Metabolomics
DOI 10.1007/s11306-012-0410-z

ORIGINAL ARTICLE

Variable complementary network: a novel approach for identifying biomarkers and their mutual associations

Hong-Dong Li · Qing-Song Xu · Wan Zhang · Yi-Zeng Liang

Received: 23 September 2011 / Accepted: 15 February 2012

Abstract Biological variables involved in a disease process often correlate with each other through for example shared metabolic pathways. In addition to their correlation, these variables contain complementary information that is particularly useful for disease classification and prediction. However, complementary information between variables is rarely explored. Therefore, establishing methods for the investigation of variable's complementary information is very necessary. We propose a model population analysis approach that aggregates information of a number of classification models obtained with the help of Monte Carlo sampling in variable space for quantitatively calculating the complementary information between variables. We then assemble these complementary information to construct a variable complementary network (VCN) to give an overall visualization of how biological variables complement each other. Using a simulated dataset and two metabolomics datasets, we show that the complementary information is effective in biomarker discovery and that mutual associations of metabolites revealed by this method can provide information for exploring altered metabolic pathways. (The source codes for implementing VCN in MATLAB are freely available at: http://code.google.com/p/vcn2011/.)

Electronic supplementary material The online version of this article (doi:10.1007/s11306-012-0410-z) contains supplementary material, which is available to authorized users.

H.-D. Li · W. Zhang · Y.-Z. Liang (✉)
College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, People's Republic of China
e-mail: yizeng_liang@263.net

Q.-S. Xu
School of Mathematic Sciences, Central South University, Changsha 410083, People's Republic of China

Keywords Model population analysis · Variable complementary network · Monte Carlo sampling · Biomarker discovery · Variable selection

1 Introduction

Biological variables participating in a disease process are interlinked with each other through for example shared metabolic pathways and are thus correlated (Beasley and Planes 2007; Dunn et al. 2011; Holmes et al. 2006; Küffner et al. 2000). For example, serum level of cholesterol is closely correlated with low density lipoprotein (LDL) through the process of cholesterol transport (Arsenault et al. 2011). Correlation between variables is of importance for the understanding of physiological states, such as health and disease, and has been extensively investigated (Subramanian et al. 2005). In molecular information-based disease prediction, however, correlated variables contain a lot of redundant information that is not helpful in improving the accuracy of a predictive model (Rajalahti et al. 2009). Obviously, only the complementary information among multiple variables provides additional predictive value and should be therefore of particular use for biomarker identification and further disease prediction. Loosely speaking, in a prediction problem, for two out of p variables, their complementary information in the presence of the remaining $p - 2$ variables, refers to those predictive information that is gained by their combinatorial use. However, the complementary information is not, at least not explicitly, considered in established methods for biomarker discovery, such as genetic algorithm (Li et al. 2001), genetic programming (Liu and Xu 2009), random forest (Fan et al. 2011), recursive feature selection (Guyon et al. 2002) and so on. In all, to the

Published online: 29 February 2012

best of our knowledge, complementary information was rarely investigated.

Given a dataset of p variables, there are $p(p-1)/2$ different combinations if we analyze each pair of variables. However in this way, complementary information between different sets of multiple variables can not be considered. In fact, the total number of different combinations is 2^p, which poses an NP problem where the time needed to solve this kind of problem explodes exponentially with the increasing number of variables (Selman 2008). Since the intractability of this problem, one could only resort to other alternatives. To our knowledge, Monte Carlo sampling (MCS) may provide a possibility for the analysis of complementary information among multiple variables.

Here we report an approach, called variable complementary network (VCN), to explore and visualize the complementary information of each pair of variables conditioned on the remaining $p-2$ variables. This method is designed based on model population analysis (MPA) of which the basis is MCS (Li et al. 2009, 2010, 2011, 2012). In this method, N, e.g. 10,000, classification sub-models are first built using partial least squares-linear discriminant analysis (PLS-LDA) (Barker and Rayens 2003; Yi et al. 2006). Of note, each of these sub-models includes only Q variables random selected from the p variables. The reasons why we use only Q (usually $\ll p$) variables are mainly two folds: (1) the predictive performance of a variable in a model with a small number of variables can be more accurately assessed than in a model with a large number of variables and (2) it can lower the computational cost.

Next, to make the computation of complementary information conditioned on other variables feasible and also to take into account effects of variable combinations, we propose to define the complementary information between each pair of variables based on regression coefficients and predictive performances in terms of the prediction error of each classification sub-model. The rationale is that regression coefficients reflect the relative importance of each variable in a multivariate sense and prediction errors are a measure of the overall predictive ability of the Q variables in each model. Looking on the surface, the calculated complementary information is restricted to each pair of the Q variables. However, because the definition is based on a multivariate classification model, it indeed takes effects of multiple variable combinations into account. Finally, the computed complementary information between each pair of variables of all the N sub-models is summed to form a comprehensive VCN consisting of all the p variables. The performance of the proposed method is illustrated using one simulated dataset and two metabolomics datasets. The results show that the method is effective in analyzing how variables complement each other and further in singling out a sub-network consisting of variables associated with the disease under investigation.

