BMC Systems Biology

RESEARCH Open Access

Revealing metabolite biomarkers for acupuncture treatment by linear programming based feature selection

Yong Wang[1,4*†], Qiao-Feng Wu[2†], Chen Chen[1], Ling-Yun Wu[1,4], Xian-Zhong Yan[3], Shu-Guang Yu[2], Xiang-Sun Zhang[1,4*], Fan-Rong Liang[2*]

From The 5th IEEE International Conference on Computational Systems Biology (ISB 2011)
Zhuhai, China. 02-04 September 2011

Abstract

Background: Acupuncture has been practiced in China for thousands of years as part of the Traditional Chinese Medicine (TCM) and has gradually accepted in western countries as an alternative or complementary treatment. However, the underlying mechanism of acupuncture, especially whether there exists any difference between varies acupoints, remains largely unknown, which hinders its widespread use.

Results: In this study, we develop a novel Linear Programming based Feature Selection method (LPFS) to understand the mechanism of acupuncture effect, at molecular level, by revealing the metabolite biomarkers for acupuncture treatment. Specifically, we generate and investigate the high-throughput metabolic profiles of acupuncture treatment at several acupoints in human. To select the subsets of metabolites that best characterize the acupuncture effect for each meridian point, an optimization model is proposed to identify biomarkers from high-dimensional metabolic data from case and control samples. Importantly, we use nearest centroid as the prototype to simultaneously minimize the number of selected features and the leave-one-out cross validation error of classifier. We compared the performance of LPFS to several state-of-the-art methods, such as SVM recursive feature elimination (SVM-RFE) and sparse multinomial logistic regression approach (SMLR). We find that our LPFS method tends to reveal a small set of metabolites with small standard deviation and large shifts, which exactly serves our requirement for good biomarker. Biologically, several metabolite biomarkers for acupuncture treatment are revealed and serve as the candidates for further mechanism investigation. Also biomakers derived from five meridian points, Zusanli (ST36), Liangmen (ST21), Juliao (ST3), Yanglingquan (GB34), and Weizhong (BL40), are compared for their similarity and difference, which provide evidence for the specificity of acupoints.

Conclusions: Our result demonstrates that metabolic profiling might be a promising method to investigate the molecular mechanism of acupuncture. Comparing with other existing methods, LPFS shows better performance to select a small set of key molecules. In addition, LPFS is a general methodology and can be applied to other high-dimensional data analysis, for example cancer genomics.

* Correspondence: ywang@amss.ac.cn; zxs@amt.ac.cn; acuresearch@126.com
† Contributed equally
[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[2]Acupuncture and Moxibustion College, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China
Full list of author information is available at the end of the article

Background

Acupuncture, an important therapeutic method in Traditional Chinese Medicine (TCM), has been used to treat various diseases for thousand years in China. Recently it has been gradually accepted in western countries as an alternative or complementary treatment. However, how the acupuncture works remains an open question though acupuncture exists as one of the oldest continuous systems of medicine dating back 4,000 years. Extensive studies have been conducted on the mechanism of acupuncture to explain the effects of acupuncture on various systems and symptoms [1-3]. Compared to the relatively widespread use of acupuncture, systems biology is a new term to describe the recent trends in biology research. It emphasizes the high-throughput measurement of biological systems and focuses on the complex interactions in biological systems [4,5]. We highly expect that systems biology, a biology-based inter-disciplinary study field, will provide tremendous opportunities for revealing acupuncture mechanism at the molecular level.

In this paper, we use systems biology method to study the acupuncture treatment effect by identifying a subset of important molecules from high-throughput metabolic data. Specifically, we separate the acupuncture from moxibustion and only study the effect of acupuncture on normal people by investigating the difference between acupuncture at particular acupoint and without acupuncture. Towards this aim, we utilize ^{1}H nuclear magnetic resonance (^{1}H NMR) to investigate the effects of acupuncture at several meridian points on plasma metabolites. Then metabolite profiles (vectors) are generated from a collection of case samples(with acupuncture at meridian point) and control samples (without acupuncture). These high-dimensional profile data is very similar to SNP (sequence data), gene expression (transcriptome), mass spectrum (proteome), and small molecules (metabolome) data in different levels. Then the straightforward task is to identify differentially expressed molecules and further classify and predict the diagonostic category of a sample, based on its metabolite profile [6].

Generally speaking, there are two difficulties in analyzing these high-dimensional profile data. First, a large number of features (metabolites in our case) are available to predict classes for a relatively small number of samples. The presence of a significant number of irrelevant features that are unrelated to the case status makes such analysis prone to the curse of dimensionality. Second, predictive accuracy is not the only goal and further biological validation and mechanism understanding call for explanatory power other than black box predictive results. Thus it is especially important to know which molecules largely contribute towards the classification. Ideally we can improve the generalization performance of classifier by identifying only the molecules that are significantly contribute to the classifier. This effect is attributable to the overcoming of the curse of dimensionality. For example, if it is possible to identify a small set of metabolites that is indeed capable of providing complete discriminatory information, inexpensive diagnostic assays for only a few metabolites might be developed and be widely deployed in clinical settings. Knowledge of a small set of diagnostically relevant metabolites may provide important insights into the mechanisms responsible for acupuncture treatment itself. Those molecules are usually termed as biomarkers. The procedure to reveal them is referred as feature selection, biomarker identification, or feature ranking.

Feature selection is known to be NP-hard [7] and the search becomes quickly computationally intractable. Suppose we treat the feature selection task in a brute force way. Given n features, we need to select m features which can get the best classification accuracy ($m << n$) regarding to a predefined cost function. Usually in classification or prediction problem, the cost function is selected as the accuracy of the prediction. The exhaustive search method goes through all the possible combinations, with the computation complexity $O(n^m)$. Thus, this method is not practical for realistic applications.

Existing feature selection strategies can be roughly categorized into three types [8]. Considering the partial ordering properties of the subset space, we can either start with an empty set and successively add features, or start with the set of all features and successively filter them. The former type is referred to as forward selection while the latter is referred to as backward elimination. The third type is the combination of the two approaches. However, all the above methods relies on the greedy strategy. As an example of forward feature selection, we might first look for the single most discriminative feature using any classifier. Then we could search the single additional feature that gives the best class discrimination when considered along with the first feature. Keeping augmenting the feature set iteratively in this greedy fashion, we stop until cross-validation error estimates are minimized. As a result, the global optimal solution usually cannot be guaranteed.

In this paper, we proposed a novel linear programming (LP) model to address this important problem. Feature selection problem is cast into an optimization problem with two objectives, one is to minimize the number of chosen features and the other is to maximize the predictive accuracy based on the centroid classification framework. In other words, our feature selection method simultaneously improves classification accuracy

and selects features. Comparing with several state-of-the-art feature selection methods, our Linear Programming based Feature Selection (LPFS) method can select a small set of features by applying strong regularization while keeping high accuracy. We then apply our method to analyze the metabolite profile data generated for acupuncture treatment. We identify important molecules (biomarkers) related to the acupuncture treatment for several meridian points. Further characterization of the biomarkers and the common and difference among several meridian points provide biological insights for acupuncture mechanisms at molecular level. Preliminary results in this paper were presented in our conference paper [9]. In this extended paper, we further provide the detailed derivation of the method and comparison with existing works. In addition, we performed more analysis and descriptions for the biological insights of the acupuncture biomarkers.

Method

Analytic workflow

In this paper, the acupuncture treatment effect is investigated in the framework of systems biology. The basic analytic workflow is shown in Figure 1. As the first step, metabolite profiles are originally generated by ^{1}H NMR from control samples and the samples with acupuncture treatment in meridian points. Then we develop a linear programming based feature selection method to compare the two groups of metabolite profiles. Finally, a small set of metabolites are selected as biomarkers for acupuncture treatment effect.

Overview of the linear programming based feature selection

To investigate the high-dimensional data for acupuncture treatment effect, we develop a novel method, LPFS, to select a small set of metabolites to characterize acupuncture treatment effect. The schematic illustration of LPFS is shown in Figure 2. LPFS performs feature selection based on the nearest centroid classifier. On one hand, we want to attain the best classification accuracy by minimizing the loss function. On the other hand, feature selection algorithms should be robust to noise and outliers in the data by applying strong regularization. Here we use the parsimony principle (also known as Occam's razor) by minimizing the number of selected features. Then the feature selection problem is formulated as a multi-objective programming. The next step is to convert this multi-objective programming into a single-objective linear programming by applying the ε method. After solving the linear programming model in an efficient way, the optimal features can be selected.

