Journal of Biological Systems, Vol. 17, No. 3 (2009) 329–347

World Scientific
www.worldscientific.com

NEW DRUG R&D OF TRADITIONAL CHINESE MEDICINE: ROLE OF DATA MINING APPROACHES

HONGJUN YANG,*,// JIANXIN CHEN,†,// SHIHUAN TANG,* ZHENKUN LI,* YISONG ZHEN,‡ LUQI HUANG*,¶ and JIANQIANG YI§

*Institute of Chinese Materia Medica
China Academy of Chinese Medical Sciences
No. 16 Nanxiaojie, Dongzhimennei, Beijing, 100700, P. R. China

†Beijing University of Chinese Medicine
Beijing, 100029, P. R. China

‡Key Laboratory for Clinical Cardiovascular Genetics of the Ministry of Education
FuWai Hospital and Cardiovascular Institute
Chinese Academy of Medical Sciences and Peking Union Medical College
Beijing, P. R. China

§The Key Laboratory of Complex Systems and Intelligence Science
Institute of Automation, Chinese Academy of Sciences
Beijing, 100190, P. R. China
¶huangluqi@263.net

Invited 19 November 2008
Received 24 April 2009
Accepted 20 May 2009

Traditional Chinese Medicine (TCM) documented about 100,000 formulae during past 2500 years. To use and customize them by modern pharmaceutical industry, we make an interdisciplinary effort to study the activity of new drug research and development (R&D) in TCM by introducing data mining approaches to it. We used the migraine formulae as a training set to investigate the possibility of developing new prescription by means of data mining. The activity of new drug R&D of TCM consists of two steps. The first step is to discover new prescriptions (candidates for drugs) from migraine formulae. We present an unsupervised clustering approach based on data mining theory to address the problem in the first step and automatically discover ten new prescriptions from the formulae data. The second step is to develop and optimize the prescriptions discovered by current biomedical approaches. Since *Ligusticum chuanxiong Hort* (LCH), a kind of herb, is often used to treat migraine and appears in the new prescriptions, we use it as an example and apply supervised regression method based on data mining theory to study the drug R&D activity of TCM. We revised two linear regression methods in order to establish the nonlinear association between three chemical ingredients of LCH and corresponding pharmacological activity and used it to predict the activities. The association is validated by *in vitro* experiments and we found that the experimental results are consistent with the prediction. Unsupervised clustering and supervised regression cover most part of data mining theory, which means that data mining approaches play

¶Corresponding author.
// These authors contribute equally to the work.
Mailing address for delivery of offprints: Dr. Jianxin Chen, Information Center, Beijing University of Chinese Medicine, 11 Bei San Huan Dong Lu, Chao Yang District, Beijing 100029, P. R. China.

a crucial role in new drug R&D in TCM and present a better solution to establish the platform of drug R&D in TCM.

Keywords: Drug Discovery; Traditional Chinese Medicine; Data Mining; Unsupervised Cluster; Supervised Regression; Formulae; Association; Mutual Information.

1. Introduction

Traditional Chinese Medicine (TCM) is a unique medical knowledge system in China with a history of clinical practice spanning centuries. TCM is widely accepted and belongs to the main stream tools toward health and wellness in East Asia. In TCM, combinational herbs called formulae are used to deal with diseases for ancient Chinese[1] and nowadays there are 100,000 formulae based on the continuous clinical records. There are mainly two kinds of research efforts to study and utilize those formulae that have therapeutic efficacy: one is to uncover the working mechanism of them by the-state-of-art biomedical methods;[1,2] the other is to use certain formulae to discover new prescriptions that hidden in the formulae data. A formula is a prescription that is validated by pharmacology and clinics. The two research directions belong to drug R&D of TCM. It is found that data mining approaches play critical roles in the modern drug discovery activities.[3] Therefore, data mining approaches are believed to provide solutions to some complex research problems occurred during the R&D activities of TCM.

Data mining is a systematic approach used not only to identify biomarkers for a disease but also to investigate the cellular interaction in the context of a disease to construct biological networks.[3,4] Data mining also has a crucial role to play in TCM-related drug R&D activities. By text mining, a branch of data mining approaches, the biological networks underlying cold and hot syndromes phenotypes are constructed by NEI specifications.[5] Similarly, through a combination of Chinese literatures on TCM and related English counterparts on most diseases on PubMed database, biological networks for a syndrome in TCM in the context of a disease can be automatically generated through text mining approaches.[6] In addition, several novel data mining approaches were presented to deal with various kinds of clinical or *in vivo* animal data. An unsupervised cluster algorithm called pattern discovery algorithm was developed to discover syndromes in TCM in the context of a disease, which provides the targets for formulae or prescriptions since they are prescribed based on syndromes diagnosed.[7] Furthermore, animal models for syndrome in TCM in the context of diseases were built by using supervised data mining approach to 'clone' diagnosis criterion from clinics to animals, which paves a way for *in vivo* experimental validation of a prescription.[8] However, when applying data mining approach in TCM, few research efforts are made in new drug R&D activities of TCM, it is important to investigate the role of data mining approaches in them. In this paper, we investigated two crucial steps of new drug R&D in TCM by applying data mining methods to them and found that they covered two aspects of data mining theory, i.e. unsupervised clustering and supervised regression (or classification).

We present a clustering method and revised two regression methods to fit the characteristics of data in the two steps of R&D. Our results show that unsupervised and supervised data mining approaches play a crucial role in the two steps respectively and contribute significantly to new drug R&D in TCM.

The paper is organized as the following. Section 2 is devoted to presenting a hierarchical clustering algorithm to discover novel prescriptions in TCM from formulae data. Section 3 focuses on how to develop the prescriptions from pharmacology by revision of two regression methods. Section 4 concludes the paper and discusses the role of data mining approaches in the new drug R&D in TCM.

2. Unsupervised Prescription Discovery

Prescription discovery plays a most important role in new drug R&D of Chinese medicine. Prescription is defined as combinations of herbs with proper proportion. There are more than one million formulae recorded during past 3000 years of TCM practice; most formula are still used in clinics. Prescription discovery is based on a hypothesis that new prescriptions are hidden in data composed of all formulae. Therefore, prescription discovery is performed by presenting data mining methods to retrieve useful information through complex computation from formula database. Here, we use formulae which treat migraine as an example to substantiate the important role of data mining methods in prescription discovery.

