Normalization of metabolomics data using multiple internal standards

Matej Orešič[1]

[1] VTT Technical Research Centre of Finland, Tietotie 2,
FIN-02044 Espoo, Finland
matej.oresic@vtt.fi

Abstract. Success of metabolomics as the phenotyping platform largely depends on its ability to detect various sources of biological variability. Removal of platform-specific sources of variability such as systematic error is therefore one of the foremost priorities in data pre-processing. However, chemical diversity of molecular species included in typical metabolic profiling experiments leads to different responses to variations in experimental conditions, making normalization a very demanding task. We present an approach that utilizes variability information from multiple internal standard compounds to find optimal normalization factor for each individual molecular species detected by metabolomics approach. The method is demonstrated on mouse liver lipidomic profiles using Ultra Performance Liquid Chromatography coupled to high resolution mass spectrometry. We compared its performance to two commonly utilized normalization methods: normalization by l_2 vector norm and by retention time region specific standard compound profiles. Our approach proved superior in its ability to reduce the effect of systematic error across the full spectrum of metabolite peaks.

Keywords: Metabolomics, mass spectrometry, normalization, maximum likelihood, calibration

1 Introduction

Metabolomics is a discipline dedicated to the global study of metabolites, their dynamics, composition, interactions, and responses to interventions or to changes in their environment, in cells, tissues, and biofluids [1]. Concentration changes of specific groups of metabolites may be descriptive of systems responses to environmental or genetic interventions, and their study may therefore be a powerful tool for characterization of complex phenotypes [2, 3] as well as for development of biomarkers for specific physiological responses [4, 5].

Study of the variability of metabolites in different states of biological systems is therefore an important task of systems biology. As we are primarily interested in systems responses resulting in metabolite level regulation as related to diverse genetic or environmental changes, it is important to separate such *interesting* biological variation from *obscuring* sources of variability introduced in experimental studies of

metabolites. Since multiple experimental platforms are commonly applied in the study of metabolites [6], the sources of the obscuring variation are many and platform specific. Such sources include variability rising from inhomogeneity of samples, their lability and inevitable minor differences in sample preparation. In mass spectrometry based detection, the sources include the variations in the ion source as well as matrix specific effects such as ion suppression. Following the measurement, the data pre-processing steps such as peak detection, peak integration and alignment may introduce an additional error.

Removal of non-biological sources of variability introduced by specific platforms or instrumentation and determining true concentrations of metabolites in biological samples is one of the foremost priorities in metabolomics data pre-processing. Chemical diversity of metabolites, leading for example to different recoveries during extraction and responses during ionization in mass spectrometer, makes this a formidable task. Commonly, quantitative analytical methods have relied on utilization of isotope labeled internal standard for each metabolite measured. However, in broad profiling approach this is prohibitive, since the number of metabolites is too large and many of them may not even be known. Currently applied approaches for normalization can be divided into two major categories: (1) Statistical models used to derive optimal scaling factors for each sample based on complete dataset, such as normalization by unit norm [7] or median [8] of intensities, or the maximum likelihood method [3] adopted from the approach developed for gene expression data [9], and (2) adoption of single or multiple internal standards based on empirical rules, such as specific regions of retention time [10]. The lack of an absolute concentration reference is a weakness of the former approach, while the latter does not take into account diversity of compounds in the biological matrix and is too dependent on quality of the individual standard profiles. In this paper we present a method that combines the two approaches by deriving a model based on covariance of multiple internal standards.

2 Derivation of the normalization model

The un-normalized metabolomics data resulting from first stages of pre-processing, usually including peak detection and alignment, can be represented by a matrix of N variables (metabolite peaks) and M objects (samples). For example, in liquid chromatography mass spectrometry (LC/MS) based profiling; each peak is represented by mass to charge ratio (m/z) and retention time (rt).

In the rest of the text we will use the following notation:

- i parameterizes peaks: $i \to \{\mathrm{m/z}, \mathrm{rt}\}$ and $i = 1 \ldots N$.
- s parameterizes peaks from internal standard compounds: $s \to \{\mathrm{m/z}, \mathrm{rt}\}$ and $s = 1 \ldots S$.
- j parameterizes experiment runs: $j = 1 \ldots M$.

- Intensity matrix for all peaks: $\mathbf{X} = \{X_{ij}\}$.
- Intensity matrix for all internal standard peaks: $\mathbf{Z} = \{Z_{sj}\}$.