Centroid classification prototype

A fast and simple algorithm for classification is the centroid method [6,10]. This algorithm assumes that the target classes correspond to individual (single) clusters and uses the cluster means (or centroids) to determine the class of a new sample point (see Figure 2). A prototype pattern for class C_j is defined as the arithmetic mean:

$$\mu_{C_j} = \frac{1}{|C_j|} \sum_{s_i \in C_j} x_i \tag{1}$$

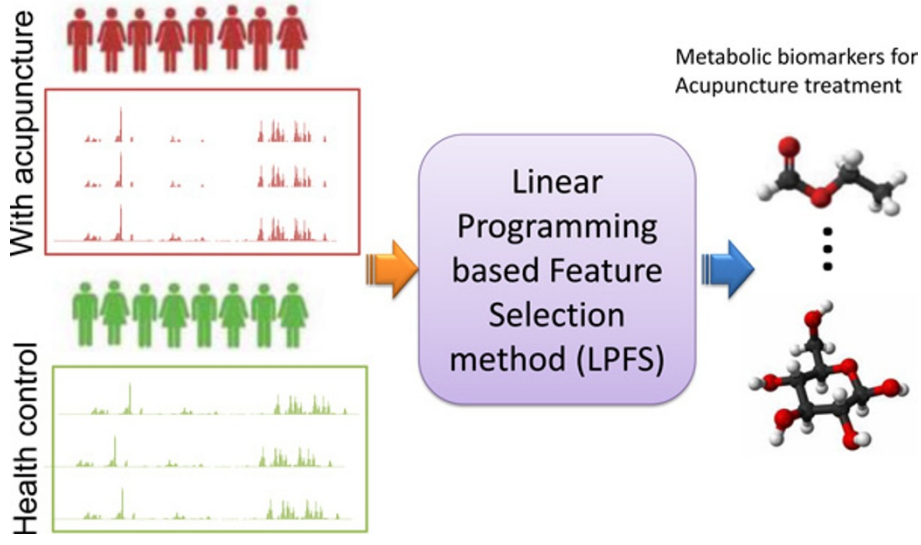

Figure 1 Our analytic workflow to identify biomarkers for acupuncture treatment effect. Metabolite profiles are originally generated by ^{1}H NMR from control samples and the samples with acupuncture treatment. Then we develop a linear programming based feature selection method to compare the two groups of metabolite profiles. Finally a small set of metabolites are selected as biomarkers for acupuncture effect.

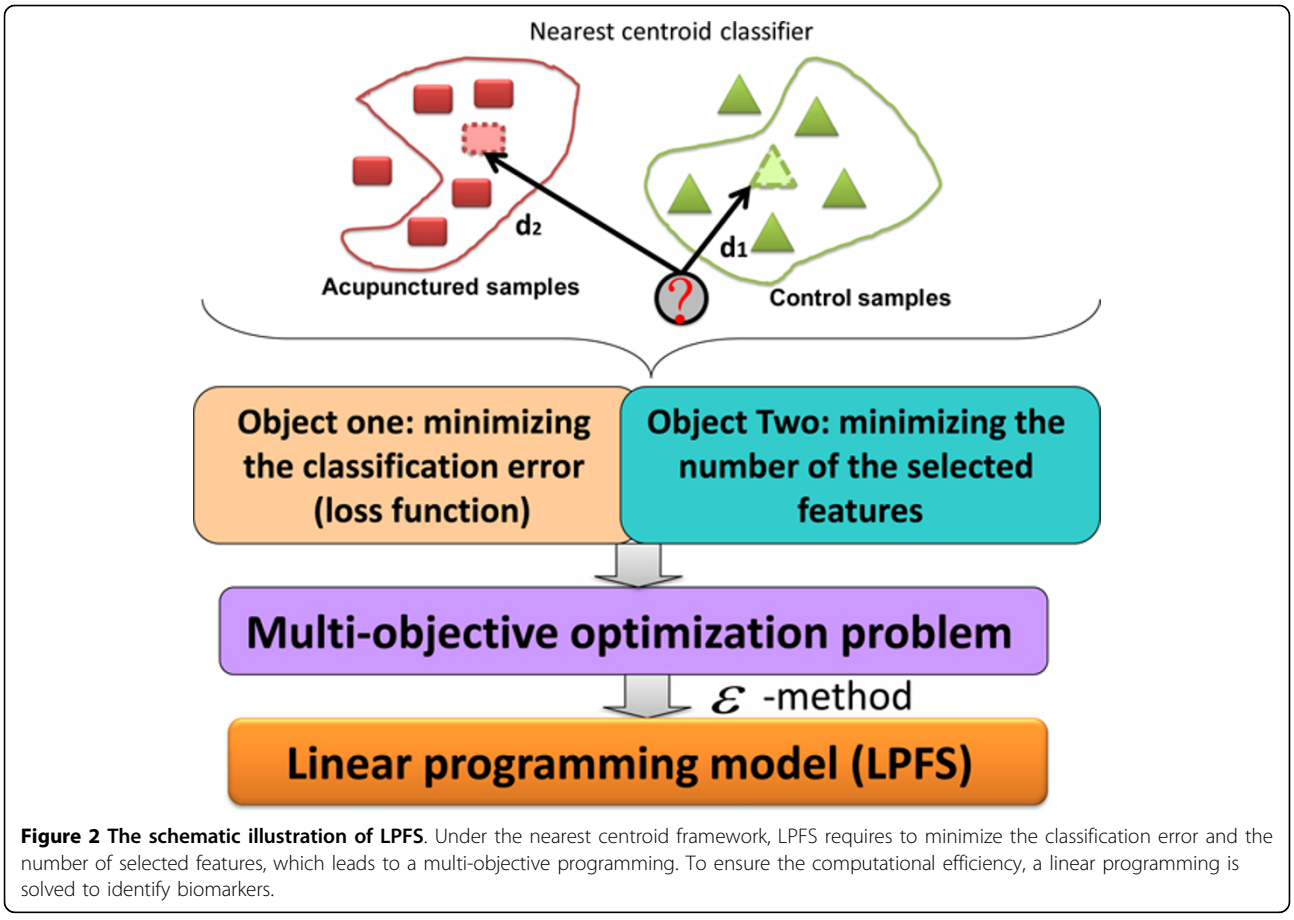

Figure 2 The schematic illustration of LPFS. Under the nearest centroid framework, LPFS requires to minimize the classification error and the number of selected features, which leads to a multi-objective programming. To ensure the computational efficiency, a linear programming is solved to identify biomarkers.

where $\boldsymbol{s}_i$ is the i-th training sample labeled as class C_j. Recall that the training sample is a metabolite spectra represented as a multi-dimensional vector (denoted in bold). In a similar fashion, we can obtain a prototypical vector for all the other classes. During classification, the class label of an unknown sample $\boldsymbol{s}$ is determined as:

$$C(s) = \arg\min_{C_j} dis(\mu_{C_j}, s) \tag{2}$$

where $dis(\boldsymbol{x}, \boldsymbol{y})$ is a distance function or:

$$C(s) = \arg\max_{C_j} sim(\mu_{C_j}, s) \tag{3}$$

where $sim(\boldsymbol{x}, \boldsymbol{y})$ is a similarity metric. This simple classifier will form the basis of our LPFS method. The advantages is that it works with any number of features. And its run-time complexity is proportional to the number of features and the complexity of the distance or similarity metric used. According to the experiments in [11], we select L_1 distance metric, which is robust to outliers and most appropriate for the centroid classification algorithm. It is defined by:

$$L_1(s, \mu) = \|s - \mu\|_1 \tag{4}$$

with $\|y\|_1 = \Sigma_i |y(i)|$, and $y(i)$ being the i-th element of vector $\boldsymbol{y}$. The value $L_1(\boldsymbol{s}, \boldsymbol{\mu})$ has a linear cost in the number of features. In this study, data sets contain two classes and hence the number of calls to the distance metric is also two. Therefore, the centroid classifier, at run-time, is linear in the number of features. During training, two prototypes are computed and the cost of computing each prototype is $O(mN)$, where N is the number of features and m is the number of training samples which belong to a given class. Note that m only varies between data sets and not during training or feature selection processes. Thus, we can view m as a constant and the centroid classifier has $O(N)$ cost in the training phase.

Multi-objective optimization model

Suppose we have two groups in the training dataset, the case group and the control group as the gold-standard data to classify new samples. We denote them set T and F respectively. Supposing $|T| = m_1$, $|F| = m_2$, and the computed centroids are $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_F$ respectively. A simple classification scheme is as follows. Given a new sample

$\boldsymbol{s}$, we want to decide which group it belongs to. The L_1 discrepancy between the sample $\boldsymbol{s}$ and the groups T and F can be calculated as $\|\boldsymbol{s} - \boldsymbol{\mu}_T\|_1$ and $\|\boldsymbol{s} - \boldsymbol{\mu}_T\|_1$. Thus a simple rule is

$$s \in T \quad \text{if} \quad \|s - \mu_T\|_1 < \|s - \mu_F\|_1 \tag{5}$$

$$s \in F \quad \text{if} \quad \|s - \mu_T\|_1 > \|s - \mu_F\|_1 \tag{6}$$

Let the feature number be n. We introduce the variables for feature selection as $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, where $x_i = 0,1$. When $x_i = 1$, it means feature i is selected in the biomarker set. Otherwise it is not selected.

Suppose the test dataset is U. And it is composed by the case group U_T and control group U_F. $U = U_T \cup U_F$, and $|U_T| = l_1$, $|U_F| = l_2$. Given a case sample $\boldsymbol{s_l} = (s_{l1}, s_{l2}, ..., s_{ln})$, $l \in \{1, 2 ..., l_1\}$, if it is classified correctly, we should have

$$\sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i < \sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \tag{7}$$

Where $\boldsymbol{t_k} = (t_{k1}, t_{k2}, ..., t_{kn}) \in T$, $k = 1, 2, ..., m_1$ and $\boldsymbol{f_k} = (f_{k1}, f_{k2}, ..., f_{kn}) \in F$, $k = 1, 2, ..., m_2$.