2 Methods

MPA was proposed as a general framework for developing data analysis methods (Li et al. 2009). The basics and applications of MPA were recently reviewed (Li et al. 2012). As described in our previous work (Li et al. 2009, 2012), MPA works in three steps: (1) MCS is used to randomly draw N sub-datasets, e.g. 10,000, (2) For each sub-dataset, a sub-model is built, and (3) the last but not the least, an outcome of interest of all the N sub-models are statistically analyzed. This interesting outcome can be for example prediction errors associated with samples, regression coefficients of variables etc. By studying the distribution of this outcome, algorithms can be designed. As an example, we proposed a variable selection method by statistically comparing the two distributions of prediction errors before and after each variable is permuted (Li et al. 2010). Of note, MCS only serves as a technique for drawing sub-datasets. The core of MPA is the statistical analysis of an interesting outcome of all N sub-models. The proposed VCN approach will be introduced according to these three steps below.

2.1 Sub-datasets sampling

Let $\mathbf{X}$ of size $n \times p$ denote the sample matrix consist of n samples and p variables and $\mathbf{y}$ the class label vector of size $n \times 1$, with elements equal to 1 or -1 in a binary classification case. Assume that the number of MCS is set to N, e.g. 10,000 and the number of variables randomly sampled at each MCS is Q ($Q < p$), e.g. 10, one can perform MCS in the variable space of $\mathbf{X}$ in the following procedure. At each MCS Q out of the p variables are randomly selected, thus obtaining a sub-dataset of size $n \times Q$. Repeating this procedure for N times, altogether N sub-datasets can be drawn. Denote the sampled N sub-datasets as $(\mathbf{X}_{\text{sub}}, \mathbf{y}_{\text{sub}})_i$, $i = 1, 2, 3, \ldots, N$.

2.2 Sub-models building using PLS-LDA

PLS-LDA is powerful in handling highly correlated variables and has seemingly become one of the most commonly used modeling methods in metabolomics. Thus we choose PLS-LDA for modeling in this work. Using each sub-dataset, a PLS-LDA model is constructed. The predictive performance of each model is assessed by misclassification error resulting from fivefold cross validation (Stone 1974). In doing so, altogether N sub-models and their associated N prediction errors are obtained.

