# Identifying Biomarkers for Acupuncture Treatment via an Optimization Model

Yong Wang*, Qiao-Feng Wu†, Chen Chen*, Xian-Zhong Yan‡, Shu-Guang Yu†,
Xiang-Sun Zhang*, Fan-Rong Liang †

*National Center for Mathematics and Interdisciplinary Sciences,
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing, China 100190
Email: ywang@amss.ac.cn, zxs@amt.ac.cn
†School of Acupuncture‖moxibustion and Massotherapy
Chengdu University of Traditional Chinese Medicine
Chengdu, China 610075
Email: qiaofengwu@yahoo.cn
‡National Center for Biomedical Analysis
Beijing, China 100850

***Abstract*—Identifying biomarkers for acupuncture treatment is crucial to understand the mechanism of acupuncture effect at molecular level. In this study, we investigate the metabolic profiles of acupuncture treatment on several meridian points in human. To identify the subsets of metabolites that best characterize the acupuncture effect for each meridian point, a linear programming based model is proposed to identify biomarkers from the high-dimensional metabolic data. Specifically, we use nearest centroid as prototype to simultaneously minimize the number of selected features and leave-one-out cross validation error of the classifier. As a result, we reveal novel metabolite biomarkers for acupuncture treatment. Our result demonstrates that metabolic profiling might be a promising method to investigating the molecular mechanism of acupuncture. Comparison with other existing methods shows the efficiency and effectiveness of our new method. In addition, the method proposed in this paper is general and can be used in other high-dimensional applications, such as cancer genomics.**

## I. Introduction

Acupuncture, an important therapeutic method in Traditional Chinese Medicine (TCM), has been used to treat various diseases for thousand years in China. 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 researches have been conducted on the mechanism of acupuncture to explain the effects of acupuncture on various systems and symptoms [1]. Compared to acupuncture, systems biology is a relatively new term to describe the recent trends in bioscience research. It emphasizes the high-throughput measurement of biological systems and focuses on the complex interactions in biological systems[2], [3]. 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 study the acupuncture treatment effect by identifying a subset of important molecules. 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 (with acupuncture treatment in meridian point) and control samples (without acupuncture treatment). 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, large number of features (metabolites) 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 somewhat prone to the curse of dimensionality. Second, predictive accuracy is not the only goal and further biological validation and mechanism understanding call for intuition other than black box predictive results. Thus it is especially important to know which molecules contribute towards the classification. Ideally we can improve the generalization performance of our classifier by identifying only the molecules that are relevant 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

YW and QFW contributed equally to this work. XSZ and FRL are co-corresponding authors.

2011 IEEE International Conference on Systems Biology (ISB)
978-1-4577-1666-9/11/$26.00 ©2011 IEEE



Zhuhai, China, September 2–4, 2011

widely deployed in clinical settings. Knowledge of a small set of diagnostically relevant metabolites may provide important insights into the mechanisms responsible for acupuncture effect itself. Those molecules are usually termed as biomarkers. The procedure to reveal them is referred as feature selection, biomarker identification, or feature ranking.

If 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, the method is not practical for realistic applications.

Existing feature selection strategies can be roughly categorized into three types. Exploiting the partial ordering properties of the space of subsets, we can either start with an empty set and successively add features, or start with the set of all features and successively remove 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. As an example of forward feature selection, we might first look for the single most discriminative feature using any classifier design algorithm. Then we could search the single additional feature that gives the best class discrimination when considered along with the first feature. By keeping augmenting the feature set iteratively in this greedy fashion we stop until cross-validation error estimates are minimized.

However we cannot obtain the global optimal solution by adopting the above greedy strategies. 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. In other words, our feature selection method simultaneously improving classification accuracy and selecting features based on the centroid classification framework. We then apply our method to analyze the metabolite profile data. 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.