Similarly given a control sample $\boldsymbol{s_l} = (s_{l1}, s_{l2}, ..., s_{ln})$, $l \in \{l_1 + 1, l_1 + 2 ..., l_1 + l_2\}$, if it is classified correctly, we should have

$$\sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i > \sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \tag{8}$$

With the above constraints for variable $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, the objective function is to choose as few as features, i. e.,

$$\min_{x_1, x_2, \cdots, x_n} \sum_{i=1}^{n} x_i \tag{9}$$

Thus the feature selection problem is formulated as an integer linear programming problem in Equation (10).

$$\begin{aligned} \min_{x_1, x_2, \cdots, x_n} \quad & \textstyle\sum_{i=1}^{n} x_i \\ s.t. \quad & \textstyle\sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i < \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i, \\ & p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in U_T, l \in \{1, 2, \cdots, l_1\}, \\ & \textstyle\sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i > \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\ & p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in U_F, l \in \{1, 2, \cdots, l_2\}, \\ & x_i = 0, 1, \quad i \in \{1, 2, \cdots, n\} \end{aligned} \tag{10}$$

When we consider the noise in the measured data, not all the test samples can be classified exactly. We introduce the tolerable error $\boldsymbol{y} = \{y_1, y_2, ..., y_{l1+l2}\}$ for every sample in $U_T \cup U_F$. And $y_i \geq 0$, $i \in \{1,2, ..., l_1 + l_2\}$. When y_i is not equal to zero, it means sample i is wrongly classified. Otherwise this sample should be correctly classified.

Given a case sample $\boldsymbol{s_l} = (s_{l1}, s_{l2}, ..., s_{ln})$, $l \in \{1, 2 ..., l_1\}$, we should have the following constraint considering the tolerable error

$$\sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i - \gamma_l < \sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \tag{11}$$

Similarly given a control sample $\boldsymbol{s_l} = (s_{l1}, s_{l2}, ..., s_{ln})$, $l \in \{l_1 + 1, l_1 + 2 ..., l_1 + l_2\}$, we should have the following constraint considering the tolerable error

$$\sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i + \gamma_l > \sum_{i=1}^{n} \left| s_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \tag{12}$$

Thus the objective function composes two parts, i. e., we want to choose as few as features $\min_{x_1, x_2, \cdots, x_n} \sum_{i=1}^{n} x_i$ and at the same time we want to reduce the classification error (loss function) $\min_{\gamma_1, \gamma_2, \cdots, \gamma_{l_1+l_2}} \sum_{i=1}^{l_1+l_2} \gamma_i$. In general, there is a trade-off relationship between the classification error and the number of features. Hence, the feature selection problem can be formulated as a multi-objective optimization problem with discrete variables $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ and continuous variables $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_{l_1+l_2})$ as shown in Equation (13).

$$\begin{aligned} \text{vector - minimize}_{(x,\gamma)} \quad & \left\{ \textstyle\sum_{i=1}^{n} x_i, \sum_{i=1}^{l_1+l_2} \gamma_i \right\}, \\ \text{subject to} \quad & (11)(12) \quad \text{with } x_i \in \{0, 1\}, i \in \{1, 2, \cdots, n\}, \\ & \gamma_i \geq 0, i \in \{1, 2, \cdots, l_1 + l_2\} \end{aligned} \tag{13}$$

The first term of objective function in Equation (13) is to minimize the number of chosen features, and the second one is to minimize the total classification error.

Mixed integer linear programming

The optimal solutions of the two-objective optimization problem consist of a Pareto set, which can be solved by transforming the two objectives of (13) into a single objective. One typical technique is the ϵ-method, which alternates a positive scalar parameter λ to obtain the Pareto set, with the formulation in Equation (14).

$$\begin{aligned} \min_{x,\gamma} \quad & \textstyle\sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{l_1+l_2} \gamma_i \\ s.t. \quad & \textstyle\sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i - \gamma_l < \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\ & p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in U_T, l \in \{1, 2, \cdots, l_1\}, \\ & \textstyle\sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i + \gamma_l > \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\ & p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in U_F, l \in \{1, 2, \cdots, l_2\}, \\ & x_i = 0, 1, \quad i \in \{1, 2, \cdots, n\}, \gamma_j \geq 0, \quad j \in \{1, 2, \cdots, l_1 + l_2\} \end{aligned} \tag{14}$$

(14) is a mixed integer programming (ILP). The objective function in (14) is $\sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{l_1+l_2} y_i$. Theoretically, we can obtain all optimal solutions belonging to the Pareto set by changing the parameter λ for the single-objective optimization problem (14). Clearly, λ transforms the number of chosen features into equivalent classification error in (14), and controls the balance between them.

By solving the proposed mixed integer linear programming model (14), we can get solutions for the feature selection variables x_i, $i \in \{1, 2, \cdots, n\}$, and classification error variables y_j, $j \in \{1, 2, ..., l_1 + l_2\}$. Checking if x_i is equal to 1, we can know if the corresponding feature should be selected in the classifier. Meanwhile checking the values of all the y_j, we can estimate the classification accuracy. For example, suppose the number of all j such that $y_j = 0$ is N_1 and the number of all j such that $y_j > 0$ is N_2. We can simply calculate the classification accuracy by N_1/l_1+l_2 and N_2/l_1+l_2.

Leave-one-out cross validation

The above model (14) is based on the general idea of cross validation, thus it depends on the choice of T and F. We noticed that there are different ways to do cross validation in feature selection [12]. One way is that the feature selection is done with all the samples and the cross-validation is only done for the classification procedure. This kind of cross validation may severely bias the evaluation in favor of the studied method due to "information leak" in the feature selection step. Another way is to include the feature selection procedure in the cross validation, i.e., to leave the test sample(s) out from the training set before undergoing any feature selection. Our feature selection model allows the freedom to choose the suitable cross validation procedure according to the practical need. If more information is preferred to select biomarkers due to the scarcity of samples, we can use all the samples to estimate the cross validation error in the following resubstitution validation in Equation (15).

$$\begin{aligned}
\min_{x,y} \quad & \sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{l_1+l_2} y_i \\
s.t. \quad & \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i - y_l < \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\
& p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in T, l \in \{1, 2, \cdots, m_1\}, \\
& \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i + y_l > \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\
& p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in F, l \in \{1, 2, \cdots, l_2\}, \\
& x_i = 0, 1, \quad i \in \{1, 2, \cdots, n\}, y_j \geq 0, \quad j \in \{1, 2, \cdots, l_1 + l_2\}
\end{aligned} \tag{15}$$

Resubstitution error rate indicates only how good are our biomarkers on the training data. However, this model has "information leak" and will underestimate the classification error. In our implement, we choose the model for leave-one-out cross validation in Equation (16).

$$\begin{aligned}
\min_{x,y} \quad & \sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{l_1+l_2} y_i \\
s.t. \quad & \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1-1} t_{ji}/(m_1-1) \right| x_i - y_l < \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\
& p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in T, l \in \{1, 2, \cdots, l_1\}, t_k = (t_{k1}, t_{k2}, \cdots, t_{kn}) \in T \backslash \{p_l\}, k \in \{1, 2, \cdots, m_1\} \backslash \{l\} \\
& \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i + y_l > \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2-1} f_{ji}/(m_2-1) \right| x_i \\
& p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in F, l \in \{1, 2, \cdots, l_2\}, f_k = (f_{k1}, f_{k2}, \cdots, f_{kn}) \in F \backslash \{p_l\}, k \in \{1, 2, \cdots, m_2\} \backslash \{l\} \\
& x_i = 0, 1, \quad i \in \{1, 2, \cdots, n\}, y_j \geq 0, \quad j \in \{1, 2, \cdots, l_1 + l_2\}
\end{aligned} \tag{16}$$

We adopt leave-one-out experiment since this particular form of cross validation is an unbiased estimator of the generalization performance of classifier. It makes the best use of the available data and involves no random subsampling. Every time we pick out one sample ($l_1 = 1$ or $l_2 = 1$) from the training data and try to classify it correctly. And by doing $m_1 + m_2$ times test we add $m_1 + m_2$ constraints.

Linear programming approximation

In general, mixed integer linear programming is difficult to solve. To ensure the computational efficiency, mixed ILP in Equation (16) can be relaxed to the corresponding linear programming (LP). Linear programming is the simplest type of mathematical programming and has been widely used in systems biology study [5,13,14]. Therefore, sophisticated algorithm can be adopted to efficiently solve this LP. In terms of computational complexity, the proposed approach makes the computation of biomarker tractable. Finally we construct the LPFS model in Equation (17).

$$\begin{aligned}
\min_{x,y} \quad & \sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{l_1+l_2} y_i \\
s.t. \quad & \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1-1} t_{ji}/(m_1-1) \right| x_i - y_l < \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2} f_{ji}/m_2 \right| x_i \\
& p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in T, l \in \{1, 2, \cdots, l_1\}, t_k = (t_{k1}, t_{k2}, \cdots, t_{kn}) \in T \backslash \{p_l\}, k \in \{1, 2, \cdots, m_1\} \backslash \{l\} \\
& \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_1} t_{ji}/m_1 \right| x_i + y_l > \sum_{i=1}^{n} \left| p_{li} - \sum_{j=1}^{m_2-1} f_{ji}/(m_2-1) \right| x_i \\
& p_l = (p_{l1}, s_{l2}, \cdots, p_{ln}) \in F, l \in \{1, 2, \cdots, l_2\}, f_k = (f_{k1}, f_{k2}, \cdots, f_{kn}) \in F \backslash \{p_l\}, k \in \{1, 2, \cdots, m_2\} \backslash \{l\} \\
& x_i \geq 0, \quad i \in \{1, 2, \cdots, n\}, y_j \geq 0, \quad j \in \{1, 2, \cdots, l_1 + l_2\}
\end{aligned} \tag{17}$$

After relaxing to continuous value, the value of the optimal solution of x_i (LPFS score) indicates the importance of feature i in the nearest centroid classifier. It should be noted that we can use other distance definitions instead of L_1 in our model to achieve the nonlinear classification effect. The parameter λ can be determined by checking the output leave-one-out predictive accuracy. We also notice that LPFS model can be extended to multi-classification task and n-fold cross validation.

