Pooled sample-based workflow and software for medical data calibration

Tianlu Chen*,
Center for Translational Medicine
Shanghai Jiao Tong University
Affiliated Sixth People’s Hospital,
Shanghai 200233, China
chentianlu@sjtu.edu.cn

Guiyun Tian,
School of Electrical and Electronic
Engineering, Merz Court, Newcastle
University, Newcastle upon Tyne,
NE1 7RU, UK
g.y.tian@ncl.ac.uk

Weiwei Quan
Department of Cardiology
Shanghai Jiao Tong University
Affiliated Rui Jin Hospital, Shanghai
200240, China
springqww@aliyun.com

***Abstract*—Systematic nonbiological variation frequently occurs within and between batches in medical data sets. We developed a seven-step workflow and software to calibrate metabolomic data sets by pooled samples as quality controls (PQCs). The software can adjust both within- and between-batch variance and perform rough, medium, and precise calibrations of the data set. Calibrated data sets can be combined directly with one another and with complementary information from various sources. It is user friendly with an integrated GUI, has lots of options, and is independent to data sources. It has been used successfully in numerous metabolomics studies and is available on request.**

Keywords*—data calibration, data integration, pooled samples, quality control

I. INTRODUCTION

Metabolomics is a powerful tool in systems biology and translational medicine for understanding phenotypes and discovering biomarkers [1, 2]. In long term or large scale mass spectroscopy (MS) based metabolomics studies, it is important to have consistent measurements with minimal systematic nonbiological biases. Unwanted variation in metabolomic data have experimental causes such as faulty sample pretreatment, instrumental drift, and human error [3]. Unwanted variation can also appear when data from different analytical platforms, sample types, and replicates are combined in an attempt to profile the whole metabolome of a living system, or at least quantify the widest possible range of metabolites.

Two types of data combinations can be envisaged. The first type combines data sets representing the same set of objects (e.g., a group of healthy volunteers) but different sets of measured variables derived from different sample types (e.g., serum, urine, and tissue) and/or different analytical techniques. The second type of data combination combines data sets representing different sets of objects (e.g., several groups of healthy volunteers from different laboratories) but the same measured variables. It is believed that the combination of different, or partially overlapping, sets of variables and objects greatly enhance the biological interpretations of the variability present in the study population. Moreover, data combinations are valuable when (a) more than one platforms or laboratories participated in the study, (b) the number of samples is too large to be measured in one batch or in one laboratory, (c) additional samples become available in the course of the study while previously collected samples have already been measured, or (d) additional samples are measured for validation following a successful pilot experiment [4]. Consequently, between-batch variation attracts more and more attention with an attempt to get a well-combined data.

Various approaches have been used to reduce systematic nonbiological error. Some of them may be conducive to data combination. Direct scaling using the total signal intensity, labeled internal standards for every analyte, or a representative Internal Standard tends to suppress the sensitivity of the analyses, leading to a loss of information. Reducing systematic nonbiological error by periodically analyzing pooled samples (PQCs) along with the subject samples is gaining acceptance as a quality control strategy in metabolic profiling [5-8]. Despite their success, PQC-based calibration methods have a limited capacity to adjust batch and injection order effects lacking the pretreatment of the PQC data beforehand. Additionally, the calibration process is laborious and prone to human error.

Here, a workflow and a user-friendly software application, the Metabolomic Data Calibration Tool (MDCT), were developed to pretreat PQCs and calibrate sample data with a variety of options. After pretreatment, PQCs are more appropriate for calibration. Data sets calibrated by the MDCT can be directly combined with one another for further statistical analysis. MDCT was written in Matlab (version 7.1). The software is rapid and platform-independent. The source code and some example data sets are freely available upon request.

II. METHODS

A. *Flowchart of MDCT and PQC pretreatment*

Figure 1a shows the seven-step MDCT flowchart: data import, parameter setting, PQC pretreatment, variable selection, sample data normalization by IS and/or PQC, Box-Cox transform, and result export. The Box-Cox transform is optional which makes the data more normal distribution-like and thus closer to the basic assumption of most statistical approaches.

