Carter, G. T. (2005). The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery, 4*(3), 206-220.

Liu, L., Cheng, Y., & Zhang, H. (2004). Phytochemical analysis of anti-atherogenic constituents of Xue-Fu-Zhu-Yu-Tang using HPLC-DAD-ESI-MS. *Chem Pharm Bull (Tokyo), 52*(11), 1295-1301.

Margaritis, D., & Thrun, S. (1999). Bayesian Network Induction via Local Neighborhoods, *Advances in Neural Information Processing Systems*. Denver, Colorado.

Monforte, M. T., Trovato, A., Kirjavainen, S., Forestieri, A. M., Galati, E. M., & Lo Curto, R. B. (1995). Biological effects of hesperidin, a Citrus flavonoid. (note II): hypolipidemic activity on experimental hypercholesterolemia in rat. *Farmaco, 50*(9), 595-599.

Pieters, L., & Vlietinck, A. J. (2005). Bioguided isolation of pharmacologically active plant components, still a valuable strategy for the finding of new lead compounds? *Journal of Ethnopharmacology, 100*(1-2), 57-60.

Tsamardinos I, Aliferis CF, & A, S. (2003). Algorithms for large scale Markov blanket discovery, *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference FLAIRS* (pp. 376-381).

Wermuth, C. G. (2004). Multitargeted drugs: the end of the 'one-target-one-disease' philosophy? *Drug Discovery Today, 9*(19), 826-827.

Xue, T. H., & Roy, R. (2003). Studying traditional Chinese medicine. *Science, 300*(5620), 740-741.

Zhang, H. J., & Cheng, Y. Y. (2006). An HPLC/MS method for identifying major constituents in the hypocholesterolemic extracts of Chinese medicine formula 'Xue-Fu-Zhu-Yu decoction'. *Biomed Chromatogr, 20*(8), 821-826.