## II. Methods

### A. Metabonomics data generation

To investigate the acupuncture treatment effects, we originally generated metabonomics data for Yangming meridian points and other meridian points on plasma metabolites in healthy males 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. As a result, most

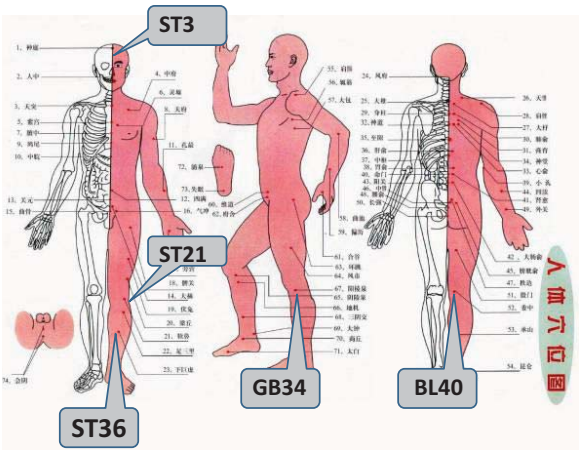



Fig. 1. Metabolite profiles are originally generated by $^1$H NMR from five meridian points.

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 to be 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, in total 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 1. Among the five points, Zusamli, liangmen,and juliao are on the same meridian.). Each group contains 10 personsInside each group the corresponding meridian points were separately acupunctured for 5 consecutive daysIn 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 profilesFurthermore 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 [4]. 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).


978-1-4577-1666-9/11/$26.00 ©2011 IEEE

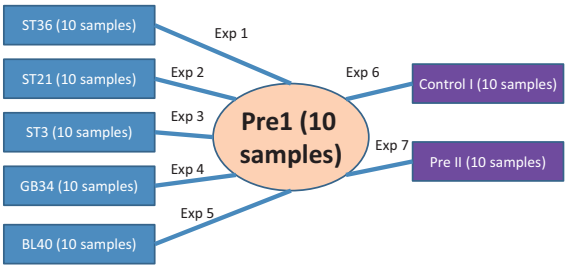



Fig. 2. Overall design of the biomarker identification experiments.

### B. 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 2. We categorize eighty samples into 8 groups shown as the circles in Figure 2). 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 Pre1 as the common control set, we have seven binary classification tasks (Exp1 to Exp7) shown as the lines in Figure 2. 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. 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.

### C. Centroid classification prototype

A fast and simple algorithm for classification is the centroid method [6], [7]. 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. A prototype pattern for class $C_j$ is defined as the arithmetic mean:

$$\boldsymbol{\mu}_{C_j} = \frac{1}{|C_j|} \sum_{\boldsymbol{s}_i \in C_j} \boldsymbol{x}_i \tag{1}$$

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:

$$C(\boldsymbol{s}) = \text{argmin}_{C_j} dis(\boldsymbol{\mu}_{C_j}, \boldsymbol{s}) \tag{2}$$

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

$$C(\boldsymbol{s}) = \text{argmax}_{C_j} sim(\boldsymbol{\mu}_{C_j}, \boldsymbol{s}) \tag{3}$$

where $sim(\boldsymbol{x}, \boldsymbol{y})$ is a similarity metric. This simple classifier will form the basis of our studies. It works with any number of features. 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 [8], we select $L_1$ distance metric, which is most appropriate for the centroid classification algorithm. It is defined by:

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

with $\|y\|_1 = \sum_i |y(i)|$, and $y(i)$ being the value of the $i$-th feature. 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.

### D. Feature selection by linear programming

Suppose we have two groups in the training dataset, the case group and the control group as the gold standard to classify new samples. We denote them $T$ 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 normalized 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

$$s \in T \quad \text{if} \quad \|\boldsymbol{s} - \boldsymbol{\mu}_T\|_1 < \|\boldsymbol{s} - \boldsymbol{\mu}_F\|_1 \tag{5}$$

$$s \in F \quad \text{if} \quad \|\boldsymbol{s} - \boldsymbol{\mu}_T\|_1 > \|\boldsymbol{s} - \boldsymbol{\mu}_T\|_1 \tag{6}$$

Let the feature number be $n$. We introduce the variables for feature section as $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, where $x_i = 0, 1$. When $x_i = 1$, it means feature $i$ is selected in the biomarker set. Otherwise it is note selected.