Most of the errors mentioned so far are intensity (or metabolite concentration) dependent Therefore, it is reasonable to assume that the true metabolite levels are modified by a multiplicative correction factor. Formally:

$$X_{ij} = m_i \times r_{ij}\left(\{Z_{sj}\}\right) \times e_{ij}\ , \tag{1}$$

where m_i is the intensity independent of the run (i.e. the real intensity value), r_{ij} is the correction factor, and e_{ij} is the random error. The basic premise of our approach is that the systematic variation in each individual metabolite $X_{i.}$ can be modelled as a function of variation of standard compounds. Based on this assumption, the correction factors r_{ij} can be determined from the profiles of standard compounds.

Due to the multiplicative error model it is more appropriate to work in logarithmic space:

$$\log \mathbf{X} = \mathbf{Y}, \log \mathbf{Z} = \mathbf{\Omega}, \log \mathbf{m} = \boldsymbol{\mu}, \log \mathbf{r} = \boldsymbol{\rho}, \log \mathbf{e} = \boldsymbol{\varepsilon} \tag{2}$$

where the model is additive:

$$Y_{ij} = \mu_i + \rho_{ij}\left(\{\Omega_{sj}\}\right) + \varepsilon_{ij}\ . \tag{3}$$

We assume the random error $\boldsymbol{\varepsilon}$ is Gaussian with zero mean and independent variables and parametrize $\boldsymbol{\rho}$ as a linear function of internal standard variation:

$$\rho_{ij} = \sum_s \beta_{is}\left(\Omega_{sj} - \langle \Omega_{s.} \rangle\right), \tag{4}$$

where the average $\langle\ \rangle$ is taken over the samples $j = 1 \ldots M$. It is the parameters $\boldsymbol{\beta}$ that control how the variability of internal standard intensities will affect the variability of intensities of other metabolite peaks.

It is clear from the above equations that Y_{ij} is normally distributed, so parameters μ and $\boldsymbol{\beta}$ can be calculated under assumption of normality using the maximum

likelihood estimate. Omitting the straightforward derivation, maximizing the (log)likelihood of observing the data leads to the following solutions:

$$\mu_i = \langle Y_{i.} \rangle \tag{5}$$

and

$$\boldsymbol{\beta} = \Sigma \times \hat{\Sigma}^{-1}, \tag{6}$$

where

$$\Sigma_{is} = \sum_j \left(Y_{ij} - \langle Y_{i.} \rangle \right) \left(\Omega_{sj} - \langle \Omega_{s.} \rangle \right) \tag{7}$$

correlates internal standards and other peaks, while

$$\hat{\Sigma}_{st} = \sum_j \left(\Omega_{sj} - \langle \Omega_{s.} \rangle \right) \left(\Omega_{tj} - \langle \Omega_{t.} \rangle \right) \tag{8}$$

is covariance matrix for internal standards. The normalized intensities for each peak can be calculated as

$$\widetilde{X}_{ij} = X_{ij} \times \exp\left(- \sum_s \beta_{is} \left(\Omega_{sj} - \langle \Omega_{s.} \rangle \right) \right), \tag{9}$$

where $\boldsymbol{\Omega}$ is obtained from profiles of identified internal standards found in spectra, while $\boldsymbol{\beta}$ is then calculated from Equation **(6)**.

Since the matrix $\boldsymbol{\beta}$ relates the variability of each individual metabolite in biological matrix with that of internal standards for a specific platform and biological matrix, it is possible that the parameters $\boldsymbol{\beta}$ are obtained from a separate repeatability experiment involving large number of repeated measurements. This may often be desirable due to large number of normalization parameters ($N \times S$) to be determined by the method. The correction factors from Equation (9) in a real biological application then include the matrix $\boldsymbol{\beta}$ obtained independently and the measured levels of internal standards $\{\Omega_{sj}\}$ from the biological experiment.

3 Application of the method

In order to evaluate the performance of the method, we performed lipidomics analysis of mouse liver with Ultra Performance Liquid Chromatography coupled to high resolution mass spectrometry with QTof Premier instrument (Waters, Inc.). We run 16 replicates of the same biological sample, corresponding to 3 different extracts of 10, 3, and 3 sample injections each, respectively. Total 6 internal standard compounds were utilized of distinct chemical and functional characteristics, and 1470 monoisotopic peaks corresponding to lipid molecular species were detected using MZmine [11] software.

The method was compared to two commonly utilized methods. The first approach was normalization by l_2 vector norm [7] (abbreviated as L2N). The second method utilized three internal standards assigned to peaks for normalization based on retention time [10] (abbreviated as *3STD*).