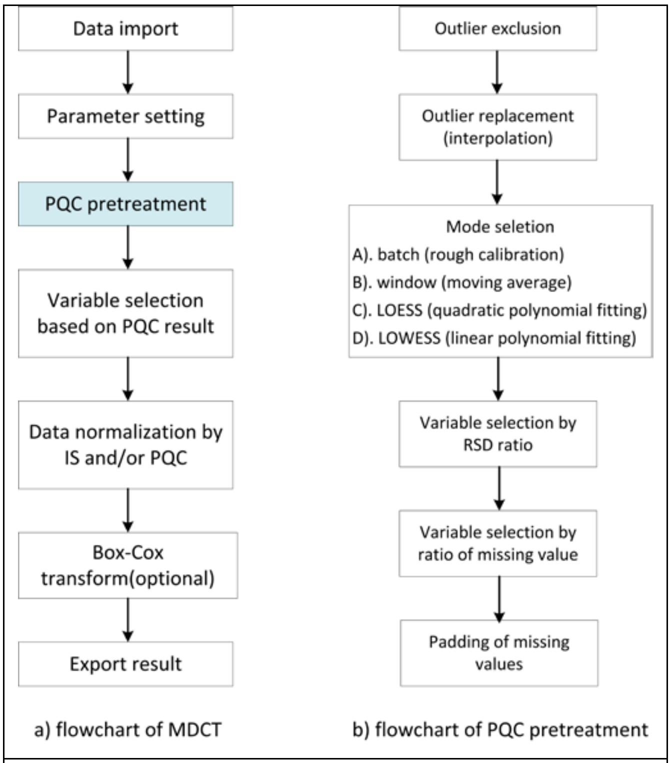

Data import
Parameter setting
PQC pretreatment
Variable selection based on PQC result
Data normalization by IS and/or PQC
Box-Cox transform(optional)
Export result
a) flowchart of MDCT
Outlier exclusion
Outlier replacement (interpolation)
Mode seletion
A). batch (rough calibration)
B). window (moving average)
C). LOESS (quadratic polynomial fitting)
D). LOWESS (linear polynomial fitting)
Variable selection by RSD ratio
Variable selection by ratio of missing value
Padding of missing values
b) flowchart of PQC pretreatment

Figure 1. Flowcharts of the (a) MDCT workflow and the (b) PQC pretreatment

PQCs are pooled samples. Therefore, they contain the same compounds as the subject samples and are supposed to reflect the average metabolite concentrations within a study. PQCs are pretreated according to the same protocols as the subject samples and are evenly injected throughout the analyses. The performances of the pretreatment and the analytical platform can be assessed using the PQCs. Consequently, high-quality PQCs are the solid foundation of a good calibration. The PQC pretreatment is the core step of MDCT.

Figure 1b shows a flowchart of the six-step PQC pretreatment. First, outliers are identified and excluded. This can be done manually or automatically with the aid of one- (run order vs. t1) and two- (t1 vs. t2) dimensional plots of the PCA scores from the PQCs and the samples or from the PQCs only. In automatic mode, PQCs with t1 larger than three times the standard deviation of all the PQCs are considered outliers and are excluded. In manual mode, any PQC can be labeled as an outlier and excluded. The excluded PQCs are replaced by a linear interpolation of all the included PQCs with the assumption that the behavior of PQCs in the same batch is linear. The outlier replacement step is to ensure that the frequency of PQC injection is constant and the calibration is unbiased to all samples.

After the outliers are excluded, the PQCs are smoothed by one of four methods: batch, window, LOESS, or LOWESS. A smoothing parameter p is set in the parameter setting step of the workflow. In batch mode, *p* is the number of batches. The PQCs are divided into *p* groups by the run order. The new PQCs, *PQC_batchi (i=1, 2, …, p)*, are the averages of all the PQCs in the respective groups (eq. 1), where *N* is the number of PQCs, *S* is the number of samples (*S = 10N*, for example), *batchSize* is the number of PQCs in each batch, and *"ceil"* means rounded to the next integer. The batch method is recommended for a rapid and rough calibration, because the number of new PQCs will be reduced substantially as long as *p* is small.