2.1. *Construction of a formula database*

We collect 541 formulae which are used in clinics for treatment of migraine from TCM literatures and 197 herbs which are components of formula. Formula and herbs constitute a 541×197 matrix (data). Each row in the data represents a formula and each column represents a herb. If a herb appears in a formula, the corresponding value of the matrix is denoted as '1', otherwise, the counterpart is '0'. The data has only 3332 grids with a positive value in the matrix, given total of $106577(541 * 197)$ grids, which means that the data are only about 3% positive frequency and it is therefore a sparse matrix. Thus, it poses a challenge to data mining methods that cope adequately with the sparse matrix.

2.2. *Unsupervised hierarchical pattern discovery algorithm of prescription discovery*

The goal of data mining here is to discover several combinations of herbs in a self-organized way. These novel combinations, considered as potential new prescriptions, would not appear in the data and the number of them is unknown before mining. Prescription discovery is thus the goal of unsupervised data mining. However, according to the theory of TCM, most herbs that appear in formulae play distinctive role in different formulas. The special character of prescription discovery eliminates a kind of cluster methods that realize so-called 'hard cluster', such as cluster

analysis, i.e. the kind of methods cannot realize that one herb appears in different prescriptions. Furthermore, the data is composed of categorical variables (only two categories, '0' and '1'), which greatly restricts the inclusion of cluster methods.

Here, we present a hierarchical pattern discovery algorithm to address the problems above. The algorithm, a kind of clustering method, is a revised version of pattern discovery algorithm that was developed by us[7] and is based on association delineated by mutual information (MI). It is acknowledged that mutual information contains better characteristics to measure association between categorical variables. Pattern discovery algorithm is successfully applied to discover several patterns from chronic renal failure data and realize some variables that appear in different patterns.[7]

Initially, we use unsupervised pattern discovery to discover several patterns from the migraine data. The corresponding result is given in Table 1, from which we can see that 63 patterns are retrieved by the algorithm. The maximal pattern is clusters with five herbs and the minimal pattern within a cluster have three herbs. However, when each pattern is returned to the data and counted its frequency in the total of 541 formulas, from Table 1, it is found that each pattern has a positive frequency in the data and the maximal is 74 and the minimal is one. This means that the detected 63 patterns are not the new prescriptions we need for further experimental validation. Despite this, the retrieved patterns from the data are considered as core herbs' combinations for migraine since they are obtained in a self-organized way by the pattern discovery algorithm. Prescriptions may be recombination of these 63 patterns. Therefore, hierarchical pattern discovery algorithm is developed to investigate association between patterns and to discover prescriptions.

How to precisely measure association between the patterns plays a key role in the hierarchical pattern discovery algorithm. Here, we present an encoding method for each pattern and then use mutual information to measure association between the patterns. Based on the associations, we present hierarchical pattern discovery algorithm to automatically retrieve prescriptions. Moreover, we also present a novel method to evaluate each prescription by ranking their significance in the data.

2.2.1. *Encoding each pattern in the data*

In order to measure mutual information between the 63 retrieved patterns, each pattern is needed to be encoded as a categorical variable. Since each herb variable in the data is of binary variable, a pattern with N variables is encoded as (2^N) categories. Therefore, encoding the pattern is done by assigned a number values between 1 and (2^N) to each category. The advantage of mutual information is that it can measure two variables with different number of categories. Most importantly, mutual information is independent from the numbers that assigned to categories of a pattern. The following mathematical proof is devoted to demonstrating the good performance of mutual information to measure association between patterns here.

Table 1. The 63 patterns discovered by the pattern discovery algorithm. Their frequencies and significances are calculated from the formulae data.

No.	Pattern					Fre.	Sig.
1	*ZhiCaoWu*	*MoYao*	*WuLingZhi*	*MaQiangZi*	*CanSha*	2	0.7611
2	*BaiQiang*	*BaiBu*	*DaQinYe*	*ZiyYuan*		1	1
3	*BoZiRen*	*GouTeng*	*GuiJia*	*MuDanPi*		1	0.6667
4	*CheQianZi*	*JuHong*	*YinChen*	*JiuHuangQi*		1	1
5	*ChenXiang*	*GuaLouZi*	*QinDai*	*ShiHu*		1	1
6	*ChenXiang*	*GuaLouZi*	*ShaRen*	*ShiHu*		1	0.8333
7	*ChuanXiong*	*HaoBen*	*ManJingZi*	*XiXin*		16	−0.0093
8	*ChuanShanJia*	*MoYao*	*WuLingZhi*	*MaQiangZi*		1	0.5
9	*DouKou*	*ZeXie*	*ZhuLing*	*ShenQu*		1	0.8333
10	*HaQiao*	*GuaLouZi*	*QinDai*	*ShiHu*		1	1
11	*JiangCan*	*QuanXie*	*ZhiChuanWu*	*ZhiTianNanXing*		5	−0.4135
12	*JiangCan*	*QuanXie*	*ZhiChuanWu*	*XiongHuang*		5	0.0288
13	*JuHua*	*ShiGao*	*XuanFuHua*	*ZhiKe*		7	0.0903
14	*KuShen*	*WuQiaoShe*	*ShuiNiuJiao*	*LongGu*		2	1
15	*KuShen*	*WuQiaoShe*	*YuLiRen*	*LongGu*		2	1
16	*RuXiang*	*MoYao*	*WuLingZhi*	*MaQianZi*		3	0.9444
17	*BaiQian*	*BaiBu*	*WuWeiZi*			1	1
18	*BaiShao*	*CaiHu*	*DangGui*			7	−0.47619
19	*BaiShao*	*DangGui*	*DiHuang*			8	−0.041667
20	*BaiShu*	*BanXia*	*ChenPi*			5	−0.66667
21	*BaiShu*	*ChenPi*	*FuLing*			5	−0.53333
22	*BaiShu*	*FuLing*	*RenShen*			7	−0.28571
23	*BaiShu*	*HuangQi*	*RenShen*			8	0.41667
24	*BaiZhi*	*BoHe*	*JingJie*			17	0.11765
25	*BanXia*	*TianMa*	*ZhiTianNanXing*			16	0.29167
26	*BingPian*	*ZhiBaiHuZi*	*ZhuSha*			5	0
27	*CangShu*	*HongHua*	*HuangBo*			2	0.33333
28	*CangShu*	*HuangBo*	*ShengMa*			5	0.6
29	*CaiHu*	*HuangLian*	*HuangQin*			4	−0.75
30	*ChangTui*	*ShiJueMing*	*SheTui*			1	0.66667
31	*ChuanBeiMu*	*DanShen*	*XiaKuCao*			1	1
32	*ChuanNiuXi*	*RouCongRong*	*DingXiang*			1	1
33	*ChuanNiuXi*	*SheXiang*	*DingXiang*			1	0.66667
34	*ChuanNiuXi*	*FangFeng*	*GanCao*			67	0.42289
35	*ChuanQiong*	*FangFeng*	*HaoBen*			25	−0.26667
36	*ChuanQiong*	*GanCao*	*XiXin*			74	0.56757
37	*DaiHuang*	*JuHe*	*QingMengShi*			1	1
38	*DanShen*	*JiangXiang*	*XiYangShen*			1	1
39	*DiGuPi*	*XuanShen*	*ShiHu*			1	0.66667
40	*DiHuang*	*HuangLian*	*HuangQin*			5	0.13333
41	*DouKou*	*QingPi*	*FuLing*			1	1
42	*FangFeng*	*GanCao*	*QiangHuo*			44	0.47727
43	*FangFeng*	*HaoBen*	*QiangHuo*			15	−0.55556
44	*FangFeng*	*JingJie*	*QiangHuo*			25	0.17333
45	*FuLing*	*QianHu*	*ZhiQiao*			5	0.4
46	*FuLing*	*RenShen*	*ShuiNiuJiao*			5	−0.066667
47	*HaQiao*	*YuanZhi*	*ZiYuan*			1	1
48	*HeShouWu*	*TianHuaFen*	*CanSha*			2	1
49	*HouPu*	*MaiYa*	*ShanZha*			2	1
50	*HuangBo*	*HuangQi*	*ShengMa*			5	0.53333
51	*HuoMaRen*	*LaiFuZi*	*MuZei*			1	1
52	*JiangXiang*	*MaiYa*	*XiYangShen*			1	1
53	*JieGeng*	*JuHe*	*QingMengShi*			1	1