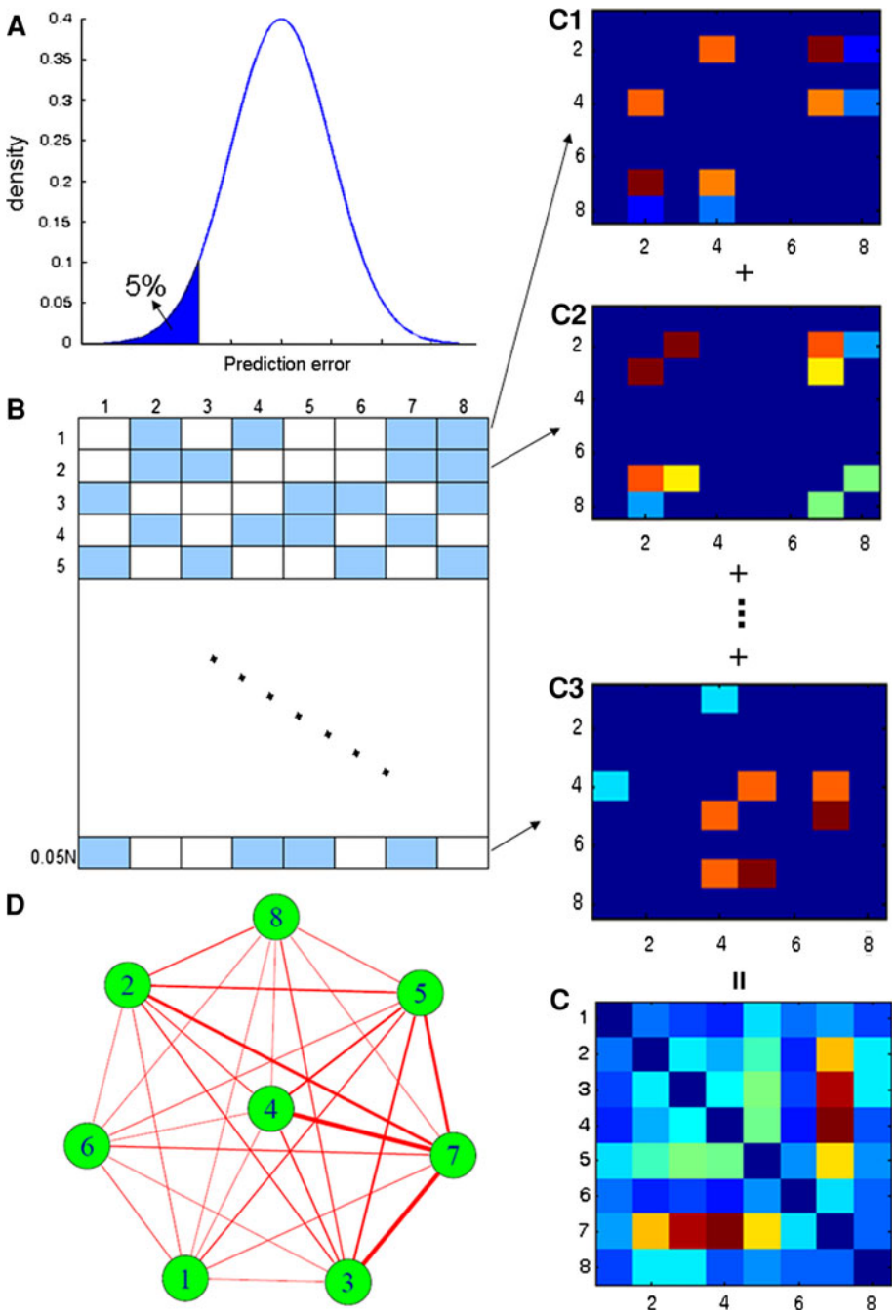

Fig. 1 Illustration of the procedure for computing VCN using a dataset containing eight variables. **a** The distribution of prediction errors of N sub-models and the 5% sub-models associated with the lowest prediction errors marked (*blue area*). **b** The variables included in each of the 5% best sub-models (*filled squares*). The model-wise VCM corresponding to each model are shown as heatmaps in **c1**, **c2**,…, and **c3**, respectively. The total VCM is computed as the sum of all the model-wise VCMs and is shown in **c**. Using the total VCM, a VCN is constructed which is shown in **d** where the width of edges indicates complementary strength between each pair of variables. For example, Variable 4 and Variable 7 strongly complements each other, whereas the complementary strength between Variable 4 and Variable 1 is comparatively rather weak (Color figure online)

2.3 Computation and visualization of VCN

We first sort all the N sub-models according their prediction errors and choose only the best 5% models with the lowest prediction errors for further analysis, which is illustrated in a of Fig. 1. The rationale for using only the best 5% models lies in three folds: (1) 'good models' with low prediction errors are enriched in this part and optimal variable combinations are expected to be included in these "good models" with high frequencies, (2) The probability for 'bad models' with high prediction errors to include an optimal variable combination should be low, and (3) Using only 5% 'good models' can significantly reduce computational cost.

The procedure for computing a VCN is illustrated in Fig. 1 using a dataset of 8 variables and with Q set to 4. Those variables used to build each sub-model is marked as filled squares in b. Note that each sub-model only includes $Q = 4$ variables. Denote the misclassification errors corresponding to the $0.05N$ sub-models as $error^k$, $k = 1, 2,\ldots,$ $0.05N$, and also denote the maximum of these misclassification errors as $error_{\max}$. For the kth PLS-LDA model,

denote the regression coefficient vector $\boldsymbol{\beta}^k$. Then the maximal difference in $|\boldsymbol{\beta}^k|$ is

$$d^k_{\max} = \max(|\beta^k|) - \min(|\beta^k|) \quad (1)$$

where |.| denotes the absolute operator. The difference of absolute regression coefficients of the ith and jth variable (i and j is the original variable index in the range between 1 and p) is calculated as

$$d^k_{ij} = |\beta^k_i| - |\beta^k_j| \quad (2)$$

With these preparations, we define the complementary information between the ith and the jth variable using the following formula:

$$I_{ij} = \sum_{k=1}^{0.05N} \frac{(|\beta^k_i| + |\beta^k_j|)}{2} \cos\left(\frac{d^k_{ij}}{d^k_{\max}} \times \frac{\pi}{2}\right) \frac{error_{\max}}{error^k} \quad (3)$$
$$k = 1, 2, \ldots, 0.05N$$

The first term $\frac{(|\beta^k_i| + |\beta^k_j|)}{2}$ is a measure of the complementary information between two variables (If either the ith or the jth variable is not included in the kth model, $\frac{(|\beta^k_i| + |\beta^k_j|)}{2}$ is manually set to zero.). This is based on our assumption that two variables should have a high complementary information content if both are of large regression coefficients and vice versa. Using only the first term, the complementary information between a predictive variable associated with a large regression coefficient and a noise variable with a small regression coefficient may be large, which does not make sense since these two variables are expected to have little complementary information. Therefore, we introduce the second term $\cos(\frac{d^k_{ij}}{d^k_{\max}} \times \frac{\pi}{2})$ that can reduce the complementary information of two variables that are of a large difference in regression coefficients measured by d^k_{ij}, which might have a shrinkage effect on regression coefficients as ridge regression does. Finally, we introduce a model-wise factor $\frac{error_{\max}}{error^k}$ which adjusts variables' complementary information based on the predictive performance of the kth model. The lower the $error^k$ is, the more complementary information the two variables in the kth model contain. Note that the complementary information between the two variables is a sum from all the $0.05N$ models containing these two variables, which indeed take into account the effects of the other variables included in these $0.05N$ variables. In addition, it also needs to be pointed out that the complementary information between two variables cannot be compared across experiments with different Q values.

Taking the first sub-model in b as an example, the variable complementary information between each pair of the four variables, i.e. I_{24}, I_{27}, I_{28}, I_{47}, I_{48} and I_{78}, can be computed. These computed complementary information values are put into a matrix, called variable complementary matrix (VCM) of size $p \times p$, with missing elements set to zero. This VCM is shown in c1 as a heatmap. In the same way, a VCM can be computed for each sub-model. The VCMs for the second and the last sub-model are shown in c2 and c3, respectively. Next, we sum all the $0.05N$ VCMs to derive the total VCM which is shown in c. It can be found that all the $0.05N$ sub-models can reveal the complementary information between each pair of all the p variables, although each sub-model only involves a subset of Q variables. Finally using the total VCM, we construct a VCN in which each vertex represents a variable and the width of the edge that connects two variables stands for their content of complementary information.

3 Results and discussion

3.1 Simulation study

Perhaps the best way to investigate behaviors of a method is to perform simulation studies. Here we simulate a dataset of 100 samples (50 positive and 50 negative) and 30 variables. The correlation of each variable with the class label vector $\mathbf{Y}$ is shown in Fig. S1. Among these 30 variables, the first two when combined can separate the two classes of samples without any errors and are expected to have a large content of complementary information. The third, fourth and the fifth variable have higher correlations than the first two but any two-variable combination of these three can not correctly separate the two classes, indicating their lower complementary information content. The remaining 25 variables are simulated as random noise. As an example, all possible two-variable combinations of the first 6 variables are shown in Fig. 2.