Suppose the test dataset is $U$. And it is composed by the case group $U_T$ and control group $U_F$. $U = U_T \bigcup U_F$. And $|U_T| = l_1$, $|U_F| = l_2$ With the preparation, we can introduce the constraints. If there is a case sample $\boldsymbol{s_l} = (s_{l1}, s_{l2}, \cdots, s_{ln}), l \in \{1, 2 \cdots, l_1\}$, if we want it to be classified correctly, we should have

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

where $\boldsymbol{t}_k = (t_{k1}, t_{k2}, \cdots, t_{kn}) \in T, k = 1, 2, \cdots, m_1$ and $\boldsymbol{f}_k = (f_{k1}, f_{k2}, \cdots, f_{kn}) \in F, k = 1, 2, \cdots, m_2$.

Similarly if there is a control sample $\boldsymbol{s_l} = (s_{l1}, s_{l2}, \cdots, s_{ln}), l \in \{l_1 + 1, l_1 + 2 \cdots, l_1 + l_2\}$, if


978-1-4577-1666-9/11/$26.00 ©2011 IEEE

we want it to be classified correctly, we should have

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

And the object function is to choose as few as features, i.e.,

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

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

When we consider the noise in the measured data, not all the test samples can be classified exactly. We introduce the tolerable error $\boldsymbol{y}=\{y_i, i\in 1,2,\cdots,l_1+l_2\}$ for every sample in $U_T\bigcup U_F$. And $y_i\geq 0$. When $y_i$ is large, it means sample $i$ is wrongly classified. Otherwise this sample is correctly classified.

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

$$\sum_{i=1}^{n}|s_{li}-\sum_{j=1}^{m_1}t_{ji}/m_1|x_i-y_l<\sum_{i=1}^{n}|s_{li}-\sum_{j=1}^{m_2}f_{ji}/m_2|x_i \quad (11)$$

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

$$\sum_{i=1}^{n}|s_{li}-\sum_{j=1}^{m_1}t_{ji}/m_1|x_i+y_l>\sum_{i=1}^{n}|s_{li}-\sum_{j=1}^{m_2}f_{ji}/m_2|x_i \quad (12)$$

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) $\min_{y_1,y_2,\cdots,y_{l_1+l_2}}\sum_{i=1}^{l_1+l_2}y_i$. In general, there is a trade-off relation 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 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ and $\boldsymbol{y}=(y_1,y_2,\cdots,y_{l_1+l_2})$ as shown in Equation (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. 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 $\epsilon$-method, which alternates a positive scalar parameter $\lambda$ to obtain the Pareto set, with the formulation in Equation (14).

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 $\lambda$ for the single-objective optimization problem (14). Clearly, $\lambda$ transforms the number of chosen features into equivalent classification error in (14), and controls the balance between them.

By solving the proposed linear programming model (14), we can get the solutions for the feature selection variables $x_i,\quad i\in\{1,2,\cdots,n\}$, and classification error variables $y_j,\quad j\in\{1,2,\cdots,l_1+l_2\}$. Checking if $x_i$ is equal to 1, we can know if the corresponding feature should be involved in the classification. Meanwhile checking the values of all the $y_j$, we can give 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 estimate the classification accuracy by $N_1/l_1$ and $N_2/l_2$.

The above model (14) is based on the general idea of cross validation, thus it depends on the choice of $T$ and $F$. Specifically we can choose a model for leave-one-out cross validation (resubstitution test) in Equation (15).