$$PQC\_batch_i = \frac{\sum_{j=batchSize\times(i-1)+1}^{batchSize\times i} PQC_j}{batchSize} \quad (i=1,2,...p)$$
$$batchSize = ceil(N/p) \qquad (1)$$

In window mode, *p* is the window size. After smoothing, *PQC_windowi (i=1, 2, …, N)* is the average of *p* consecutive PQCs, and the step size of the moving window is set to 1 (eq. 2). The number of PQCs does not change in window mode. This mode is recommended for smoothing out short-term fluctuations and highlighting long-term trends.

LOESS and LOWESS (locally weighted scatter plot smoothing) are two closely related non-parametric regression methods that combine multiple regression models in a k-nearest-neighbor-based model [9]. LOESS using linear least-squares fitting and a second-degree polynomial is a generalization of LOWESS, which uses a first-degree polynomial. In the LOESS and LOWESS modes, regression curves (run order vs. intensity) are built for variables (one curve for one variable) to fit the variation of the variable throughout the experiment. We calibrate every sample based on its corresponding PQC (with the same injection index) on the basis of the curves. The chief attraction of the regression methods is that the regressions are based not on all the data but on several localized subsets of the data. This methodology is based on the ideas that any function can be well approximated in a small neighborhood by a low-order polynomial, and that simple models can be fit to data easily. Therefore, LOESS and LOWESS are ideal for complex or random variations where no theoretical models exist. On the other hand, because they rely on data for local fitting, they require fairly large and densely sampled data sets in order to produce good smoothing. The smoothing parameter *p* determines how much of the data is used for local polynomial fitting. Useful values of *p* lie in the range of 0.2 to 0.5 for most LOESS and LOWESS applications. In these modes, every sample has a corresponding PQC according to the run order.

$$PQC\_window_i = \begin{cases} \dfrac{\sum_{j=i}^{i+p-1} PQC_j}{p} & (i=1,2,...n-p+1) \\ PQC_j & (i=n-p+2,...p) \end{cases} \qquad (2)$$

Variables with large percentages of RSDs and/or missing values are thought to be noisy or uninformative. They should be removed to avoid introducing error during the calibration. In the field of metabolomics, an acceptable tolerance of 15-30% RSD is appropriate for MS data. We apply the same criteria in this research. Likewise, variables with up to 30% of the values missing are kept. MDCT users can set the acceptable RSDs and missing values in the parameter setting step (Figure 1a). Missing values are replaced by a value equal to 1/10 the mean of the measured values of the variable.

PQCs are more appropriate for calibration after they are processed. Variables removed from the PQCs during the pretreatment are also removed from the subject samples. Then, the variables in the subject samples can be calibrated by those of the corresponding PQCs (the closest PQC to a given sample or the PQC regression curves) as in Equation 3. *D* is the calibrated data matrix (sorted by run order) and D_{ij} is the data of the jth variable in the ith sample. *R* is the raw data matrix; *C* is the pretreated PQC matrix; *S* is the number of samples; *V* is the number of variables remaining; *Q* is the number of pretreated PQCs (*Q* is equal to *p* for batch mode, and *Q* is the number of raw PQCs in window mode); and *F* is the function of regression curves (run order is the input, and intensity is the output) by LOESS or LOWESS. It is possible to directly combine multiple batches after the PQC calibration because D_{ij} will be a number close to 1 (data sets are in the same or similar scaling space).