To run the proposed method, N is set to 20,000. The optimal Q is chosen to be 5 using tenfold cross validation (see details in Fig. S2). Each variable is standardized to have zero mean and unit variance before further analysis. Due to the fact that results from this method cannot be exactly reproduced because of the embedded Monte Carlo technique, we run the program independently 10 times and use the average of the obtained 10 VCMs to construct a VCN. The full VCN is very large and thus only shown in Fig. S3. For the sake of resolution, only a sub-network of the full VCN consisting of 10 variables associated with the highest complementary information content is displayed in Fig. 3. As is shown, the first two variables have the highest information content which is consistent with the simulation that these two variables completely separate the two classes of samples (see Fig. 2). Of note, compared to the first 2 variables, the information content among the third, fourth and fifth variables is lower although they are of higher

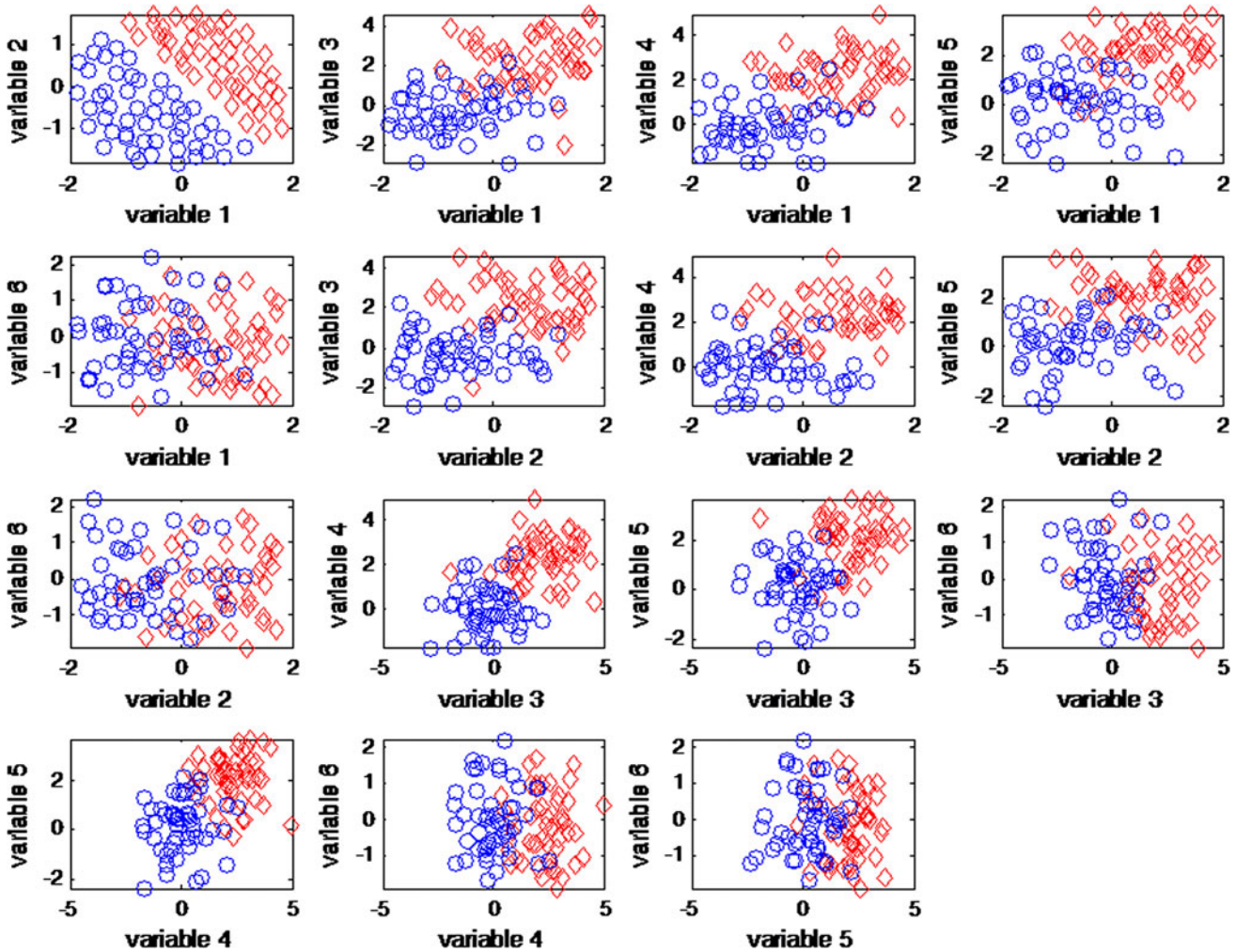

Fig. 2 Plots of the first six variables against each other for the simulated data

correlations, indicating the importance of seeking an optimal variable combination for classification. Interestingly, the five noise variables (index: 14, 15, 20, 23 and 26) possess mutually little complementary information, which makes sense because the simulated noise variables indeed provide no information for class separation.