Metabonomics profiling by ^{1}H NMR spectra

Venous blood (3ml) was collected into a heparin sodium tube and the plasma was collected by centrifugation at 1000× g at 4°C for 10 minutes. An aliquot of 300 μl plasma was mixed with 250 μl D_2O and 50 μl TSP (3-trimethylsilyl-$^2H_4$-propionic acid) in D_2O (1 mg/ml) in 5 mm NMR tube. The D_2O provided a field-frequency lock solvent for the NMR spectrometer and the TSP served as an internal reference of chemical shift. ^{1}H NMR spectra of the plasma samples were acquired on a

Varian INOVA 600 MHz NMR spectrometer at 27°C by using Carr-Purcell-Meiboom-Gill (CPMG) spin-echo pulse sequence. with a total spin-spin relaxation delay ($2n\tau$) of 320 ms. The free induction decays (FIDs) were collected into 32K data points with a spectral width of 8000 Hz and 64 scans. The FIDs were zero-filled to double size and multiplied by an exponential line-broadening factor of 0.5 Hz prior to Fourier transformation (FT). In addition, diffusion-edited experiments were also carried out with BPP-LED (bipolar pulse pair longitudinal eddy current delay) pulse sequence [15,16]. The gradient amplitude was set at 35.0 G cm^{-1}, with a diffusion delay of 100 ms. A total of 128 transients and 16k data points were collected with a spectral width of 8000 Hz. A line-broadening factor of 1 Hz was applied to FIDs before Fourier transformation.

All plasma ^{1}H NMR spectra were manually phased and baseline corrected using VNMR 6.1C software (Varian, Inc.). For CPMG spectra, each spectrum over the range of 0.4 to 4.4 was data-reduced into integrated regions of equal width (0.01 ppm). For BPP-LED data, each spectrum over the range of 0.1 to 6.0 was segmented into regions of equal width (0.01 ppm). The regions containing the resonance from residual water (4.6-5.1) were excluded. The integral values of each spectrum was normalized to constant sum of all integrals in a spectrum in order to reduce any significant concentration differences between samples [17,18]. Identification of metabolites in spectra was accomplished based on literatures and the Chenomx NMR Suite 5.0 (Chenomx, Calgary, Canada).

Results

Metabonomics data generation

To investigate the acupuncture treatment effects, we originally obtained metabonomics data of plasma metabolites in healthy males at five meridian points using Proton NMR. Proton NMR (also named as Hydrogen-1 NMR, or ^{1}H NMR) applies nuclear magnetic resonance in NMR spectroscopy with respect to hydrogen-1 nuclei within the molecules of a substance, in order to determine the structure of the molecules [19].

As a result, most organic compounds are characterized by chemical shift values, which are usually expressed in parts per million (ppm) by frequency and are in the range +14 to -4 ppm. Chemical shift values are not precise, but typically they are regarded mainly as orientational. The exact value of chemical shift depends on molecular structure and the solvent in which the spectrum is being recorded. These chemical shift values can be mapped to eight metabolic subsets (amino acids, carbohydrates, energy, glycans, lipids, nucleotides, secondary metabolites/xenobiotics, vitamins, and cofactors). In our experiment, 400 chemical shift values are measured for their concentration in plasma, and mathematically every sample is represented by a vector in 400 dimensional space.

Fifty healthy young males were randomly allocated to Zusanli (ST36), Liangmen (ST21), Juliao (ST3), Yanglingquan (GB34), and Weizhong (BL40) groups (The locations of the meridian points are shown in Figure 3a. Among the five points, Zusanli, Liangmen, and Juliao are on the same meridian.). Each group contains 10 persons. Inside each group the corresponding meridian points were separately acupunctured for 5 consecutive days. In addition, twenty healthy young males are recruited as the blank control groups. All the twenty people are measured before the start of 5 consecutive days and additionally ten of them are measured after 5 consecutive days. Fasting venous blood was taken in all the subjects. Plasma metabolites were measured by ^{1}H NMR to derive metabolic profiles (see details in Method Section). Furthermore to exclude possible noises, all the seventy males are strictly trained to make sure their metabolic profiles are measured in very similar conditions. The detailed experimental method can be found in [20]. In summary, we have 80 samples grouped into Zusanli (10 samples, acupuncture point ST36), Liangmen (10 samples, acupuncture point ST21), Juliao (10 samples, acupuncture point ST3), Yanglinquan(10 samples, acupuncture point GB34), Weizhong (10 samples, acupuncture point BL40), Control I (10 samples, normal people measured after the consecutive 5 days), and Control II (20 samples, normal people measured before the consecutive 5 days).

Classification experiments design

With the data, we design experiments to identify biomarkers for the acupuncture treatment of each meridian point. The overall design of biomarker identification experiments is shown in Figure 3b. We categorize eighty samples into 8 groups shown as the circles in Figure 3b. ST36, ST21, ST3, GB34, and BL40 each has 10 samples. The 20 samples in Control II are naturally decomposed into two groups with equal size, Pre1 (10 samples with follow-up measurement after 5 days) and Pre2 (10 samples without follow-up measurement). Treating the Pre1 as the common control set, we have seven classification tasks (Exp1 to Exp7) shown as the lines in Figure 3b. For example, task Exp1 tries to identify a subset of metabolites to classify Pre1 as the control and ST36 as the case. In this way, Exp1 to Exp5 aim to identify the biomarkers for acupuncture treatment on ST36, ST21, ST3, GB34, and BL40 respectively. While Exp6 tries to capture the metabolite change by 5 consecutive days without acupuncture. And Exp7 tries to test if there are significant metabolite change for the people under similar condition. Exp6 and Exp7 serve as the control studies to guarantee the significance of our result.

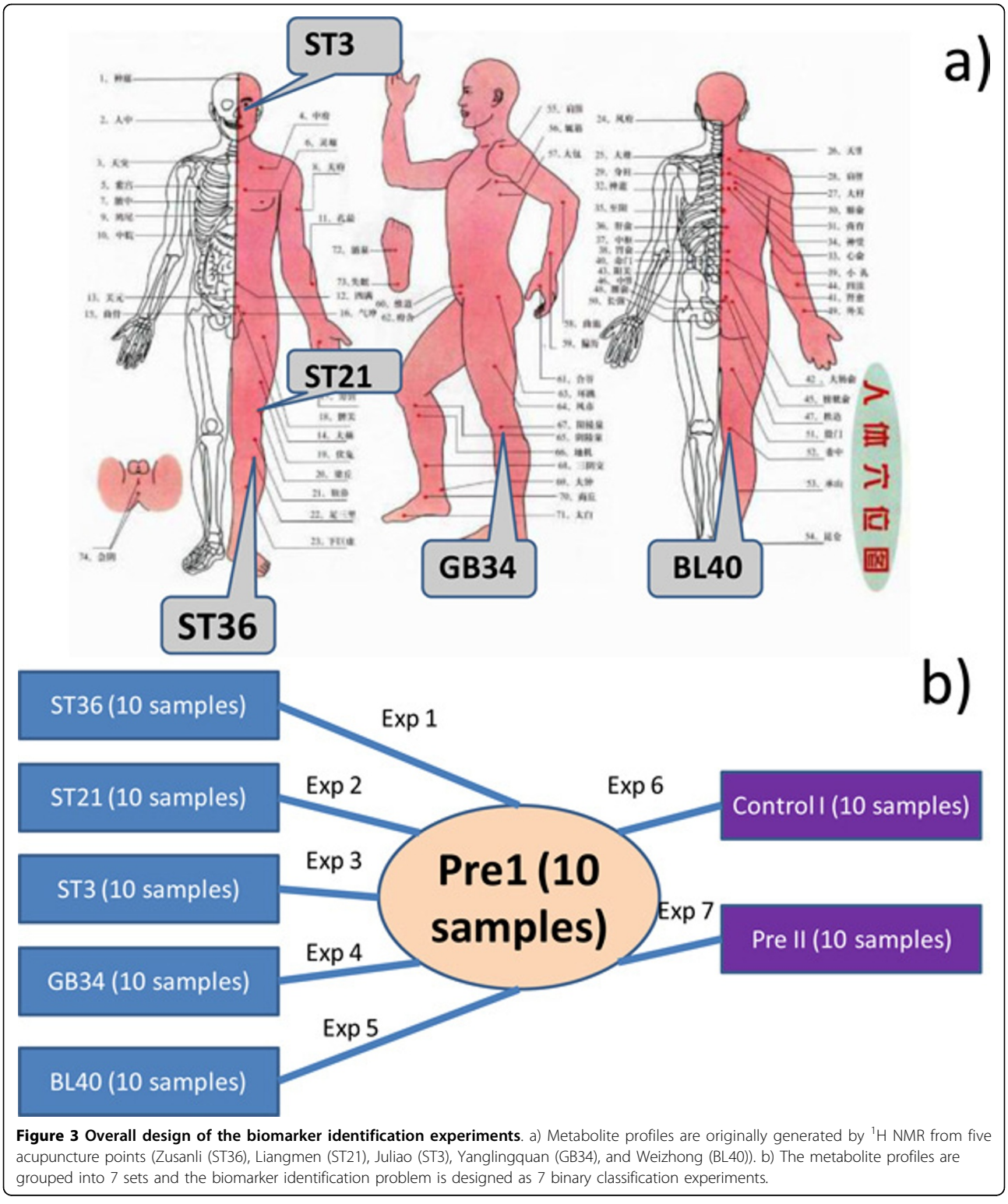

Figure 3 Overall design of the biomarker identification experiments. a) Metabolite profiles are originally generated by ^{1}H NMR from five acupuncture points (Zusanli (ST36), Liangmen (ST21), Juliao (ST3), Yanglingquan (GB34), and Weizhong (BL40)). b) The metabolite profiles are grouped into 7 sets and the biomarker identification problem is designed as 7 binary classification experiments.