Table 1. (*Continued*)

No.	Pattern			Fre.	Sig.
54	*LaiFuZi*	*MaiYa*	*ShanZha*	1	0.66667
55	*MaiYa*	*ShanZha*	*ShenQu*	1	0.66667
56	*QingDai*	*WuGong*	*YuJin*	1	1
57	*QuanXie*	*ZhiBaiFuZi*	*ZhiTianNanXing*	9	0.18519
58	*RouGui*	*ShuiZhi*	*ZhaoJiao*	1	0.66667
59	*ShanYao*	*ShanZhuYu*	*ShouDiHuang*	3	0.33333
60	*ShanYao*	*ShanZhuYu*	*YuZhu*	1	−1
61	*ShiJueMing*	*XiaKuCao*	*LingYangJiao*	1	1
62	*TianMa*	*ZhiBaiFuZi*	*ZhiTianNanXing*	8	−0.5
63	*ZhiBaiFuZi*	*SheXiang*	*ZhuSha*	7	0.42857

Given a pattern, denoted as X has N variables, the other pattern Y has M variables, definition of mutual information between X and Y is given by Eq. (1).

$$MI(X,Y) = H(X) + H(Y) - H(X \cup Y), \tag{1}$$

where $H(X)$ and $H(Y)$ denote entropy of X and Y respectively, and $H(X \cup Y)$ denote joint entropy of X and Y. They are well-defined by following equations:

$$H(X) = \sum_{i=1}^{2^N} P(X=i) \log P(X=i), \tag{2}$$

$$H(Y) = \sum_{i=1}^{2^M} P(Y=i) \log P(Y=i), \tag{3}$$

$$H(X \cup Y) = \sum_{i=1}^{2^N} \sum_{j=1}^{2^M} P(X=i, Y=j) \log P(X=i, Y=j). \tag{4}$$

When encoding the pattern X, given two distinctive assigned numbers k and q, if they are interchanged, then the definition of X would turn to be

$$\begin{aligned} H(X) &= \sum_{i=1}^{2^N} P(X=i) \log P(X=i) \\ &= \sum_{i \neq k,q} P(X=i) \log(P(X=i)) + P(X=k) \log P(X=k) \\ &\quad + P(X=q) \log P(X=q) \\ &= \sum_{i \neq k,q} P(X=i) \log(P(X=i)) + P(X=q) \log P(X=q) \\ &\quad + P(X=k) \log P(X=k) \\ &= \sum_{i=1}^{2^N} P(X=i) \log P(X=i) \\ &= H(X). \end{aligned} \tag{5}$$

It is found that $H(X)$ is independent from numbers that are assigned to categories of the pattern X.

Moreover, the form of joint entropy is turned to be

$$\begin{aligned}
H(X \cup Y) &= \sum_{i=1}^{2^N} \sum_{j=1}^{2^M} P(X=i, Y=j) \log P(X=i, Y=j) \\
&= \sum_{j=1}^{2^M} \sum_{i=1}^{2^N} P(Y=j, X=i) \log P(Y=j, X=i) \\
&= \sum_{j=1}^{2^M} \left(\sum_{i \neq k,q} P(Y=j, X=i) \log P(Y=j, X=i) + P(Y=j, X=k) \right. \\
&\quad \left. \times \log P(Y=j, X=k) + P(Y=j, X=q) \log P(Y=j, X=q) \right) \\
&= \sum_{j=1}^{2^M} \left(\sum_{i \neq k,q} P(Y=j, X=i) \log P(Y=j, X=i) + P(Y=j, X=q) \right. \\
&\quad \left. \times \log P(Y=j, X=q) + P(Y=j, X=k) \log P(Y=j, X=k) \right) \\
&= \sum_{j=1}^{2^M} \sum_{i=1}^{2^N} P(Y=j, X=i) \log P(Y=j, X=i) \\
&= H(X \cup Y). \qquad (6)
\end{aligned}$$

Therefore, the mutual information of X and Y is independent from numbers assigned to categories of the pattern X. That is to say, the encoding method presented here fits to define a pattern as a categorical variable as well as the measure of mutual information between patterns is of unique value, which form a basis for investigation of further association between patterns.

2.2.2. *Hierarchical pattern discovery algorithm*

Based on encoding method described above, the 63 patterns retrieved by the pattern discovery algorithm are treated as 63 categorical variables to be further clustered. Thanks to merit of mutual information, association between each pattern pair can be measured precisely. In order to avoid patterns discovered by hierarchical pattern discovery that still has positive frequency in the data, we revised the form of mutual information as following:

$$MI'(X,Y) = \begin{cases} H(X) + H(Y) - H(X \cup Y) & pro(X,Y) = 0 \\ H(X) + H(Y) - 2 * H(X \cup Y) & pro(X,Y) > 0 \end{cases}, \qquad (7)$$

where $pro(X, Y)$ denotes frequency of combination of pattern X and Y. Once a combination of two patterns has positive frequency in the original data, it means that the combination is full or part of an existing formula and not a novel prescription, therefore, such association would be eliminated.