To measure the overall performance of variables in terms of complementary information, we calculate the total complementary information (TCI), which is the sum of the information content between a given variable and all the other variables based on the full VCN. The TCI of each variable for this simulated dataset is shown in Fig. 3. Clearly, the first two variables stand out, whereas the TCI for the third, fourth and fifth variable is much lower. In contrast, all the noise variables have negligible TCI values, thus not worth considering. In addition, to check whether the cos(.) term and the model-wise term are meaningful or not, we also compute the full VCN with the same setting except for using only $\frac{(|\beta_i^k|+|\beta_j^k|)}{2}$. The TCI of the 30 variables are shown in Fig. S4. Clearly, Variable 3 get the highest TCI value which is not consistent with the simulated data structure. In contrast, Variable 1 and 2 (Fig. S4A) in are well pronounced using Eq. 3, suggesting the necessity of the use of the cos(.) term and the model-wise term. Further, we compared the predictive performances of the two highest ranked variables using VCN and the variable importance in projection (VIP) in PLS (see Fig. S5), respectively. The results are shown in Table 1. It can be found that the VCN procedure indeed better scores the variable importance.

Summing up, the proposed VCN is shown to be promising in the identification of a set of variables that are mutually of high complementary information and also informative in class prediction. It provides a novel approach for the analysis and display of variable inter-relationships.

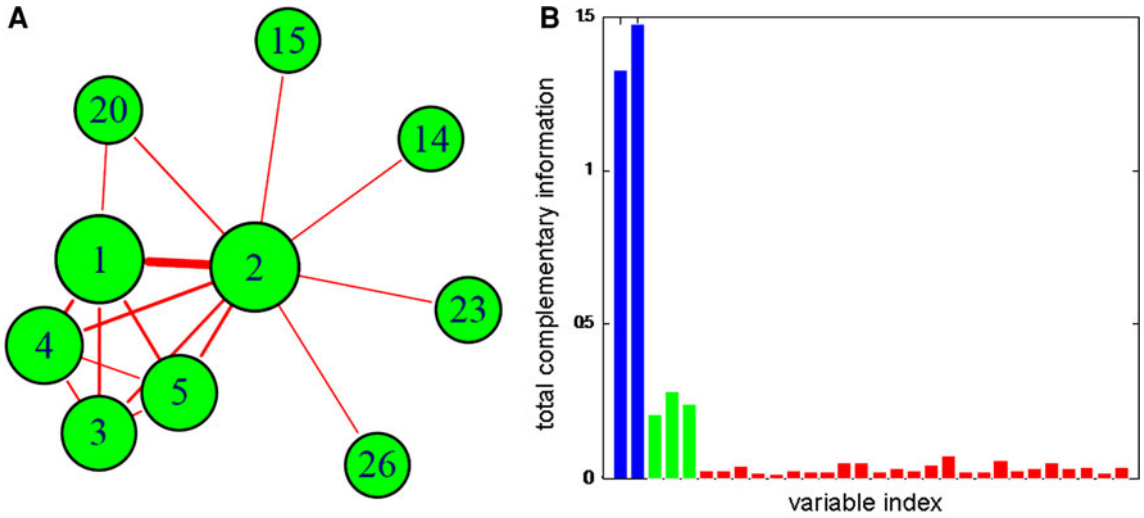

Fig. 3 **a** A sub-VCN consisting of only ten variables with the highest complementary information content. The size of *each vertex* reflects the content of TCI (computed using the full VCN) that is displayed in **b**

Table 1 Ten-fold double cross validation results of PLS-LDA models using all measured metabolites and the identified metabolites using VCN and VIP, respectively

Data	All metabolites/VCN/VIP			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	nLVs
Simulated	97.0/100/94.0	98.0/100/96.0	96.0/100/92.0	1/1/1
T2DM	93.3/96.7/96.7	91.1/95.6/95.6	95.6/97.8/97.8	3/2/1
POCD	79.2/91.7/83.7	83.3/91.7/75.0	83.3/91.7/91.7	4/1/1

nLVs number of latent variables used in PLS, *VCN* the proposed method, *VIP* variable importance projection in PLS

3.2 Type 2 diabetes mellitus data

The overnight fasting plasma samples were collected from 45 T2DM patients and 45 healthy controls from Xiangya Hospital, Changsha city of People's Republic of China. Altogether 21 metabolites were quantified using GC/MS combined with chemometrics resolution methods. Details of this dataset were described in our previous work (Tan et al. 2009).

To begin with, N is set to 20,000. The optimal Q value is 5 determined using tenfold cross validation (see details in Fig. S2). To make regression coefficients of a sub-model comparable, each metabolite is standardized to have zero mean and unit variance. By analogy, we use the average of the 10 VCMs resulting from 10 runs of the proposed procedure to construct the VCN. To have a good resolution of the network, here we only present a sub-VCN (in Fig. 4) composed of ten metabolites with the highest complementary information content (the full VCN is shown in Fig. S3).

As displayed in Fig. 4, the four metabolites, i.e. α-linolenic acid (ALA, C18:3$n-3$), eicosapentaenoic acid (EPA, C20:5$n-3$), oleic acid (OLA, C18:1$n-9$) and C18:1$n-7$, are associated with complementary information content that is significantly higher than that of others, suggesting that the changes of plasma concentration profile of these four metabolites could possibly reflect the process of development of type 2 diabetes. Thus, these four metabolites could be considered as possible biomarker candidates for type 2 diabetes. Of note, ALA is located at a central position in the network, implying its key role in distinguishing type 2 diabetes from healthy controls. Biologically, ALA is an essential fatty acid and serves as the substrate for the in vivo synthesis of EPA which is shown to be closely linked to energy metabolism and insulin sensitivity. It was found by Madigan et al. (2005) that intake of OLA can raise the level of high density lipoprotein while lowering LDL in diabetic patients. Although various aspects of type 2 diabetes mellitus were understood, it seems that our results can provide additional information for the understanding of the disease by looking at how biological variables complement each other. Besides, we also calculate the TCI for each metabolite, which is also shown in b in Fig. 4. From this plot, it can be observed that the TCI contained in ALA, EPA, OLA and C18:1n-7 is much higher than that of the other metabolites. In fact, the TCI can be seen as a measure of metabolites' overall contribution in the discrimination of two classes of samples and hence could work as a potential criterion for biomarker screening.