$$D_{ij} = \begin{cases} \dfrac{R_{ij}}{C_{kj}}, & for\ batch\ and\ window\ \mathrm{mod}es \\ \dfrac{R_{ij}}{F(run\ order\ of\ ith\ sample)_j}, & for\ LOESS\ and\ LOWESS\ \mathrm{mod}es \end{cases} \qquad (3)$$

$$(i=1,2,...S;\ \ j=1,2,...V;\ \ k=1,2,...Q)$$

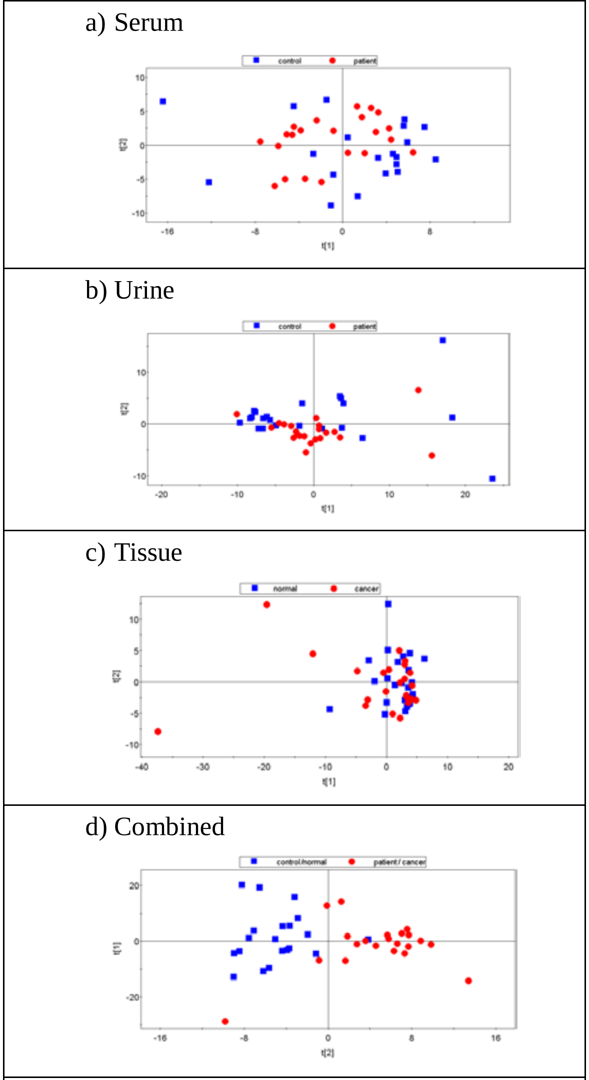
B. Performance evaluation

Unsupervised PCA is a common method in metabolomics of visually exploring the similarity of different data sets or groups. Data that are closely related to each other cluster together in a PCA scores plot. Orthogonal partial least squares (OPLS) models coupled with Variable Importance in the Projection (VIP) values are usually used to separate groups and select important variables. We evaluated the calibration and integration performance of the workflow and MDCT by using the pattern of the PCA scores plots (t1 vs. t2), the three typical OPLS model parameters, and the Euclidean distances of the group centroids before and after processing. The three typical OPLS model parameters are cumulative R^2X, R^2Y, and Q^2Y. $\mathrm{R}^2\mathrm{X}$ and $\mathrm{R}^2\mathrm{Y}$ represent the fraction of the variance of the data *X* and the group instructor variable *Y* explained by the model (square of percentage of explanation). Q^2Y estimates the predictive performance of the model (square of percentage of prediction). The values of the three parameters vary from 0 to 1 (larger values are better).

All the evaluations were completed using matlab 7.1 and SIMCA-P+ 13.0.

III. Results and Discussion

We tested the performance of the workflow and MDCT using metabolomic data sets derived from multiple types of samples. The gastric cancer data set came from serum and urine samples of 23 patients with gastric cancer and 23 age- and gender-matched healthy volunteers, as well as 23 pairs of cancerous and adjacent normal tissues. We collected and analyzed the samples by GC-TOF/MS (Leco Corporation, St. Joseph, MI) and UPLC-QTOF/MS (Waters, Manchester, U.K., positive and negative ion modes) platforms using protocols established in our lab [1, 17, 18]. PQCs were inserted between every 10 or 15 samples. This data set comprised three separate sample batches with 46 samples each and 200, 185, and 223 annotated metabolites, respectively.


a) Serum

b) Urine

c) Tissue

d) Combined

Figure 2. Two-dimensional PCA scores plots of the calibrated serum (a), urine (b), tissue (c), and serum, urine, and tissue combined (d) data

The blue boxes represent the healthy controls or the normal tissues, and the red dots represent the patients with cancer or the cancerous tissues.