We adopt leave-one-out experiment since leave-one-out is an unbiased estimator of the generalization performance of classifier. i.e., 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. Furthermore, ILP can be relaxed into the corresponding linear programming (LP). Therefore, an LP algorithm can be adopted to efficiently solve this ILP. In terms of computational complexity, the proposed approach makes the computation of biomarker tractable. Finally we construct the new model in Equation (16).

It should be noted that we can use other distances instead of $L_1$ in our model to achieve the nonlinear classification effect. The parameter $\lambda$ is determined by checking the output leave-one-out predictive accuracy. We notice that our model can be extended to multi-classification task and $n$-fold cross validation experiment.

## III. Results

### A. Global characterization of the data

We first perform hierarchical clustering on the 80 metabolic profiles. The results are shown in Figure 3. 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 3) and we cannot see clearly boundaries.

Furthermore, we calculate the centroids for the seven groups of samples in Figure 2 by averaging the 10 samples for their expression values. These centroids are plotted side by side in Figure 3, which shows that these centroids are very similar and it's very difficult to detect the 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 biomarkers to achieve clear discrimination.

### B. Biological insights for the identified biomarkers

We then applied the proposed optimization 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 effect of ST36, ST21, ST3, GB34,


978-1-4577-1666-9/11/$26.00 ©2011 IEEE

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

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

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

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

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

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 1. 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 biomarker list and exclude these metabolites in our final results. Some biomarkers identified in Table 1 are annotated as glucose and lipid. Most of them are new to us and are under further investigation.


978-1-4577-1666-9/11/$26.00 ©2011 IEEE

| 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 |

TABLE I
IDENTIFIED BIOMARKERS FOR DIFFERENT MERIDIAN POINTS.

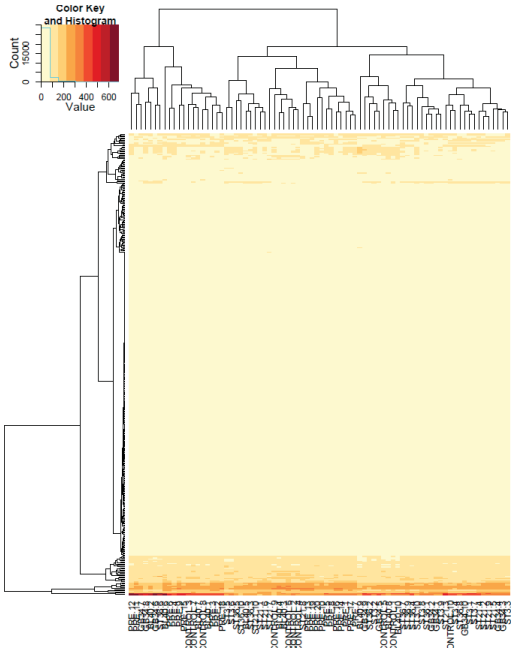



Fig. 3. Hierarchical clustering of the metabolic profiles of the 80 samples.


Fig. 4. Centroids for the seven datasets. The horizontal units are expression values for the metabolites. The metabolites are sorted by their chemical shift values.

From Table 1, we can see that acupuncture at Yangming meridian points (including acupuncture points at ST36, ST21, and ST3) influenced 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 and may become a new useful means to study the specificities of meridian points.

Our result shows that metabolite with chemical shift value 3.55 is clearly a common biomarker for ST36, ST21, ST3, and GB34. In Figure 5, 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 [13] (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 5). This demonstrates the effectiveness of our biomaker identification method.

### C. Comparison with other approaches

We compare our optimization based method with several existing methods. 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 ([9]) as

$$Ti = \frac{\bar{x}_i - \bar{y}_i}{s_i} \tag{13}$$

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 $i$.