To intuitively understand whether the metabolites identified above are of value for classification, PCA models are constructed using all the 21 metabolites and only three metabolites (ALA, EPA and OLA) of the highest

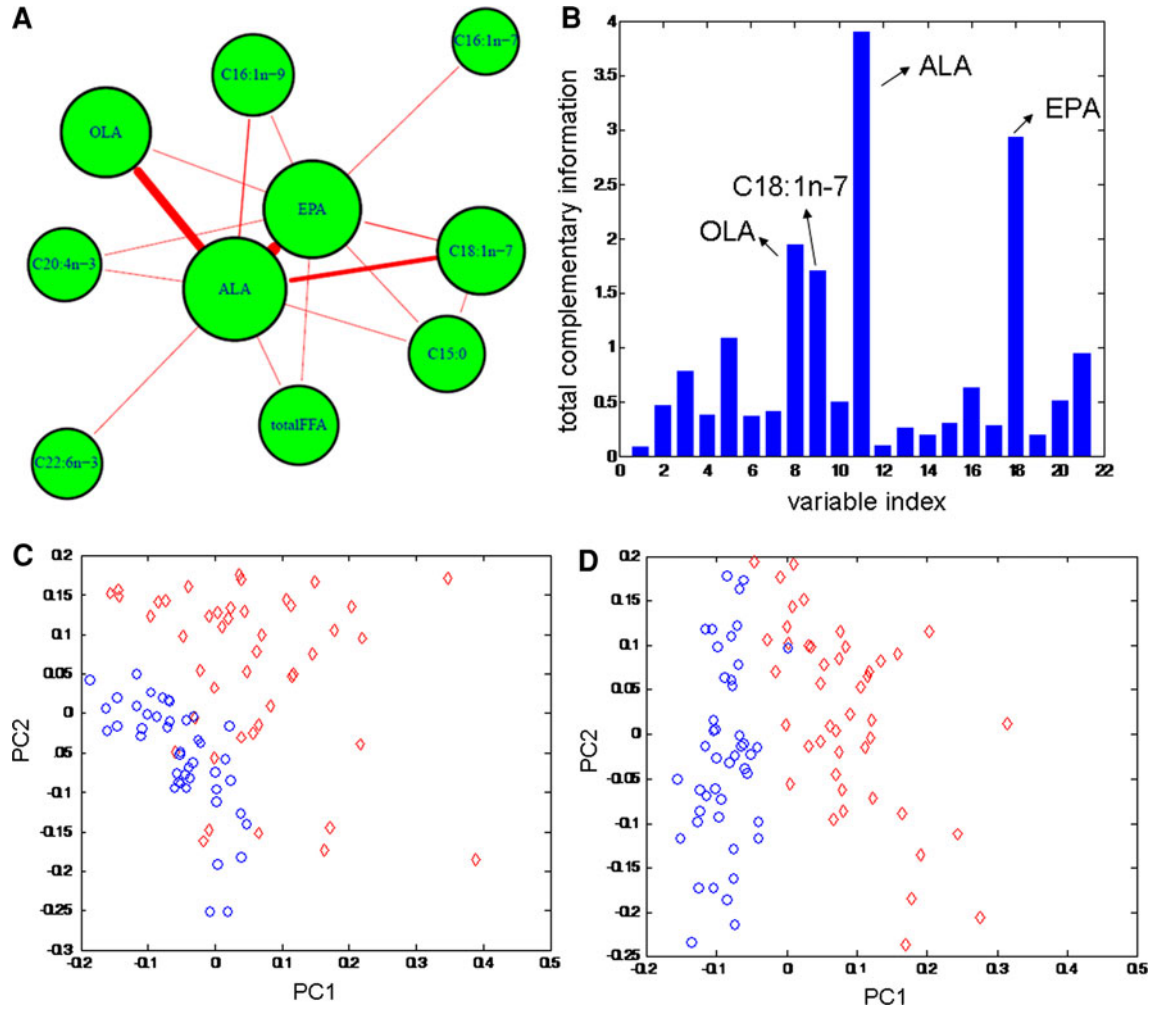

Fig. 4 **a** A sub-VCN consisting of only ten variables with the highest complementary information content for diabetes data. The size of each vertex reflects the content of TCI (computed using the full VCN) that is displayed in **b**. PCA plots using all metabolites as well as ALA, OLA and EPA marked in **b** are given in **c** and **d**, respectively

complementary information content, respectively. Scores plots are shown in Fig. 4. It can found clearly that significantly better separation between classes is achieved. In addition, we also test whether using only the metabolites that are of high complementary information content can improve the predictive performance of a PLS-LDA classification model. Using tenfold double cross validation, a subset of three metabolites, i.e. ALA, EPA and OLA, are determined as the best subset in terms of the lowest prediction error. The overall accuracy, sensitivity and specificity is 96.7, 95.6 and 97.8%, which is significantly improved compared to the PLS-LDA model established using all the metabolites (Table 1). For comparison, the prediction results based on also three metabolites with the highest VIP value (shown in Fig. S5) are presented in Table 1. For this data, the VCN method and the VIP give the same results. However, it should be noted that VCN has an additional advantage that it can reveal complementary information between variables.