Hierarchical pattern discovery algorithm is presented in two steps:

Step 1: For each pattern from 63 retrieved patterns, we choose only a pattern that most associated with it by considering value of mutual information.

Step 2: Two patterns, X and Y, are significantly associated if and only if X is most associated with Y as well as Y is most associated with Y. Since the number of patterns is limited (63 here), therefore, the algorithm will quickly converge. All significantly associated patterns are considered as efficient combinations of patterns and these are novel prescriptions.

The novel prescriptions obtained by hierarchical pattern discovery algorithm are given in Table 2; the algorithm automatically retrieves ten novel prescriptions from the 63 patterns. The maximal number of herbs is seven and the least is six.

2.3. *Evaluation of significant patterns*

For unsupervised data mining methods, evaluation of patterns significance also plays a key role in the determination of significant patterns for further analysis.

Table 2. The 10 novel prescriptions are discovered by the hierarchical pattern discovery algorithm presented in the paper, the significances of them are ranked descending. We can see that *ChuanXiong* (LCH) appears in different prescriptions with variant significances.

No.	Novel prescriptions	Sig.	No. of herbs
1	*ChuanQiong, GanCao, XiXin DanShen, JiangXiang, XiYangShen*	0.7838	6
2	*ChuanQiong, HaoBen, ManJingZi, XiXin, ChuangBeiMu, DanShen, XiaKuCao*	0.4954	7
3	*BanXia, TianMa, ZhiTianNanXing, ChanTui, ShiJueMing, SheTui*	0.4792	6
4	*HuangBo, HuangQi, ShengMa, QuanXie, ZhiBaiFuZi, ZhiTianNanXing*	0.3593	6
5	*JiangCan, QuanXie, ZhiChuanWu, XiongHuang, DiGuPi, XuanShen, ShiHu*	0.3477	7
6	*RuXiang, MoYao, WuLingZhi, MaQianZi, BaiShu, FuLing, RenShen*	0.3293	7
7	*CheQianZi, JuHong, YinChen, ZhiHuangQi, BaiShao, CaiHu, DangGui*	0.2619	7
8	*BaiShu, BanXia, ChenPi, RouGui, ShuiZhi, ZhaoJiao*	0	6
9	*BaiZhi, BoHe, JingJie, ShanWao, ShanZhuYu, YuZhu*	−0.4412	6
10	*CaiHu, HuangLian, HuangQin, ZhiBaiFuZi, SheXiang, ShuSha*	−0.1607	6

However, it is hard to evaluate the significance since we do not have information of response variables, especially for categorical variables. Here, we try to present a significance evaluation method to score a pattern. As it is known to us, for a sole variable and a pattern, frequency and mutual information can be used to score them in the data respectively. However, mutual information is not used to evaluate significance since the significance is calculated based on comparison of two groups, often called case and control. Therefore, we need to define available case and control groups for a pattern. It is observed that as the least number of variables in a pattern is three, other patterns can be viewed as combinations of patterns with three variables.

Supposed a pattern P is composed of three variables: A, B and C. Its significance can be evaluated by significance of sum of three combinations (A, B with C, A, C with B and B, C with A). For A and B, we define the two groups called pseudo case and pseudo control as following.

$$\text{Pseudo-case: freq } (A = 1, B = 1, C = 1),$$
$$\text{Pseudo-control: freq } (A = 1, B = 1, C = 0),$$

where freq$(A = 1, B = 1, C = 1)$, Pseudo-control: freq$(A = 1, B = 1, C = 0)$ represents the frequencies of both two combinations in the data of interest. The larger the difference between them, the greater the significance is. Based on this, the significance of pattern P can be defined as:

$$\begin{aligned}\text{sig}(P) = &\frac{\text{freq}(A = 1, B = 1, C = 1) - \text{freq}(A = 1, B = 1, C = 0)}{\text{freq}(A = 1, B = 1, C = 1)} \\ &+ \frac{\text{freq}(A = 1, C = 1, B = 1) - \text{freq}(A = 1, C = 1, B = 0)}{\text{freq}(A = 1, C = 1, B = 1)} \\ &+ \frac{\text{freq}(B = 1, C = 1, A = 1) - \text{freq}(B = 1, C = 1, A = 0)}{\text{freq}(B = 1, C = 1, A = 1)}. \end{aligned} \tag{8}$$

The maximal value of sig(P) is 1, which means that the pattern is significantly associated the variable since they would not associated with other variables. For patterns with more than three variables, their significance of them can be evaluated by the sum of significance of permutation of patterns with three variables. The corresponding results are given in right part of Table 1, from which we can rank the prescriptions discovered in the data.

3. Supervised Prescription Optimization

Once the novel prescriptions are obtained by unsupervised data mining methods, the next step is to focus on optimizing the prescriptions to enhance therapeutic efficacy and reduce adverse effects. It is demonstrated that major chemical ingredients in the formulae or herbs would contribute significantly to related pharmacological activities.[1] However, association between active ingredients and bioactivities are

rarely investigated by data mining methods. Here, the association is established to investigate two important pharmacological issues:

1. how major ingredients interact with each other to produce bioactivities; and
2. optimizing the prescriptions or herbs to produce maximal efficacy and minimal toxicity in level of chemistry.

Supervised data mining methods facilitate access to investigation of the two issues. Regression is a kind of supervised data mining methods and usually employed to establish analytic equations of quantitative variables. Furthermore, by the regression equation, interactions among major ingredients of formulae or herbs can be clearly revealed. Aided by optimization computation algorithm, the optimal quantity of each ingredient can be obtained, which facilitates further experimental validations *in vitro* or *in vivo*.

From the novel prescriptions given in Table 2, we can see that *ChuanXiong*, i.e. *Ligusticum chuanxiong Hort* (LCH) appears in two prescriptions. Indeed, in clinics, LCH is usually used to treat migraine. Furthermore, in the literature data introduced above, the herb occupies more than 50% in all included formulae. From research efforts in phytochemistry, ligustilide (L), ferulic acid (F), and butyl phthalide (B) are the major active ingredients of LCH.[9,10] Since disorder of cerebrovascular contractility and function is core mechanism of migraine, blood vessel relaxation (BVR) can be used as the pharmacological activity to investigate interactions among L, F, and B.

We take LCH as a paradigm to investigate the role of supervised data mining methods in optimizing prescription. Here, we present a framework not only to study association between chemical components of the herbal and pharmacological activity, but also to optimize pharmacological activity to discover the optimal quantity of each ingredient. As shown in Fig. 1, the main part of the framework consists of four sections: experimental design; *in vitro* experiment system; data mining; and experimental validation.