978-1-4577-1666-9/11/$26.00 ©2011 IEEE

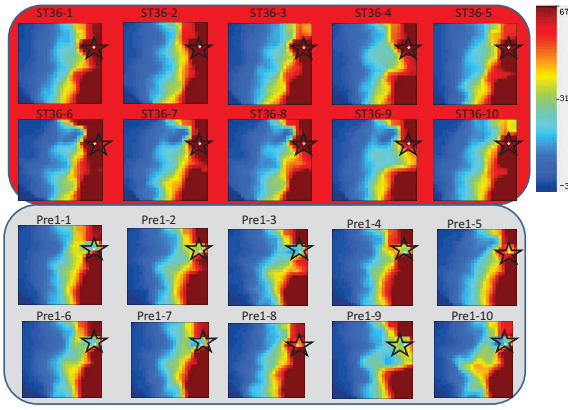



Fig. 5. Metabolic sample is visualized as a two dimensional graph. Each grid denotes a group of metabolites with similar profiles. Red color means the highly expressed metabolite group and blue color means the lowly expressed metabolite group. Particularly metabolite with chemical shift value 3.55 is highlighted in white color and indicated by the star.

The standard definition of the fold-change ([9]) for metabolite $i$ is

$$FC_i = \frac{\bar{x}_i}{\bar{y}_i} \tag{14}$$

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.

Since our method simultaneously optimize classification accuracy and the number of selected features, we 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 optimizing a single posterior objective function (see [10] and [11]). SMLR has been extensively applied in problems in systems biology [12]. 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 the Exp1 (ST36) as an example. The results obtained from 4 methods are listed and compared in Table 2. The t-test based method identifies 84 metabolites if we choose a cutoff 2.84 (corresponding p-value 0.005). Top 10 are listed in Table 2. The fold change based method identifies 97 metabolites by choosing a cutoff 4 (top 10 are listed in Table 2). While SMLR select 37 features to achieve the 100% leave-one-out predictive accuracy. Our LP based method finally selected 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 classifier, we use SVM to do five-fold cross validation, the predictive accuracy is still 100%. This demonstrate that we can select a small set of important features really matters by applying strong regularization.

In Figure 6, we compare different methods 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. We find that the ordinary t-statistic selects genes with low standard deviations. The fold-changes select genes with large shifts between control and treatment. While our LP method tends to reveal the metabolites with small standard deviation and large shifts, which exactly serves our requirement for good biomarker.

## IV. 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 data), and other areas of computational biology. Due to the number of features is much larger than the number of observations, simple and highly regularized approaches are in pressing need. Here, we proposed a novel linear programming (LP) 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 an integer linear programming problem. Then the model is further relaxed to linear programming to ensure the efficient identification a feature-subset in a fast way. We can solve the in-essence combinatorial optimization problem in a computational reasonable way. In summary, Our LP based 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 other scenarios.

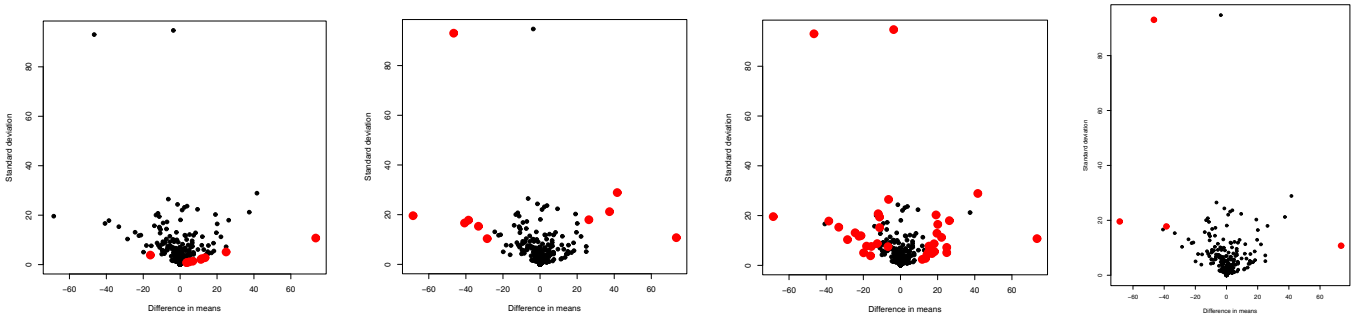
We extensively compared our LP based method with existing methods in some real datasets on acupuncture treatment. We find that, 1). our method can select the fewest features while achieve accurate predictions. 2). our method is free of arbitrary threshold choice. 3). close check of the selected feature 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 to interpretation by considering a group of biological meaning 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 the prior information can be easily incorporated into our optimization model either by adding some constraints or penalizing in the objective function.