3.3 Postoperative cognitive dysfunction data

The serum samples were collected from 12 POCD rats and 12 none-POCD rats after isoflurane anesthesia from Xiangya Hospital, Changsha city of People's Republic of China. Altogether 44 metabolites were quantified using GC/MS combined with chemometrics resolution methods (details of this data have not been published yet).

By analogy with the previous data, each metabolite is standardized to zero mean and unit variance, N is set to 20,000 and the optimal Q value is chosen to be 8 using tenfold cross validation (see details in Fig. S2). The VCN is constructed using the averaged VCMs from 10 runs of the proposed procedure and is shown in Fig. S3. To get a good

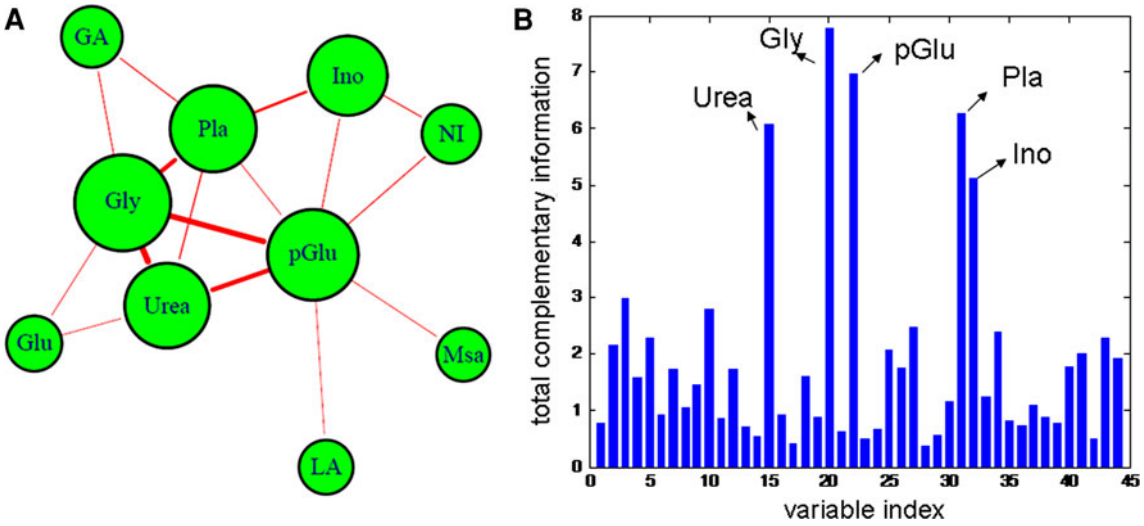

Fig. 5 **a** A sub-VCN consisting of only ten variables with the highest complementary information content for the POCD data. The size of each vertex reflects the content of TCI (computed using the full VCN) that is displayed in **b**

resolution of the network, only a sub-VCN composed of ten metabolites with the highest complementary information content is given here in Fig. 5. The full names of each metabolite are given in Table S1.

As is shown in Fig. 5, the five metabolites, i.e. urea, glycine (Gly), pyroglutamic acid (pGlu), palmitic acid (Pla) and inositol (Ino), are associated with high complementary information content (a) as well as high TCI (see b of Fig. 5) content, suggesting that these five metabolites may be potential biomarkers of which the plasma concentration pattern could be used for the screening of patients with postoperative cognitive dysfunction in clinical practice. Indeed, Gly indentified by our method has been shown to affect synapse function which is associated with cognitive decline (Xie and Tanzi 2006). It was reported that pGlu, also called 5-oxyproline, can inhibit brain energy metabolism (Escobedo and Cravioto 1973), thus maybe leading to its association with POCD. Indeed, the concentration of pGlu of the 12 POCD samples and the 12 none-POCD samples are 0.276 ± 0.053 and 0.219 ± 0.041, which is consistent with previous findings. Besides, it was reported that POCD is linked to Alzheimer's disease (AD). pGlu, Pla and Ino have been shown to be associated with AD (Barak et al. 1996; Patil et al. 2008), thus being indirectly linked to POCD. Of interest, it can be observed from Fig. 5 that pGlu, Gly and urea are mutually tightly associated, indicating their possible joint contributions to the development of POCD. In general, the molecular mechanism of POCD is largely unknown and reports on metabolomics investigation of POCD as well as associated biomarker discovery are quite limited, what we provide here is explorative, aiming at providing some initial metabolic findings that are related to POCD.

We also compared PCA models constructed using all the 44 metabolites and the five metabolites identified above, respectively. The scores plots are displayed in Fig. S6. Apparently, the two classes of samples are much better separated using only five metabolites, suggesting that complementary information is of predictive value. We also compared the predictive performances of the PLS-LDA classification model established using all variables, these 5 variables with high TCI value and the highest ranked 5 variables by VIP score (Fig. S5), respectively. The results based on tenfold double cross validation are shown in Table 1. The overall accuracy, sensitivity and specificity of this selected mode is 91.7, 91.7 and 91.7%, which is significantly better than the model using all 44 metabolites (79.2, 83.3 and 83.3%) and the model selected by VIP (83.7, 75.0 and 91.7%). This comparison demonstrates that the metabolites that possess high variable complementary information content are indeed useful for distinguishing POCD samples from the non-POCD ones and the proposed approach is promising in identifying potential biomarkers associated with the disease analyzed.