3.1. *Experimental design and uniform design*

We choose uniform design method (UD) to design the experimental scheme for three active ingredients since it can establish association between chemical components and pharmacological activity by less experimental times than other design methods, such as orthogonal design.[11] Using UD, the experimental scheme for three components with nine levels is illustrated in Table 3.

3.2. *In vitro experiment system*

3.2.1. *Experiment animals*

Sprague-Dawley (SD) male rats are provided by medicine department of Peking University, Beijing, China and their weight is 220 ± 20 g.

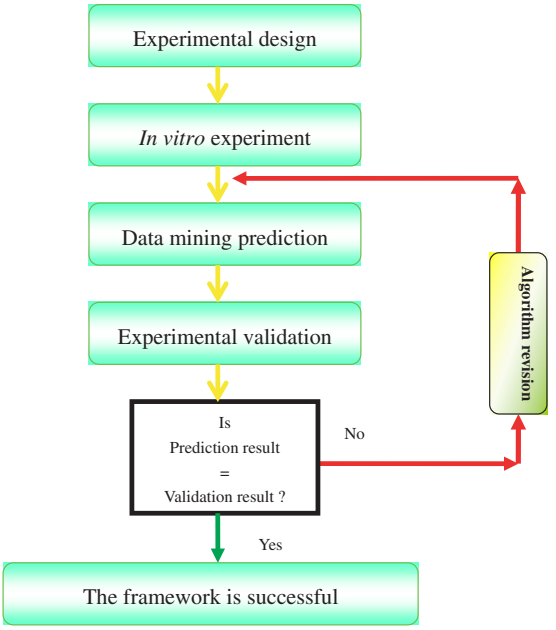

Fig. 1. The framework of optimizing a novel prescription.

Table 3. Experimental scheme for three active ingredients with nine levels generated by uniform design method. Concentration of each ingredient is equally divided into nine levels.

No.	L	F	B
1	9	7	3
2	8	4	6
3	7	1	9
4	6	8	2
5	5	5	5
6	4	2	8
7	3	9	1
8	2	6	4
9	1	3	7

3.2.2. *Experiment reagents*

Ligustilide (size 92%, China Academy of Chinese Medicine, Beijing, China), ferulic acid (size 98%, Nanjing Qingze medicine science and technology company, Jiangsu, China) and butyl phthalide (size 94%, China Academy of Chinese Medicine, Beijing, China). Acetycholine, (Ach, Sigma Company of U.S.A). Krebs-Henseleit (Composed of: NaCl, 118.96 mmo1 $\cdot$ L^{-1}; KCl 4.73 mmo1 $\cdot$ L^{-1}; KH2P041.17 $\cdot$ L^{-1}; MgSO4 1.17 mmol $\cdot$ L^{-1}; NaHCO3 25.0 mmol $\cdot$ L^{-1}; CaCl2 2.54mmol $\cdot$ L^{-1}; Dextrose 11.1mmo1 $\cdot$ L^{-1}). Tartaric acid noradrenalin inject liquid (NE) 2mg/mL (Shanghai Hefeng pharmacy Company, Shanghai, China).

3.2.3. *Experiment equipment*

ALC–M *in vitro* tissue apparatus experiment system (Shanghai ALC biology Science and Technology Company, Shanghai, China). Electrical balance (BP110S, Germany Sartorius Company)

3.2.4. *Experiment methods*

1. Rat thoracic aorta ring preparation method
   SD rats, cervix disjoint, cut thorax, take out chest quickly, place into K-H liquid with 4°, shuck off fat and connective tissue on the thoracic aorta, and chop into blood vessel circles with 2–3 cm lengths. Append the circle in the flume with 5ml K-H liquid. Keep the temperature around 37.0 ± 0.2°C and give off mixed gas with 95% O_2 and 5% CO_2 into the liquid. Fix one end of the circle, connect the other end to ALC–M experiment system with tension energy-exchange equipment, and record variation of tension during experiment process. Start from zero tension, after 30 minutes, tune the tension to 1g and remain for 60 minutes and change the K-H liquid each quarter in the duration. Once thoracic aorta ring remains steady, Kcl with $60\,\text{mmol} \cdot \text{L}^{-1}$ is used to stimulate it; after shrinkage range is steady, use K-H liquid to wash and doff.
2. Measure of blood vessel activity
   Clear the blood vessel circle using K-H liquid several times to keep basic tension; then use KCl ($60\,\text{mmo1} \cdot \text{L}^{-1}$) to induce the maximal shrinkage range of 100%. The blood vessel activity of the three ingredients combinations is measured by Eq. (8), given as following:

$$BVR(\%) = \frac{X_1 - X_2}{X_1 - \text{const}} \times 100, \tag{9}$$

   where X_1 denotes maximal tensility of blood vessel during the experiments, X_2 is responsible for the tensility of blood vessel after adding the chemical component(s), and const is the basis tensility, often is set as 1 g.

3.2.5. *Experiment results*

Uniform design needs nine *in vitro* experiments to guarantee successful association established between three active ingredients and pharmacological activity. The corresponding outcome of each experiment is shown in Table 4.

In order to investigate combinational efficacy of L, F and B, solo efficacy of each ingredient needs to be studied and compared with the former counterpart. We find that there exists an almost linear association between sole ingredient and BVR. The association is given by following three equations.

$$BVR_L = 0.0565 * X_L - 0.5562, \tag{10}$$

$$BVR_F = 0.1190 * X_F + 4.3989, \tag{11}$$

$$BVR_B = 0.1042 * X_B - 0.2452, \tag{12}$$

Table 4. Concentrations of L, F and B are in accord with the uniform design table. The corresponding experimental activities are produced by combinations of three active ingredients.*

No.	L (μg/ml)	F (μg/ml)	B (μg/ml)	Total (μg/ml)	BVR (%)
1	139	106	49	294	55 ± 11.59
2	118	57	94	269	61 ± 9.44
3	112	16	140	268	51 ± 4.56
4	92	125	34	251	44 ± 4.89
5	78	77	82	237	61 ± 6.59
6	69	26	125	220	56 ± 11.02
7	50	140	17	207	31 ± 2.46
8	38	121	65	224	49 ± 2.47
9	22	46	112	180	36 ± 2.99

*There is a minor difference in qualities of design and actual experiment due to metage. The difference does not affect the association established. Part of the data has been published in a Chinese journal.

where $X_i (i = L, F, B)$ and $BVR_i (i = L, F, B)$ denote concentration of each ingredient and corresponding solo pharmacological activity respectively. With these linear associations, the comparison of efficacy can be carried out on five aspects: Solo BVR at total concentration of each of the three ingredients, Statistical BVR calculated by using the concentration of each ingredient in the Table 4, and Experimental BVR. The five aspects are denoted as L, F, B, S and L + F + B respectively. We investigated the five aspects on the nine experiments and found that experimental BVRs are obviously more than anther BVRs, which suggests that interactions among three ingredients are positive and association between them and BVR are nonlinear (Fig. 2).