Global characterization of the data

We first perform hierarchical clustering on the 80 metabolic profiles. The results are shown in Figure 4a. If the samples can be clearly discriminated by global pattern, the 80 samples should be clustered by with or without acupuncture treatment and then by their meridian points. However, all the sample labels are mixed in the clustering result (Figure 4a) and we cannot see clearly boundaries.

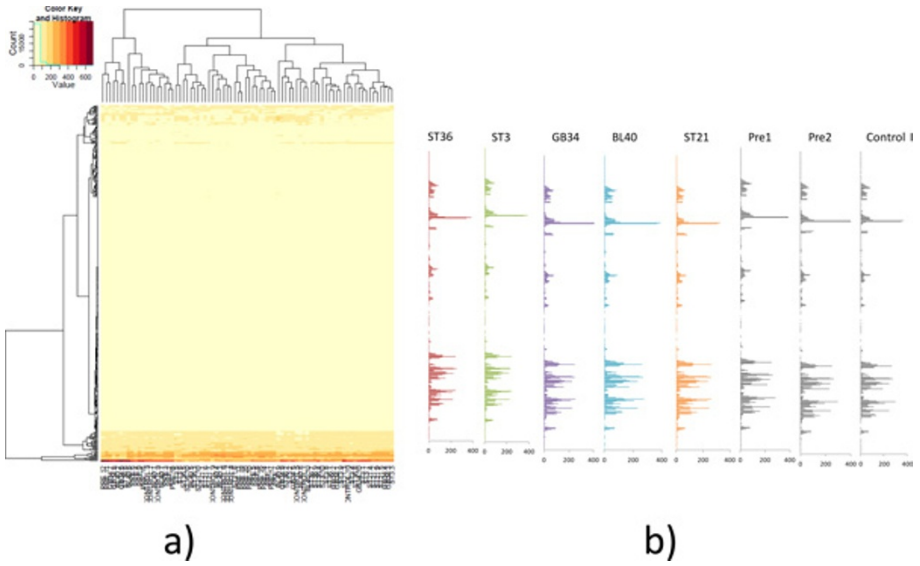

Figure 4 Global characterization of the metabolite profiles. a) Hierarchical clustering of the metabolic profiles of the 80 samples. b) Centroids for the seven datasets. The horizontal units are expression values for the metabolites. The metabolites are sorted by their chemical shift values.

Furthermore, we calculate the centroids for the seven groups of samples in Figure 4b by averaging the 10 samples for their metabolite expression values. These centroids are plotted side by side in Figure 4b, which shows that these centroids are very similar and it's very difficult to detect the global difference.

The above results together demonstrate that global pattern in metabolic profiles cannot discriminate the Zusanli, Yanglingquan, Liangmen, Juliao, Weizhong, Pre1, Pre2, and Control I groups. Thus it is necessary to find the local pattern in the profile data. Our strategy is to find a subset of metabolites as biomarkers to achieve clear discrimination.

Comparison with other approaches

Before we conduct the acupuncture biomarker identifications, we benchmark our LPFS method by comparing with several existing state-of-the-art methods. There are many existing feature selection methods and they can be roughly categorized into three types, filter, wrapper, and embedded methods. To make the comparison simple and comprehensive, we pick out some representative methods in each type to compare in the same dataset.

Filter methods select features as a preprocessing step and feature selection part is independent of a machine learning algorithm (classifier). This is computationally efficient. Fold change and t-test are the simplest and popular methods to identify biomarker. They are usually the representative methods for filter methods.

Let x_{ij} and y_{ij} denote the log expression values of metabolite i in sample j in the case and control, respectively. We define the ordinary two-sample t-statistic [21] as

$$T_i \frac{\bar{x}_i - \bar{y}_i}{s_i} \tag{18}$$

Where $\bar{x}_i, \bar{y}_i$, and s_i are the mean of case, mean of control, and the standard deviation of the samples for metabolite i. From t-statistic T_i we can easily calculate a p-value. Usually a feature is selected if its corresponding p-value is smaller than a predefined threshold 0.05.

The standard definition of the fold-change [21] for metabolite i is

$$FC_i = \frac{\overline{x}_i}{\overline{y}_i} \tag{19}$$

Where $\acute{x}_{ij}$ and $\acute{y}_{ij}$ are the raw expression values of metabolite i in sample j in the case and control, respectively. In our implementation, we computes the difference of means (i.e. the fold-change for log-transformed data) and then rank the metabolites by their absolute values. We choose a cutoff 2 to select the significant ones.

On the other hand, wrapper method ranks features based on their effects on classification accuracy. It takes dependencies of the feature subset on the learning algorithm into account and is computationally more demanding. Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is one of the most successful wrapper method based algorithm in the feature selection [22]. SVM-RFE conducts feature selection in a

sequential backward elimination manner, which starts with all the features and discards one feature at a time by checking the SVM accuracy. It has been widely used and extended in high-dimensional data analysis [23]. We compare our LPFS method with SVM-RFE in metabolite data.

Since our LPFS method is an embedded method and simultaneously optimize classification accuracy and the number of selected features, we specifically choose to compare with an existing method with similar strategy, called sparse multinomial logistic regression approach (SMLR). It was developed to jointly and simultaneously identify the optimal nonlinear classifier, and select the optimal set of features via the optimization of a single posterior objective function (see [24] and [25]). SMLR has been extensively applied in problems in systems biology [26]. SMLR is freely available at http://www.cs.duke.edu/~amink/software/smlr/ and we take the default values for the parameters in our calculation.

Without loss of generality, we take Exp1 in Figure 3b as an example to identify a subset of metabolites to discriminate ST36 and Pre1. We select biomarkers from the metabolic profiles and compare the results in two ways.

Firstly, we compare different methods in Figure 5, by assessing the quality of the selected biomarkers. We simply plot all the metabolites by their standard derivation versus the difference of mean expression value. This is a scatter-plot used in [27], which plots significance versus fold-change on the y- and x-axes, respectively. Conceptually, it is very similar to the volcano plot in statistics [28]. Plotting points in this way results in two regions of interest in the plot: those points that are found towards the bottom of the plot and far to either the left- or the right-hand side. These represent values that display large magnitude fold changes as well as high statistical significance with small standard derivation.

The t-test based method identifies 172 metabolites if we choose a cutoff 1.73 (corresponding p-value 0.05). A strict threshold will still select 84 metabolites with cutoff 2.84 (corresponding p-value 0.005). We list the top 10 in Table 1 and plot them in Figure 5a. We find that the ordinary t-statistic selects metabolites with low standard deviations.

The fold change based method identifies 159 metabolites if we choose a commonly used cutoff 2 [28]. And there are still 97 metabolites by choosing a cutoff 4. Again the top 10 are listed in Table 2 and plotted in Figure 5b. It's clear that the fold-changes select metabolites with large shifts between control and treatment.

While SMLR selects 37 features to achieve the 100% leave-one-out predictive accuracy. These 37 metabolites

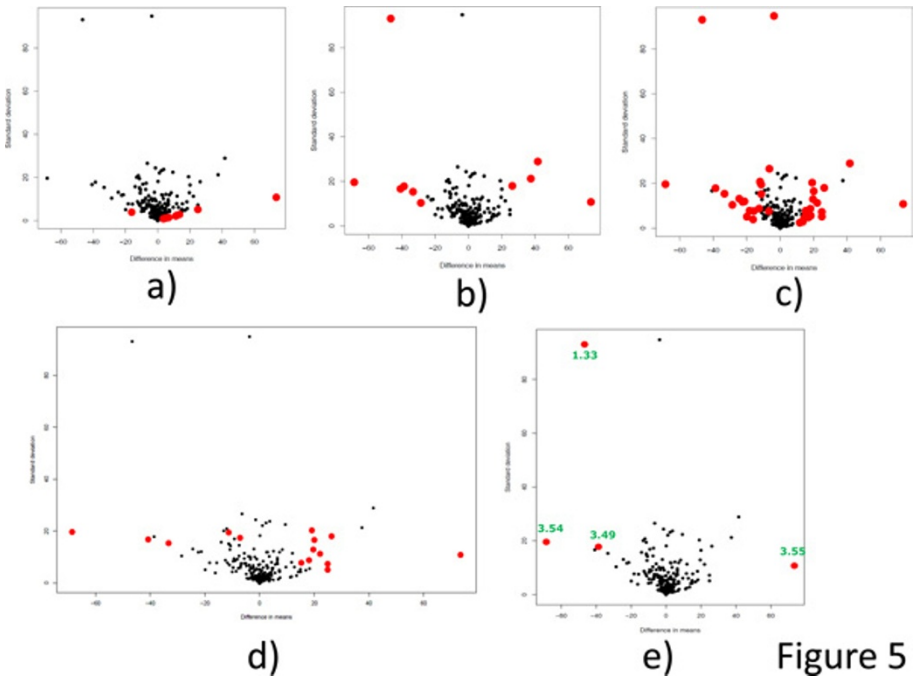

Figure 5 Methods comparison via volcano like plots. Comparison of our LPFS method with existing methods regarding to the identified biomarkers. All the 400 metabolites are plotted into a two dimensional plane. The selected biomarkers are highlighted in red. The x-axis denotes the difference of means and the y-axis denotes the standard derivation. Good biomakers should locate either in the left bottom corner or in the right bottom corner. a) volcano like plot of t-test method. The top 10 features are in red. b) volcano like plot of fold change method. The top 10 features are in red. c) volcano like plot of SMLR method. d) volcano like plot of SVM-RFE method. e) volcano like plot of our LPFS method.