4 Conclusions

In the present work, a MPA approach is developed for the quantitative analysis of variable complementary information which has been rarely explored. A VCN is then constructed by using the complementary information to give an overall structure displaying how biological variables complement each other. The performance of the proposed method is first tested using a simulated dataset. The results demonstrated that the proposed VCN can reveal an optimal

combination of variables that collectively shows high predictive performances. When applied to two real world metabolomics datasets on type 2 diabetes and postoperative cognitive dysfunction, it is found that the metabolites displaying high complementary information content are not only biologically interpretable but also are of high predictive performance of the clinical outcome under investigation. We anticipate that variable complementary information will gain much attention and the proposed VCN will find applications in a variety of fields, such as genomics and proteomics.

Acknowledgments This work was financially supported by the National Nature Foundation Committee of People's Republic of China (Grants Nos. 20875104, 21075138 and 21105129) and the Graduate degree thesis Innovation Foundation of Central South University (CX2010B057). The studies meet with the approval of the university's review board. The study is approved by the review board of Central South University.

References

Arsenault, B. J., Boekholdt, S. M., & Kastelein, J. J. P. (2011). Lipid parameters for measuring risk of cardiovascular disease. *Nature Reviews Cardiology, 8*, 197–206.

Barak, Y., Levine, J., Glasman, A., Elizur, A., & Belmaker, R. H. (1996). Inositol treatment of Alzheimer's disease: A double bund, cross-over placebo controlled trial. *Progress in Neuropsychopharmacology and Biological Psychiatry, 20*, 729–735.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*, 166–173.

Beasley, J. E., & Planes, F. J. (2007). Recovering metabolic pathways via optimization. *Bioinformatics, 23*, 92–98.

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., & Griffin, J. L. (2011). Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews, 40*, 387–426.

Escobedo, M., & Cravioto, J. (1973). Studies on the malabsorption syndromes. inhibition of (Na^+–K^+) ATPase of small intestine microvilli by pyrrolidone carboxylic acid. *Clinica Chimica Acta, 49*, 147–151.

Fan, Y., Murphy, T. B., Byrne, J. C., Brennan, L., Fitzpatrick, J. M., & Watson, R. W. G. (2011). Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. *Journal of Proteome Research, 10*, 1361–1373.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*, 389–422.

Holmes, E., Cloarec, O., & Nicholson, J. K. (2006). Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: Application to $HgCl_2$ toxicity. *Journal of Proteome Research, 5*, 1313–1320.

Küffner, R., Zimmer, R., & Lengauer, T. (2000). Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics, 16*, 825–836.

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics, 17*, 1131–1142.

Li, H.-D., Liang, Y.-Z., Xu, Q.-S., & Cao, D.-S. (2009). Model population analysis for variable selection. *Journal of Chemometrics, 24*, 418–423.

Li, H.-D., Zeng, M.-M., Tan, B.-B., Liang, Y.-Z., Xu, Q.-S., & Cao, D.-S. (2010). Recipe for revealing informative metabolites based on model population analysis. *Metabolomics, 6*, 353–361.

Li, H.-D., Liang, Y.-Z., Xu, Q.-S., Cao, D.-S., et al. (2011). Recipe for uncovering predictive genes using support vector machines based on model population analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8*, 1633–1641.

Li, H.-D., Liang, Y.-Z., Xu, Q.-S., & Cao, D.-S. (2012). Model population analysis and its applications in chemical and biological modeling. *Trends in Analytical Chemistry*. doi:10.1016/j.trac.2011.11.007.

Liu, K.-H., & Xu, C.-G. (2009). A genetic programming-based approach to the classification of multiclass microarray datasets. *Bioinformatics, 25*, 331–337.

Madigan, C., Ryan, M., Owens, D., Collins, P., & Tomkin, G. H. (2005). Comparison of diets high in monounsaturated versus polyunsaturated fatty acid on postprandial lipoproteins in diabetes. *Irish Journal of Medical Science, 174*, 8–20.

Patil, S., Balu, D., Melrose, J., & Chan, C. (2008). Brain region-specificity of palmitic acid-induced abnormalities associated with Alzheimer's disease. *BMC Research Notes, 1*, 20.

Rajalahti, T., Arneberg, R., Kroksveen, A. C., Berle, M., Myhr, K.-M., & Kvalheim, O. M. (2009). Discriminating variable test and selectivity ratio plot: Quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical Chemistry, 81*, 2581–2590.

Selman, B. (2008). Computational science: A hard statistical view. *Nature, 451*, 639–640.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society Series B, 36*, 111–147.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of USA, 102*, 15545–15550.

Tan, B.-B., Liang, Y.-Z., Yi, L.-Z., Li, H.-D., et al. (2009). Identification of free fatty acids profiling of type 2 diabetes mellitus and exploring possible biomarkers by GC–MS coupled with chemometrics. *Metabolomics, 6*, 219–228.

Xie, Z., & Tanzi, R. E. (2006). Alzheimer's disease and post-operative cognitive dysfunction. *Experimental Gerontology, 41*, 346–359.

Yi, L.-Z., He, J., Liang, Y.-Z., Yuan, D.-L., & Chau, F.-T. (2006). Plasma fatty acid metabolic profiling and biomarkers of type 2 diabetes mellitus based on GC/MS and PLS-LDA. *FEBS Letters, 580*, 6837–6845.