Figure 2 shows the PCA scores plots of the data of the serum (2a), tissue (2b), urine (2c), and serum, tissue, and urine combined (2d). The blue boxes represent the healthy controls or the normal tissues, and the red dots represent the patients with cancer or the cancerous tissues. We could not separate the groups using the data from the serum, urine, or tissue alone. The distances of the group centroids were 2.59, 1.67, and 3.28, respectively. When we calibrated and combined the three data sets, the distance of the group centroid increased to 9.96, which is 3 to 6 times that of the single batches. Thus, we could discriminate between the two groups based on PC2 [t(2) in Figure 2d] using the combined data set.

IV. Conclusions

Systematic nonbiological variance within and between batches is unwanted but often unavoidable in medical especially metabolomic data processing. We presented a PQC based workflow for calibrating metabolomic data sets. They were evaluated comprehensively using three clinical data sets coupled with various criteria. PQCs pretreated for outlier exclusion and replacement, compression or smoothing, variable selection, and missing value padding produced more reproducible results, suggesting that the pretreatments improved the subsequent calibration and analyses. Data sets calibrated using PQCs can be directly combined, because all of the data values are close to 1 (i.e., in the same scale space). Statistical analysis and biomedical discussion based on calibrated and combined data sets are more powerful and convincing because of the larger sample sizes and complementary information from multiple sources.

The MDCT is a data-source independent software that can inspect raw data, pretreat PQCs, and calibrate on rough, medium, and precise scales. It is user friendly with an integrated GUI (graphical user interface) and lots of optionsand has been successfully applied to several clinical metabolomic data sets. We are currently integrating the MDCT into our automated data analysis pipeline [10, 11].

Acknowledgment

Supported by the European Marie Curie Actions (CoNHealth 294927), the National Natural Science Foundation of China (Grant no. 81170760) and Drug Innovation Program of National Science and Technology Project (no. 2011ZX09307-001-02).

References

[1] Nicholson, J.K. and J.C. Lindon, *Systems biology: Metabonomics.* Nature, 2008. **455**(7216): p. 1054-6.

[2] Viant, M.R. and U. Sommer, *Mass spectrometry based environmental metabolomics: a primer and review.* Metabolomics, 2013. **9**(1): p. 144-158.

[3] Livera, A.M.D., et al., *Normalizing and Integrating Metabolomics Data.* Analytical Chemistry, 2012. **84**: p. 10768−10776.

[4] Draisma, H.H.M., et al., *Equating, or Correction for Between-Block Effects with Application to Body Fluid LC-MS and NMR Metabolomics Data Sets.* Analytical Chemistry, 2010. **82**(3): p. 1039–1046

[5] Wang, S.-Y., C.-H. Kuo, and Y.J. Tsen, *Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Methods.* Analytical Chemistry, 2012. **85**: p. 1037-1046.

[6] Kamleh, M.A., et al., *Optimizing the Use of Quality Control Samples for Signal Drift Correction in Large-Scale Urine Metabolic Profiling Studies.* Analytical Chemistry, 2012. **84**: p. 2670−2677.

[7] Zelena, E., et al., *Development of a Robust and Repeatable UPLC-MS Method for the Long-Term Metabolomic Study of Human Serum.* Analytical Chemistry, 2009. **81**(4): p. 1357–1364.

[8] Kloet, F.M.v.d., et al., *Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping.* Journal of Proteome Research, 2009. **8**: p. 5132–5141.

[9] Cleveland, W.S. and S.J. Devlin, *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting.* Journal of the American Statistical Association, 1988. **83**: p. 596-610.

[10] Ni, Y., et al., *ADAP-GC 2.0: Deconvolution of Coeluting Metabolites from GC/TOFMS Data for Metabolomics Studies.* Analytical Chemistry, 2012. **84**: p. 6619−6629.

[11] Jiang, W., et al., *An Automated Data Analysis Pipeline for GC-TOF-MS Metabonomics Studies.* Journal of Proteome Research, 2010. **9**: p. 5974–5981.