For example, for experiment No. 1 in Table 4,

$$L = BVR_L|_{X_L=294}, \quad F = BVR_F|_{X_F=294}, \quad B = BVR_B|_{X_B=294}$$
$$S = BVR_L|_{X_L=139} + BVR_F|_{X_F=106} + BVR_B|_{X_B=49}.$$

L + F + B represents the actual experiment BVR. All nine results jointly indicate that there are positive interactions among three ingredients to enhance pharmacological activity.

Moreover, as illustrated in Fig. 3, by ranking nine total concentrations ascendingly, the tendencies of BVR activity is investigated from nine experiments in actual BVR and statistical BVR. We find that, in actual experiments, BVR activity dose not always increase with the raise of total concentration. Alternatively, the maximal BVR in the nine experiments is only 65, far from maximal BVR with 100. Both findings suggest that optimization of BVR activity and optimal concentration among the three ingredients do exist and need to be discovered by supervised data mining methods.

3.2.6. *Supervised data mining methods and revised two regression methods*

The goal of data mining here is to establish accurate association between three ingredients and pharmacological activity, and to optimize the activity. Unlike unsupervised prescription discovery, the association establishment belongs to

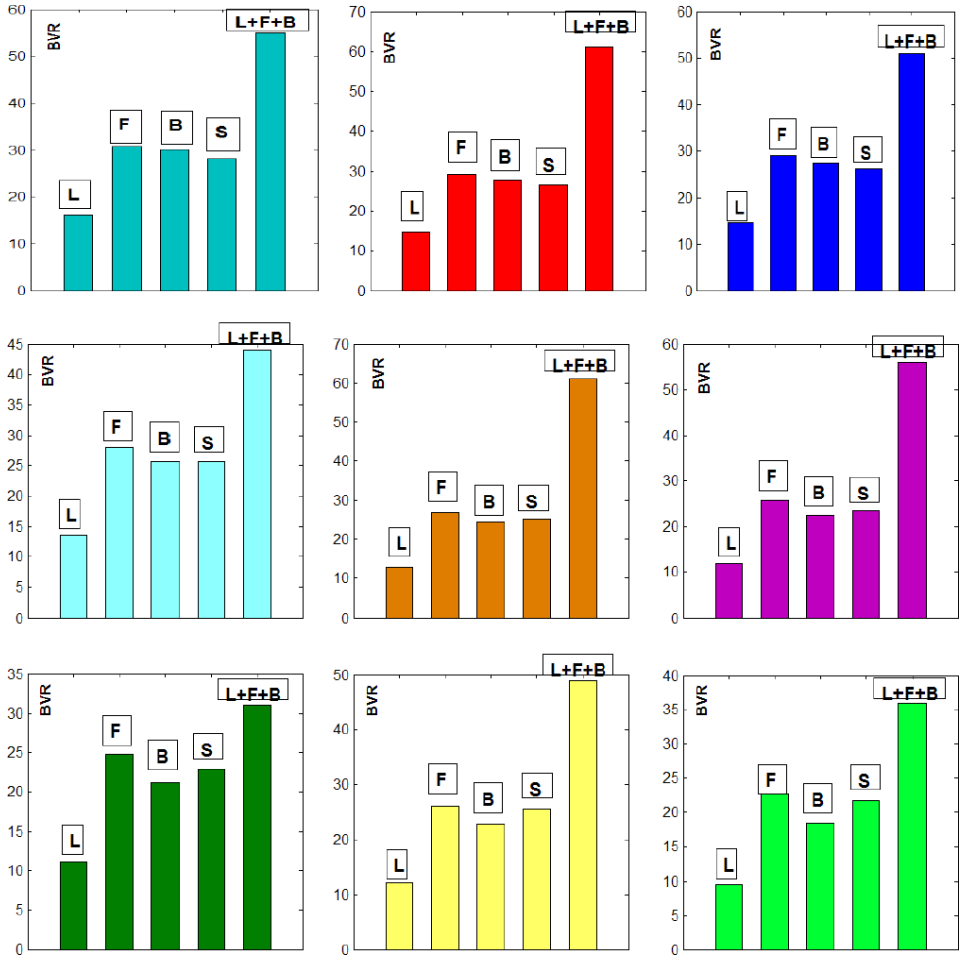

Fig. 2. Comparisons of five BVRs in the nine experiments represented by different colors. L, F and B in the figures denote solo BVR of ingredient L, F and B at total concentration of each experiment. S denotes statistical BVR.

supervised category in data mining methods since response variables are involved during the computation process. Furthermore, the response variable (BVR) and three independent variables (L, F and B) are continuous variables. Regression methods from supervised data mining methods are most fit to deal with the data here. LASSO and Least Angle Regression (LARS) are two linear regression methods that have been successfully applied in other research fields.[12,13] The basic idea of them is to solve the following mathematic expressions.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \quad \text{such that} \quad \sum_{j=1}^{p} |\beta_j| \leq c, \tag{13}$$

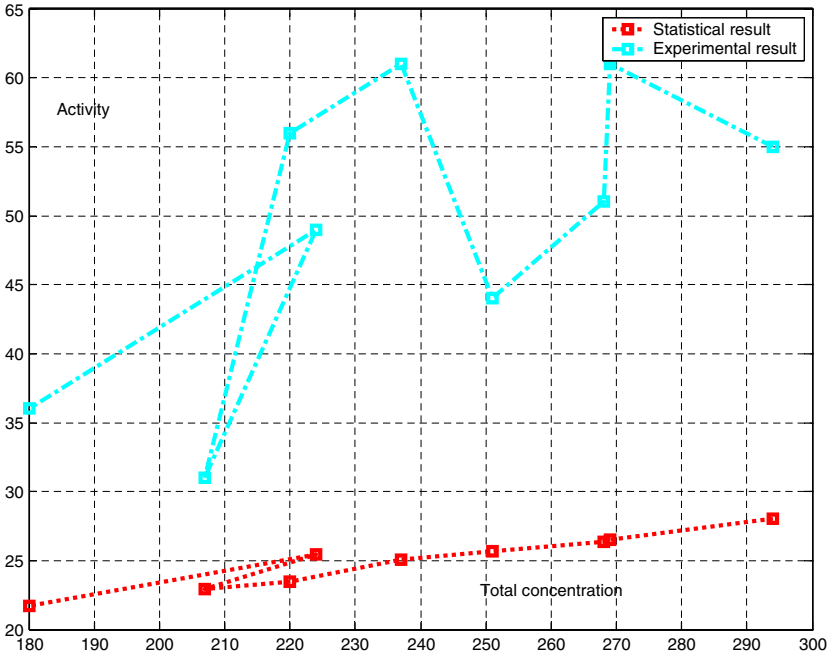

Fig. 3. The total concentration of three chemical components is not consistently increasing with vasodilatation activity in the experimental results (cyan dotted line). But in the statistical data the relation between total concentration and activity is monotone (red dotted line).