Table 1 The top ten identified biomarkers by different methods on the ST36 meridian point.

Student t-test			Fold change			SMLR		SVM-RFE		Our LPFS method			
ID	ppm	t-score	ID	ppm	FC-score	ID	ppm	ID	ppm	ID	ppm	LPFS score	Metabolite name
86	3.55	15.29	86	3.55	73.48	45	4	86	3.55	86	3.55	0.015	
195	2.46	11.91	87	3.54	68.52	50	3.95	68	3.73	92	3.49	0.008	
251	1.9	11.07	308	1.33	46.58	52	3.93	87	3.54	87	3.54	0.006	a-glucose/glycine
45	3.96	10.96	310	1.31	41.61	58	3.87	69	3.72	308	1.33	0.002	lactate
229	2.12	10.86	70	3.71	40.75	60	3.85	102	3.39				
81	3.6	10.80	92	3.49	38.61	67	3.78	70	3.71				
18	4.23	10.03	293	1.48	37.45	68	3.77	295	1.46				
127	3.14	9.75	295	1.46	33.26	69	3.76	116	3.25				
17	4.24	9.71	71	3.7	28.55	70	3.75	229	2.12				
232	2.09	9.35	69	3.72	26.30	71	3.74	88	3.53				

are plotted in Figure 5c and 10 of them (ranked by ID) are listed in Table 1. Figure 5c demonstrates that SMLR is quite efficient to cover almost all the metabolites with large fold change and small standard derivation. However, 37 metabolites is too much from the viewpoint of practical usage of biomarkers.

SVM-RFE selects 15 metabolites in total to achieve the 100% leave-one-out predictive accuracy. These metabolites are illustrated in Figure 5d and the top ten are listed in Table 1. Figure 5d shows that SVM-RFE favors the metabolites with small standard derivation.

Our LPFS method finally selects 4 features as the biomarkers to discriminate ST36 and Pre1. By using only 4 features we can achieve 100% leave-one-out predictive accuracy. To show these four important biomarkers are not dependent on the nearest centroid classifier, we use SVM to do five-fold cross validation, the predictive accuracy is still 100%. This demonstrates that we can select a small set of important features really matters by applying strong regularization. The selected 4 metabolites are listed in Table 1 and scatter plotted in Figure 5e. Figure 5e also shows that our LPFS method tends to reveal the metabolites with small standard deviation and large shifts, which exactly serves our requirement for good biomarker. The four metabolites are with ppm 3.55, 3.54, 3.49, and 1.33. ppm 3.54 and 1.33 are annotated as a-glucose/glycine and lactate respectively. The expression level of ppm 3.54, 3.49, and 1.33 goes down after acupuncture treatment while the expression level of ppm 3.55 goes up.

Secondly, we compare the results of these five methods in a venn diagram in Figure 6. Figure 6a shows the overlap among the results from fold change, SVM-RFE, SMLR, and LPFS. The 4 metabolites selected by our LPFS, metabolites with ppm 3.55, 3.54, 3.49, and 1.33, also correctly selected by fold change, SVM-RFE, and SMLR. Interestingly, all the results obtained by SVM-RFE, SMLR, and LPFS are included in the results by fold change. Figure 6b shows the overlap among the results from t-test, SVM-RFE, SMLR, and LPFS. Again, metabolites with ppm 3.55, 3.54, 3.49 are consistently supported by SVM-RFE, SMLR, and t-test. Metabolite with ppm 1.33 is not included in the t-test result but supported by SVM-RFE and SMLR. SVM-RFE and SMLR also select some metabolites which are not included by t-test results. The venn diagram demonstrates that fold change method is more consistent with current feature selection methods when the sample number is not so large.

Table 2 Identified biomarkers from different meridian points by our LPFS method.

Zusanli ST36			Liangmen ST21			Juliao ST3			Yanglingquan GB34			Weizhong BL40		
Metabolite	ppm	ID	Metabolite	ppm	ID	Metabolite	ppm	ID	Metabolite	ppm	ID	Metabolite	ppm	ID
	3.55	86		**2.11**	230		**3.55**	86		**3.55**	86		**3.78**	63
a-glucose/ glycine	**3.54**	87		**0.88**	353	a-glucose/ glycine	**3.54**	87	a-glucose/ glycine	**3.54**	87		**3.99**	42
	3.49	92	histidine/taurine	**3.25**	116				threonine	**1.32**	309		**3.88**	53
lactate	**1.33**	308		**3.55**	86							lipid	**1.3**	311
			lactate	**1.33**	308							lysine/ arginine	**1.91**	250
			a-glucose/ glycine	**3.54**	87								**3.92**	49
				3.2	121								**3.49**	92
													3.2	121

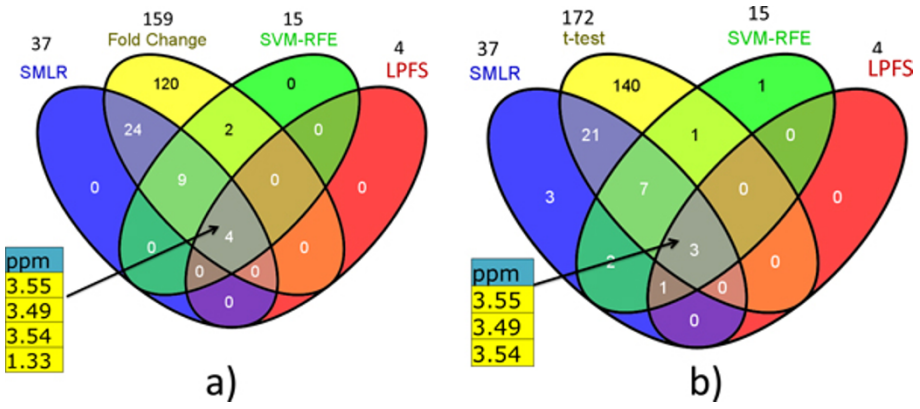

Figure 6 Venn diagram for the results obtained by different methods. a) Comparing t-test, SMLR, SVM-RFE, and LPFS methods by checking the overlaps of their selected biomarkers. b) Comparing fold change, SMLR, SVM-RFE, and LPFS methods.

In addition to the overall venn diagram, the top ten biomarkers obtained by the t-test, fold change, SVM-RFE methods are summarized in Table 1. When comparing the overlap with top 10 list, LPFS still gets consistent results with other methods. Metabolite with ppm 3.55 has a t-test score 15.29 and fold change score 75.38. It is ranked the first by all the three methods with rank information. Metabolites with ppm 3.54 and 3.49 rank high in fold change, SVM-RFE. ppm 1.33 ranks high in the result obtained by fold change.

Biological insights for the identified biomarkers

We then applied the proposed LPFS method to identify the biomarkers from the designed seven experiments. As a result, we identified 4, 7, 2, 3, and 8 biomarkers for the acupuncture treatment effects of ST36, ST21, ST3, GB34, and BL40 respectively. These selected biomarkers can achieve 100%,100%,100%,100%, and 95% leave-one-out cross validation accuracy. The results are summarized in Table 2. As expected, Exp7 fails to find any biomarkers. Exp6 finds several metabolites due to the fact that the expression values of these metabolites vary after consecutive 5 days. So we carefully check the obtained metabolites list by Exp6 and exclude these metabolites in our final results. Some biomarkers identified in Table 2 are annotated as glucose and lipid. Most of them need further investigation on their chemical structures and biological functions.

From Table 2, we can see that acupuncture at Yangming meridian points (including acupuncture points at ST36, ST21, and ST3) influence mainly plasma micromolecular metabolites and was closely related to energy metabolism pathway. Acupuncture at Yanglingquan influences mainly plasma macromolecular metabolites and is closely related to lipid metabolism and transport. Acupuncture at Weizhong doesn't largely influence plasma metabolites. This study suggests that Yangming meridian points have certain characteristics, which are different from those of both Yanglingquan and Weizhong. Metabonomics techniques based on ^{1}H NMR and biomarker identification method provide experimental evidence for distinguishing between Yangming meridian points and other meridian points from the metabolic aspect. This fact may become a new useful information source to study the specificities of meridian points.