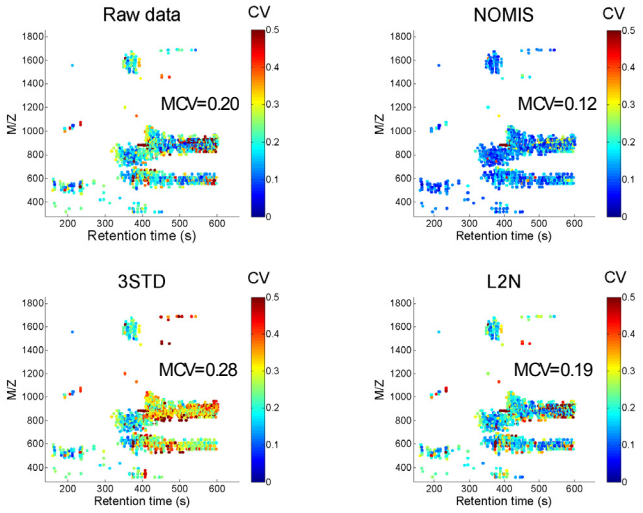

Fig. 1. Liver metabolomics 16-sample repeatability experiment. Coefficients of variance for raw data and three normalization methods: NOMIS (introduced in this paper), L2N, and 3STD. The median coefficient of variance is also shown (MCV).

The performance of our normalization method (abbreviated as NOMIS) on liver lipidomics dataset, as compared to the raw data and to L2N and 3STD methods, is shown in Fig. 1. Along with median coefficient of variance over all peaks (MCV), the coefficients of variance for all peaks in two-dimensional spectral representation (mass-to-charge *vs*. retention time) are shown using the color bar. It is evident NOMIS method is superior in its ability to reduce the variability due to systematic error across the whole spectrum. In contrast, the 3STD method performed particularly poorly for higher values of retention time. This was due to variable internal standard utilized for normalization in that specific retention time region.

4 Conclusions

In this paper we introduced a new approach to normalize metabolomics data using multiple internal standards. Compared to currently applied approaches, it combines the commonly utilized multiple internal standard approach with the optimal selection of normalization parameters for each individual metabolite based on the standard profiles. The method proved superior to two other commonly utilized normalization strategies in its ability to reduce variability across the full spectrum of metabolites. While we demonstrated the method on LC/MS based approach, we believe the same strategy can be applied to other analytical platforms used in metabolomics, as well as to other levels of molecular profiling such as mass spectrometry based proteomics.

Acknowledgments. We thank Tuulikki Seppänen-Laakso for performing the metabolomics experiment utilized in this paper.

References

1. Oresic, M., A.J. Vidal-Puig, and V. Hänninen, *Metabolomic approaches to phenotype characterization and applications to complex diseases.* Expert Rev. Mol. Diag., 2006. **6**(4) [in press].
2. Raamsdonk, L.M., et al., *A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.* Nat. Biotechnol., 2001. **19**(1): p. 45-50.
3. Oresic, M., et al., *Phenotype characterization using integrated gene transcript, protein and metabolite profiling.* Appl. Bioinformatics, 2004. **3**: p. 205-217.
4. Pauling, L., et al., *Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography.* Proc. Nat. Acad. Sci. U S A, 1971. **68**: p. 2374-2376.
5. Clayton, A.T., et al., *Pharmaco-metabonomic phenotyping and personalized drug treatment.* Nature, 2006. **440**(7087): p. 1073-1077.
6. van der Greef, J., P. Stroobant, and R.v.d. Heijden, *The role of analytical sciences in medical systems biology.* Curr. Opin. Chem. Biol., 2004. **8**(5): p. 559-565.
7. Scholz, M., et al., *Metabolite fingerprinting: detecting biological features by independent component analysis.* Bioinformatics, 2004. **20**(15): p. 2447-2454.
8. Wang, W., et al., *Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards.* Anal. Chem., 2003. **75**: p. 4818 - 4826.
9. Hartemink, A.J., et al., *Maximum likelihood estimation of optimal scaling factors for expression array normalization*, in *Microarrays: optical technologies and informatics. Proceedings of SPIE (vol. 4266)*. M. Bittner, Y. Chen, and A. Dorsel, Editors. 2001. p. 132-140.
10. Bijlsma, S., et al., *Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation.* Anal. Chem., 2006. **78**(2): p. 567-574.
11. Katajamaa, M., J. Miettinen, and M. Oresic, *MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data.* Bioinformatics, 2006. **22**(5): p. 634-636.