Table 5. Nonlinear association is supposed to embody in the seven interactions of the three ingredients.

x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
L^2	F^2	B^2	$L*F$	$L*B$	$F*B$	$L*F*B$

where y denotes BVR and x denotes three ingredients, the second expression means that the regression coefficients are bounded.

However, as we investigate the experimental results above, association between BVR and L, F as well as B is nonlinear. We need to revise the two regression methods to uncover the nonlinear association. We suppose that the nonlinear association between BVR is embodied in seven combinations given in Table 5. Besides L, F, B are denoted as x_1, x_2, x_3 respectively, the other seven are denoted as $x_4 - x_{10}$.

By uniform design, we have nine samples and ten independent variables, since we do not know how L, F and B are interacted with each other to produce BVR, i.e. the number of non-zero regression coefficients is unknown before regression. Fortunately, both LASSO and LARS are so called feature selection based regression methods in the supervised data mining methods,[12] they can select informative variables to establish association with BVR. Furthermore, comparison study is usually

carried out in data mining field to choose better data mining methods to establish association. Here, we also compare the two regression methods and determine the better method by further experimental validation according to predicted results given by them.

The comparison study is shown in Figs. 4–6. Regression error, regression coefficients that construct the expression between the three ingredients and BVR, as well as curves of predictive value and real value of activity, are used to depict the differences between two regression methods when dealing with the data. The better method with better predictive ability is determined in the further experimental validation by comparing the predictive activity with experimental activity.

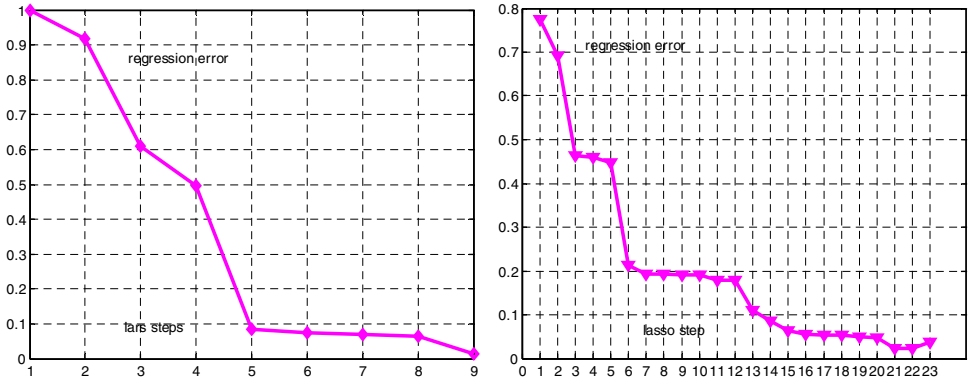

Fig. 4. The feature selection process for two regression methods is slightly different. The number of selection steps for LARS is equal to the number of sample. The counterpart for LASSO is larger than the number of sample since it takes a stepwise way to approach response variables of interest.

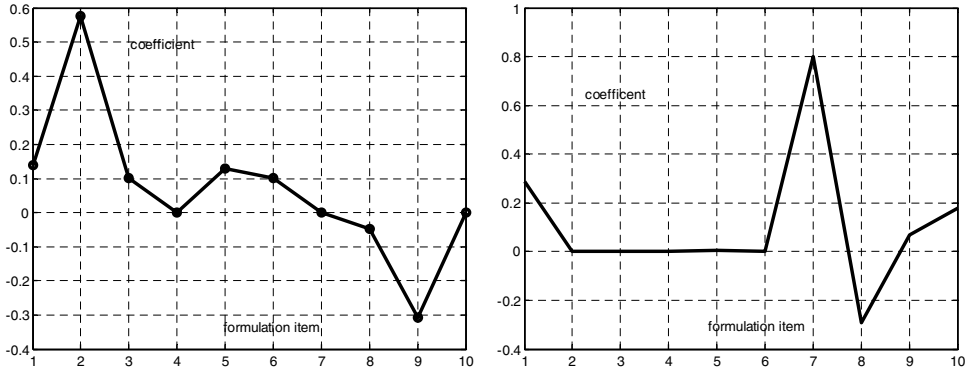

Fig. 5. The coefficients of the model for delineating the association between the three ingredients and BVR are totally different.

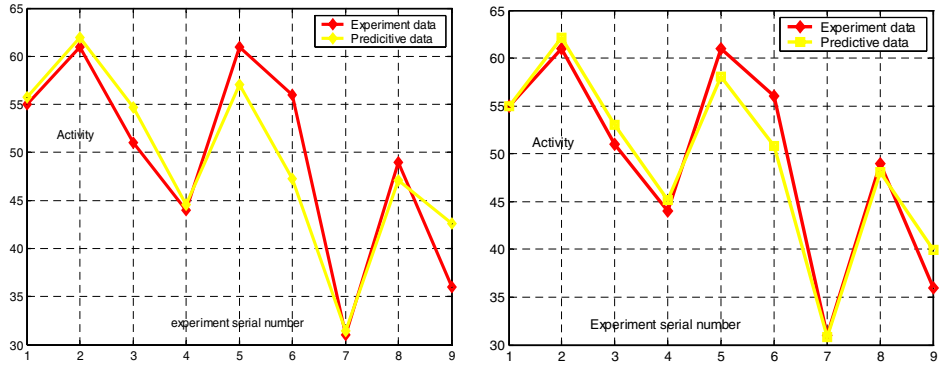

Fig. 6. For LARS, the regression equation is:

$$\text{Activity} = 0.13816 * \text{X} + 0.57593 * \text{Y} + 0.10118 * \text{Z} + 0 + 0.12853 * 0.01 * \text{x} * \text{z} + 0.10215 * 0.01 * \text{y} * \text{z} + 0 - 0.046539 * 0.01 * \text{x}^2 - 0.30718 * 0.01 * \text{y}^2 + 0;$$

For LASSO, the model is

$$\text{Activity} = 0.28604 * \text{X} + 0 * \text{Y} + 0 * \text{Z} + 0 + 0.0062586 * 0.01 * \text{x} * \text{z} + 0 + 0.79875 * 0.0001 * \text{x} * \text{y} * \text{z} - 0.29123 * 0.01 * \text{x}^2 + 0.068596 * 0.01 * \text{y}^2 + 0.17949 * 0.01 * \text{z}^2.$$

Where X, Y, Z denote L, F, B respectively.