To reveal the similarity and difference of the identified biomarkers regarding to meridian points, we calculate the overlaps of these biomarkers and present them in Figure 7. We found that Weizhong is slightly different from other meridian points. There is no overlapped metabolites for Weizhong and other meridian points. Compared to that, Zusanli, Yanglingquan, Juliao, and Liangmen are close to each other by sharing two common metabolites: ppm 3.54 and 3.55 (Juliao is not

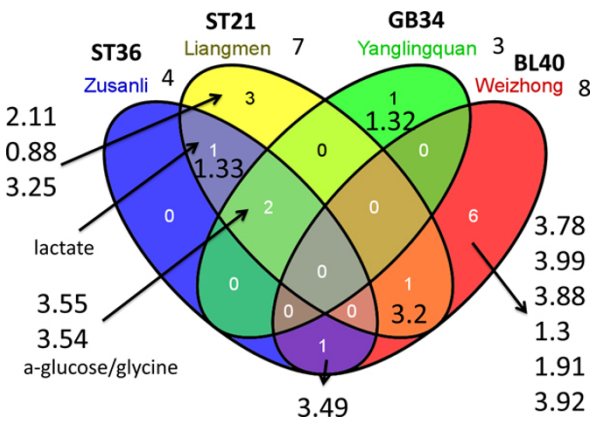

Figure 7 Venn diagram for identified biomarkers from different acupuncture points. The similarity and difference of those identified biomarkers from Zusanli, Liangmen, Yanglingquan, and Weizhong are shown in a venn diagram. The overlapped biomarkers are indicated by their ppm and known annotation.

shown in Figure 7. The two biomarkers from Juliao are ppm 3.55 and 3.54 and totally included by Zusanli, Yanglingquan, and Liangmen). In them 3.54 is annotated as a-glucose/glycine. Among the five points, Zusanli, Liangmen, and Juliao are on the same meridian. From the venn diagram in Figure 7, we can clearly see this trend. Yanglingquan has a unique biomarker with ppm 1.32. Our biomarker analysis indicates that acupuncture has some common molecules in metabolic level. At the same time, acupuncture at different meridian points has different molecular response. Our result is consistent with the theory on specificities of meridian points.

Our results show that metabolite with chemical shift value 3.55 is clearly a common biomarker for ST36, ST21, ST3, and GB34. In Figure 8, we visualize the metabolic profiles as a two-dimensional graph and highlight this important molecule. The two dimensional graph, called the GEDI-"mosaics", provide a unique, one-glance visual engram that gives each high-dimensional sample a face. A characteristic of GEDI's analysis is that it does not prejudicate any particular structure in the data (such as clusters or hierarchical organization). Thus, it allows the researcher to use human pattern recognition to perform a global first-level analysis of the data [29] (GEDI is downloaded from http://www.childrenshospital.org/research/ingber/GEDI/gedihome.htm). It is clear that the highlighted metabolite has distinct expression value in case and control group (ST36 and Pre1 in Figure 8). This demonstrates the effectiveness of our biomaker identification method.

Importantly, our LPFS method reveals the metabolite with ppm 1.33 as a biomarker for meridian points ST36 and ST21. This molecule is annotated as lactate. Lactate has been extensively studied over years for many important functions. For example, the lactate has always been regarded as the central nervous system metabolic waste and a sign of hypoxia [30]. Also in recent years, evidences show that the role of lactate in brain energy metabolism should be re-recognized. Firstly, investigations find that lactate is not only a very important

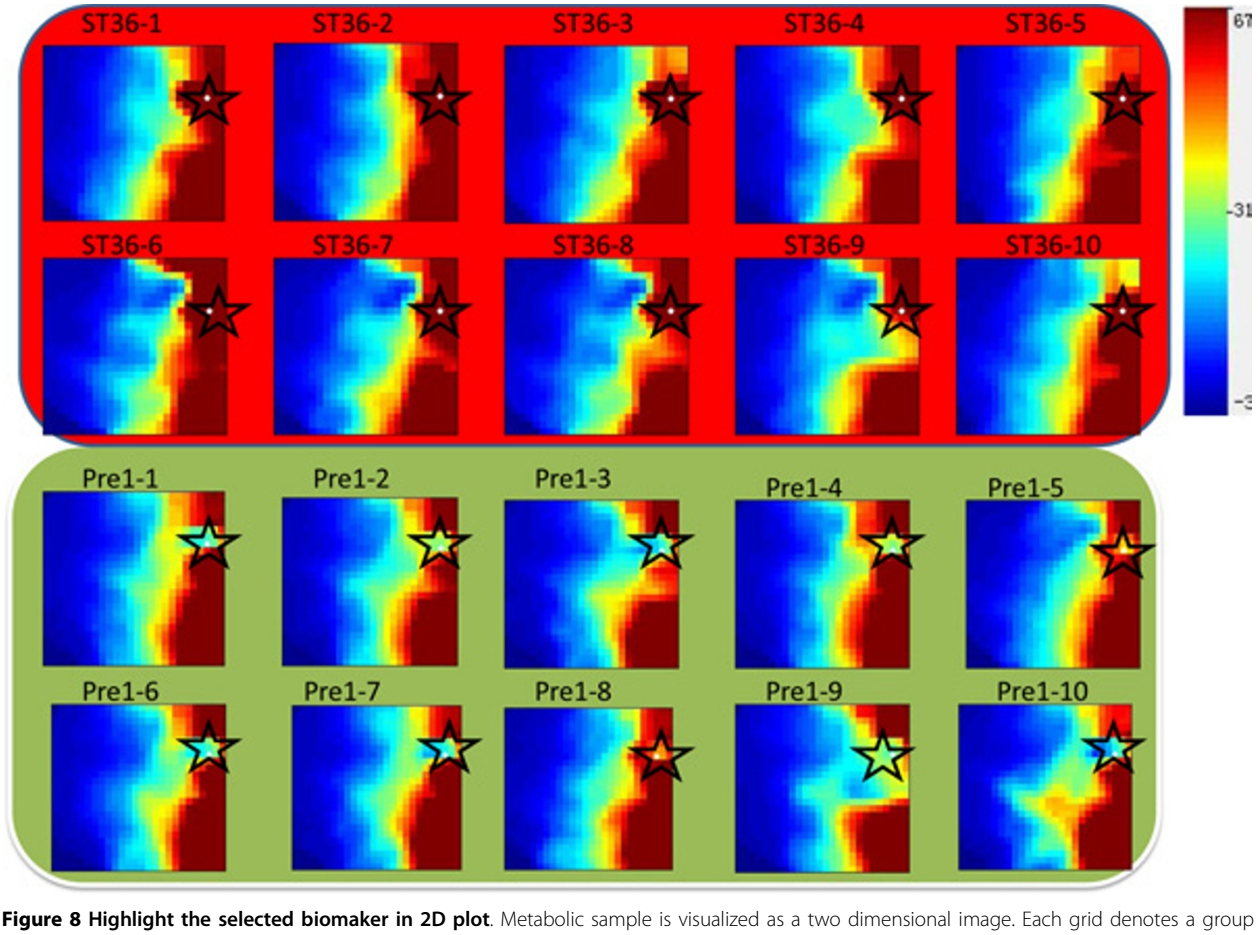

Figure 8 Highlight the selected biomaker in 2D plot. Metabolic sample is visualized as a two dimensional image. Each grid denotes a group of metabolites with similar profiles. Red color means the highly expressed metabolite group and blue color means the lowly expressed metabolite group. In particular, metabolite with chemical shift value 3.55 is highlighted in white color and indicated by the star.

energy resource for brain, but a sensor of brain energy homeostasis [31]; secondly, lactate in adult brain mainly comes from astrocytes, and it is a collaborative carrier for neuron and astrocytes [32,33]; thirdly, lactate plays an important role in coupling energy metabolism and functional activity [34]; and lastly, lactate has neuroprotective effect in a number of pathological conditions. It's interesting to see lactate as a biomaker closely relating to acupuncture in our result. Whether lactate plays a substantial role in acupuncture effect at ST36 and ST21 needs follow-up biological experiments.

Discussions and conclusions

Biomaker identification or feature selection considers the problem of constructing a prediction rule from only a feature-subset and accurately classifying the context of diagnosis and treatment observations (e.g. with vs. without acupuncture treatment). Such problems have become increasing important and quite general in genomics (identifying differentially expressed genes in microarray data), proteomics (finding promising protein marker from the mass spectrometry data), metabolics (selecting metabolite markers from NMR, GC-MS data), and other areas of computational biology. Due to the number of features is much larger than the number of observations, simple, highly regularized approaches are in pressing need. Here, we proposed a novel linear programming based feature selection (LPFS) model to address this important problem. The feature selection problem is cast into an optimization problem with two objectives, one is to minimize the number of chosen features and the other is to maximize the predictive accuracy. Mathematically the feature selection problem is formulated as a mixed integer linear programming problem. Then the model is further relaxed to linear programming to ensure the efficient identification of a feature-subset. We can solve the in-essence combinatorial optimization problem in a computational reasonable way. In summary, our LPFS method can select feature and learn the classifier in a joint way and we can select a small set of features by applying strong regularization. Our methodology is general and can be easily applied to other scenarios [35].

We extensively compared our LPFS method with existing methods in the real datasets on acupuncture treatment at different acupoints. We find that, 1). Our method can select the fewest features while achieve accurate predictions. 2). Our method is free of arbitrary threshold choice. 3). Close check of the selected feature shows that our method can identify those biological meaningful features. 4). In addition, the cross-validation results show that our method can achieve relatively high accuracy in prediction.

Prior information allows further improvement of our method. Currently the identified biomarkers are independent to each other. We can move further step to interpretation by considering a group of biologically meaningful biomarkers. For example, we can incorporate the network information (interactions among features) into the feature selection procedure. As a result, a pathway or modules in the network will be finally selected instead of single molecule as the biomarker, so called network biomarker. We note that prior information can be easily incorporated into our optimization model either by adding some constraints or penalizing in the objective function.

In this paper, the biomarker identification for each acupuncture point is treated as a single binary classification task. We then compare the revealed biomarkers for their similarity and difference across different acupuncture points. We note that a multi-classifier can be developed to systematically integrate all the profiles from different points together. This topic is in progress as our further direction.