978-1-4577-1666-9/11/$26.00 ©2011 IEEE

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

TABLE II
IDENTIFIED BIOMARKERS BY DIFFERENT METHODS ON THE ST36 MERIDIAN POINT.




(a) Student t-test (b) Fold change (c) SMLR (d) LP based method

Fig. 6. Comparison of our optimization 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.

## ACKNOWLEDGMENT

The authors would like to thank Prof. Luonan Chen, Dr. Ruisheng Wang, and ZHANGroup members for insightful discussions. YW and XSZ are supported by NSFC grant 10801131 and grant kjcx-yw-s7 from CAS. QFW and FRL are supported by NSFC grant 30901933. . YW is also supported by SRF for ROCS, SEM and the Shanghai Key Laboratory of Intelligent Information Processing (No. IIPL-2010- 008).

## REFERENCES

[1] M.T. Cabỳoglu, N. Ergene, and U. Tan. *The mechanism of acupuncture and clinical applications*, International journal of neuroscience, Vol.116 No.2, 2006.

[2] L. Chen, R. S. Wang, and X.S. Zhang. *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley & Sons, Hoboken, New Jersey. July, 2009.

[3] Y. Wang, T. Joshi, X. S. Zhang, D. Xu, and L. Chen. *Inferring gene regulatory networks from multiple microarray datasets*. Bioinformatics, 2006, 22:2413–2420.

[4] Qiaofeng Wu, Shizhen Xu, et al., *Metabonomics and pattern recognition study on the specificity of foot-yangming Meridian points*, Shanghai J Acu-mox, Vol.29 No.9, 2010.

[5] Ilya Levner. *Feature selection and nearest centroid classification for protein mass spectrometry* BMC Bioinformatics 2005, 6:68 doi:10.1186/1471-2105-6-68

[6] Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*. Springer Series in Statistics. Springer Verlag, New York; 2001.

[7] Park H, Jeon M, Rosen JB: *Lower dimensional representation of text data based on centroids and least squares*. BIT 2003, 43(2):1-22.

[8] Levner I: *Proteomic pattern recognition*. Technical report, University of Alberta, No: TR04-10 2004.

[9] Daniela Witten and Robert Tibshirani. *A comparison of fold-change and the t-statistic for microarray data analysis*. Technical report, Stanford university. 2007.

[10] B. Krishnapuram, L. Carin, and A. Hartemink. *Joint Classifier and Feature Optimization for Comprehensive Cancer Diagnosis Using Gene Expression Data.* Journal of Computational Biology, 11, pp. 227C242, 2004.

[11] B. Krishnapuram, M. Figueiredo, L. Carin, and A. Hartemink. *Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27, June 2005. pp. 957C968, 2005.

[12] P. Pratapa, E. Patz, and A. Hartemink. *Finding Diagnostic Biomarkers in Proteomic Spectra.* In Pacific Symposium on Biocomputing 2006 (PSB06), World Scientific: New Jersey. pp. 279C290, 2006.

[13] G.S. Eichler, S. Huang, D.E. Ingber, *Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles*, Bioinformatics, 19(17),2321-2, 2003.


978-1-4577-1666-9/11/$26.00 ©2011 IEEE