3.2.7. *Experimental validation to choose the better regression method*

Through optimization, the optimal combination of three chemical components is: L, 62 μg/ml; F, 70; and B, 150 μg/ml. The corresponding BVR is about 108%. However, in real experiments, it is hard to make BVR reach 100% in ten minutes. Therefore, if the prediction of activity given by the regression is larger than 100%, we consider it as 100%. Otherwise, the validation would induce more errors. From Table 6 we can find that the prediction results given by LASSO accords with the experimental results well, which indicates that the revised regression approach is fitted to deal with the data and the revised LASSO has the ability to establish the nonlinear association between chemical ingredients of prescriptions or herbs and pharmacological activity. Moreover, the data mining approaches significantly enhance the pharmacological activities of the herb. It is believed that data mining methods here greatly contribute to the pharmacological R&D of herb or novel prescriptions.

Table 6. The experiment validation of predicted chemical ingredients combinations. The results show that the LASSO algorithm is better than LARS in discovering association between the three chemical ingredients and corresponding activity.

No.	L	F	B	Total	Experiment	LARS	LASSO
1	70	60	150	280	99.367	64.2299	99.644
2	70	90	120	280	95.313	68.3135	97.948
3	66	81	141	288	96.56	71.4815	100
4	81	121	86	288	95.377	61.1355	94.990
5	71	86	131	288	97.273	70.9917	100
6	62	77	150	289	96.43	71.8395	100

4. Conclusion and Discussion

In this paper, we used a data mining method to solve the problems in the new drug R&D in TCM and find that two important aspects of data mining — unsupervised clustering and supervised regression approaches — play critical roles in the research activity. The hierarchical prescriptions discovery approach for the discovery of new TCM drugs from the formulae data was presented here, and ten novel prescriptions were obtained. It is noteworthy that the proposed approach can cluster a herb into different prescriptions, which is in accordance with the basic theory of TCM. The non-parametric significance of each discovered prescription was calculated to evaluate it. The supervised regression data mining approaches were revised to develop and optimize the prescriptions by current biomedical approaches. The association between chemical ingredients of LCH and corresponding pharmacological activity was clearly revealed by the revised regression approach. Furthermore, the experimental results are consistent with the prediction counterparts by the supervised data mining approach, which demonstrates that the revised regression approach can be used to the new drug R&D in TCM. Based on the sequential two research efforts given in the paper, we can conclude that data mining approaches need to be integrated into the new drug R&D of TCM and should play an important role in the establishment of drug industry of TCM.

Traditionally, it is considered that new drug R&D of TCM is composed of three parts. The first is to discover new drugs that are composed of several herbs. The second is to identify the active chemical ingredients of a drug. Finally, pharmacological experiments are performed to validate the activities of the identified ingredients. The data mining approaches include the first and final steps of new drug R&D in TCM. Indeed, identification of the active chemical ingredients from various drugs also needs to be aided by data mining approaches and the proper mining approaches need to be presented for the intermediate steps of new drugs R&D in TCM.

Acknowledgments

The authors would like to thank the anonymous reviewer, Hang Chang from Lawrence Berkeley National Laboratory and Bing Wang from John Hopkins University School of Medicine for their excellent comments. The authors also wish to thank Hongwei Wu, Geng Li and Chang Chen for assistance with the experiments. The work was supported by National Basic Research Program of China ("973 Program") under grant No.2006CB504700 and NSFC project under grant No. 30600820.

References

1. Wang L, Zhou G-B, Liu P *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**(12):4826–4831, 2008.

2. Li Y, Qu H, Cheng Y, Identification of major constituents in the traditional Chinese medicine "QI-SHEN-YI-QI" dropping pill by high-performance liquid chromatography coupled with diode array detection-electrospray ionization tandem mass spectrometry, *J Pharmaceut Biomed Anal* **47**:407–412, 2008.
3. Yang Y, Adelstein SJ, Kassis AI, Target discovery from data mining approaches, *Drug Discov Today* **14**(3/4):147–154, 2009.
4. Lindsay MA, Target discovery, *Nat Rev Drug Discov* **2**:831–838, 2003.
5. Li S, Zhang X, Wang Y *et al.*, Understanding *Zheng* in traditional Chinese medicine in the context of neuro-endocrine immune network, *IET Syst Biol* **1**(1):51–60, 2007.
6. Zhou X, Liu B, Wu Z *et al.*, Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks, *Artif Intell Med* **41**(2):87–104, 2007.
7. Chen J, Xi G, Chen J *et al.*, An unsupervised pattern (syndrome in traditional Chinese medicine) discovery algorithm based on association delineated by revised mutual information in chronic renal failure data, *J Biol Syst* **15**(4):435–451, 2007.
8. Guo S, Chen J, Zhao H *et al.*, Build and evaluate an animal model for syndrome in traditional Chinese medicine in the context of unstable angina (myocardial ischemia) by supervised data mining methods, *J Biol Syst*, 2009 (accepted for publication).
9. Ru Y, Lia S-L, Chunga H-S *et al.*, Simultaneous quantification of 12 bioactive components of *Ligusticum chuanxiong Hort* by high-performance liquid chromatography, *J Pharm Biomed* **37**:87–95, 2005.
10. Peng C, Xie X, Wang L *et al.*, Pharmacodynamic action and mechanism of volatile oil from *Rhizoma Ligustici Chuanxiong Hort* on treating headache, *Phytomedicine* **16**:25–34, 2009.
11. Hickernell FJ, Liu MQ, Uniform designs limit alising, *Biometrika* **89**(4):893–904, 2002.
12. Efron B, Hastie T, Johnstone I *et al.*, Least angle regression, *Ann Stat* **32**(2):407–451, 2004.
13. Tibshirani, R, Regression shrinkage and selection via the lasso, *J R Stat Soc B* **58**:267–288, 1996.