Finally, the metabolic profile is known for its high variance. We note that the main source of variance is from NMR technology instead of acupuncture effect [36]. To maximally reduce the variance from metabolic profile, we carefully design our experiments. Firstly, we used the relatively stable blood samples instead of urine sample. Secondly, we specifically use the Pareto scaling and orthogonal signal correction (OSC) method [37-39] to normalize the raw data, which will reduce the variance of samples inside each group and enhance the differences among groups. Even with all these efforts, the remaining high variance may due to the change of environments and conditions and will eventually prevent the high accuracy for identification of biological meaningful biomarkers. In addition to variance, the limited number of sample in our study may also bring some potential effects on the results. From this viewpoint, these identified biomarkers should be carefully validated for their biological functions. Also, additional control study should be carefully designed to exclude other possible cofactors. Further integration of the data from other levels, such as gene expression and proteomics levels, will improve the robustness of the identified biomarker.

Acknowledgements

The authors would like to thank Prof. Luonan Chen, Dr. Ruisheng Wang, and ZHANGroup members for insightful discussions. YW, LYW, and XSZ are supported by NSFC grant 61171007, 11131009, 60970091, and CAS grant kjcx-yw-s7. QFW and FRL are supported by NSFC grant 30901933 and National Basic Research Program of China (no.2012CB518500). YW is also supported by SRF for ROCS, SEM and the Shanghai Key Laboratory of Intelligent Information Processing (No. IIPL-2010- 008).

This article has been published as part of *BMC Systems Biology* Volume 6 Supplement 1, 2012: Selected articles from The 5th IEEE International Conference on Systems Biology (ISB 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/6/S1.

Author details
[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. [2]Acupuncture and Moxibustion College, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China. [3]National Center for Biomedical Analysis, Beijing 100850, China. [4]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China.

Authors' contributions
YW proposed the computational method. QFW, XZY, SGY, and FRL designed the experimental study and generated the data. YW, CC, LYW, and XSZ implemented the method, performed the experiments and analyzed the data. YW and QFW wrote the manuscript. All authors revised the manuscript and approved the final version.

Competing interests
The authors declare that they have no competing interests.

Published: 16 July 2012

References

1. Langevin H, Churchill D, Cipolla M: **Mechanical signaling through connective tissue: a mechanism for the therapeutic effect of acupuncture.** *FASEB J* 2001, **15(12)**:2275-2282.
2. Sims J: **The mechanism of acupuncture analgesia: a review.** *Complementary Therapies in Medicine* 1997, **5(2)**:102-111.
3. Cab[001]oglu M, Ergene N, Tan U: **The mechanism of acupuncture and clinical applications.** *International journal of neuroscience* 2006, **116(2)**:115-125.
4. Chen L, Wang R, Zhang X: *Biomolecular networks: methods and applications in systems biology, Volume 10* John Wiley & Sons Inc; 2009.
5. Wang Y, Joshi T, Zhang X, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets.** *Bioinformatics* 2006, **22(19)**:2413.
6. Hastie T, Tibshirani R, Friedman J, Franklin J: **The elements of statistical learning: data mining, inference and prediction.** *The Mathematical Intelligencer* 2005, **27(2)**:83-85.
7. Amaldi E, Kann V: **On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems.** *Theoretical Computer Science* 1998, **209(1-2)**:237-260.
8. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *The Journal of Machine Learning Research* 2003, **3**:1157-1182.
9. Wang Y, Wu Q, Chen C, Yan X, Yu S, Zhang X, Liang F: **Identifying biomarkers for acupuncture treatment via an optimization model.** *IEEE International Conference on Systems Biology* IEEE; 2011, 319-326.
10. Park H, Jeon M, Rosen J: **Lower dimensional representation of text data based on centroids and least squares.** *BIT Numerical mathematics* 2003, **43(2)**:427-448.
11. Levner I: **Proteomic pattern recognition.** *Tech rep* Technical report, University of Alberta, No: TR04; 2004.
12. Zhang X, Lu X, Shi Q, Xu X, Leung H, Harris L, Iglehart J, Miron A, Liu J, Wong W: **Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.** *BMC bioinformatics* 2006, **7**:197.
13. Wang Y, Wang R, Joshi T, Xu D, Zhang X, Chen L, Xia Y: **A Linear Programming Framework for Inferring Gene Regulatory Networks by Integrating Heterogeneous Data.** In *Computational Methodologies in Gene Regulatory Networks* Das S, Caragea D, Hsu WH, Welch SM, IGI Global 2010, 450-475.
14. Wang Y, Zhang X, Chen L: **Optimization meets systems biology.** *BMC systems biology* 2010, **4(Suppl 2)**:S1.
15. Wu D, Chen A, Johnson C: **An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses.** *Journal of Magnetic Resonance, Series A* 1995, **115(2)**:260-264.
16. Beckwith-Hall B, Thompson N, Nicholson J, Lindon J, Holmes E: **A metabonomic investigation of hepatotoxicity using diffusion-edited 1H NMR spectroscopy of blood serum.** *Analyst* 2003, **128(7)**:814-818.
17. Holmes E, Nicholls A, Lindon J, Ramos S, Spraul M, Neidig P, Connor S, Connelly J, Damment S, Haselden J, *et al*: **Development of a model for classification of toxin-induced lesions using 1H NMR spectroscopy of urine combined with pattern recognition.** *NMR in Biomedicine* 1998, **11(4-5)**:235-244.
18. Waters N, Holmes E, Williams A, Waterfield C, Farrant R, Nicholson J: **NMR and pattern recognition studies on the time-related metabolic effects of *α*-naphthylisothiocyanate on liver, urine, and plasma in the rat: An integrative metabonomic approach.** *Chemical research in toxicology* 2001, **14(10)**:1401-1412.
19. Wu Q, Zhang Q, Sun B, Yan X, Tang Y, Qiao X, Chen Q, Yu S, Liang F: **1H NMR-based metabo-nomic study on the metabolic changes in the plasma of patients with functional dyspepsia and the effect of acupuncture.** *Journal of pharmaceutical and biomedical analysis* 2010, **51(3)**:698-704.
20. Wu Q, Xu S, Yan Z, Yu S, Tang Y, Liu J, Mao S, Zhou S, Liang F: **Metabonomics and pattern recognition study on the specificity of foot-yangming Meridian points.** *Shanghai J Acu-mox* 2010, **29(9)**.
21. Tibshirani R: **A comparison of fold-change and the t-statistic for microarray data analysis.** *Analysis* 2007, 1-17.
22. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines.** *Machine learning* 2002, **46**:389-422.
23. Duan K, Rajapakse J, Wang H, Azuaje F: **Multiple SVM-RFE for gene selection in cancer classification with expression data.** *IEEE Trans Nanobioscience* 2005, **4(3)**:228-234.
24. Krishnapuram B, Carin L, Hartemink A: **Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data.** *Journal of Computational Biology* 2004, **11(2-3)**:227-242.
25. Krishnapuram B, Carin L, Figueiredo M, Hartemink A: **Sparse multinomial logistic regression: Fast algorithms and generalization bounds.** *IEEE Trans Pattern Anal Mach Intell* 2005, **27**:957-968.
26. Pratapa PN, Patz EF Jr, Hartemink AJ: **Finding Diagnostic Biomarkers in Proteomic Spectra.** *Pac Symp Biocomput* 2006, 279-290.
27. Witten D, Tibshirani R: **A comparison of fold-change and the t-statistic for microarray data analysis.** *Tech rep* Department of Statistics, Stanford University technical report; 2007.
28. Cui X, Churchill G: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4(4)**:210.
29. Eichler G, Huang S, Ingber D: **Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles.** *Bioinformatics* 2003, **19(17)**:2321.
30. Song Z, Routh V: **Differential effects of glucose and lactate on glucosensing neurons in the ventromedial hypothalamic nucleus.** *Diabetes* 2005, **54**:15.
31. Patil G, Briski K: **Lactate is a critical "sensed" variable in caudal hindbrain monitoring of CNS metabolic stasis.** *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 2005, **289(6)**:R1777.
32. Castro M, Beltrán F, Brauchi S, Concha I: **A metabolic switch in brain: glucose and lactate metabolism modulation by ascorbic acid.** *Journal of neurochemistry* 2009, **110(2)**:423-440.
33. Mangia S, Simpson I, Vannucci S, Carruthers A: **The in vivo neuron-to-astrocyte lactate shuttle in human brain: evidence from modeling of measured lactate levels during visual stimulation.** *Journal of neurochemistry* 2009, **109**:55-62.
34. Dienel G, Hertz L: **Glucose and lactate metabolism during brain activation.** *J Neurosci Res* 2001, **66(5)**:824-838.
35. Srinivas P, Kramer B, Srivastava S: **Trends in biomarker research for cancer detection.** *The lancet oncology* 2001, **2(11)**:698-704.
36. Bollard M, Stanley E, Lindon J, Nicholson J, Holmes E: **NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition.** *NMR in Biomedicine* 2005, **18(3)**:143-162.
37. Keun H, Ebbels T, Antti H, Bollard M, Beckonert O, Holmes E, Lindon J, Nicholson J: **Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling.** *Analytica chimica acta* 2003, **490(1-2)**:265-276.

38. Beckwith-Hall B, Brindle J, Barton R, Coen M, Holmes E, Nicholson J, Antti H: **Application of orthogonal signal correction to minimise the effects of physical and biological variation in high resolution 1H NMR spectra of biofluids.** *Analyst* 2002, **127(10)**:1283-1288.
39. Gavaghan C, Wilson I, Nicholson J: **Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA.** *FEBS letters* 2002, **530(1-3)**:191-196.

doi:10.1186/1752-0509-6-S1-S15
Cite this article as: Wang *et al.*: Revealing metabolite biomarkers for acupuncture treatment by linear programming based feature selection. *BMC Systems Biology* 2012 **6**(Suppl 1):S15.