Single sample enrichment analysis for mass spectrometry-based omics data

Hiroyuki Yamamoto

h.yama2396@gmail.com

Japan Computational Mass Spectrometry (JCompMS) group

Abstract

In omics studies using mass spectrometry, including metabolomics, it is challenging to detect all metabolites, leading to potential biases when performing traditional enrichment analyses. In this study, we applied a single sample enrichment analysis to the metabolome data of fasting mice. This method, distinct from conventional approaches, utilizes information from both detected and undetected metabolites. It is particularly useful for omics data from mass spectrometry where it is difficult to comprehensively capture all metabolites and where there are many missing values.

Keyword single sample enrichment analysis, missing value, metabolomics

質量分析オミックスデータのための
シングルサンプルエンリッチメント解析

質量分析インフォマティクス研究会　山本　博之　h.yama2396@gmail.com

Abstract メタボロミクスをはじめとする質量分析を用いたオミックス研究では、全ての代謝物を検出することが難しく、従来のエンリッチメント解析を行う際に偏りが生じる可能性がある。そこで本研究では、シングルサンプルエンリッチメント解析を用いて、絶食マウスのメタボロームデータに適用した。本手法は、検出された代謝物と検出されなかった代謝物の情報を利用している点が従来とは異なり、全ての代謝物を完全に網羅することが難しく欠損値が多い質量分析を用いたオミックスデータに対して特に有用である。

Introduction

GSVA (Gene Set Variation Analysis)[1]をはじめとする Single sample enrichment analysis (ssEA)[2]では、それぞれのサンプルに対して遺伝子発現量の大きさで並び替えた後、その値の大きさに対してエンリッチメント解析を行い、得られたエンリッチメントスコアを行列データとし、その後、統計的仮説検定や多変量解析、機械学習などの各種解析を行う。一方メタボロミクスでは、面積値もしくは相対面積値がデータとして利用されることが多く、その値は代謝物の濃度と質量分析の感度の両方に依存し、各物質の濃度の大小関係を直接表しているとは言えないことから、その値の大きさで並べることは必ずしも適切とは言えない。

さらに、メタボロミクスをはじめとする質量分析を用いたオミックス研究では、遺伝子発現とは異なり、全ての物質を網羅的に解析することは出来ない。その結果、検出された代謝物が多く含まれる代謝物セット(代謝パスウェイなど)と少数しか含まれない代謝物セットが混在し、実際にエンリッチメント解析を行った際に、各代謝物セットに対する検出物質数のバイアスを考慮して結果を解釈する必要があるといった課題が残されていた。

そこで本研究では、新しい Single sample enrichment analysis として、サンプル毎に検出された物質とそれ以外を 2 値として Over representation analysis(ORA)[3]を行った。これにより、検出された物質が多い代謝物セットを大きく、検出された物質が少ない代謝物セットを小さく表すスコアをデータとし、代謝物セットの偏りをスコアに組み込むことで、各代謝物セットに対する検出物質数の偏りの問題を回避出来る可能性がある。実際にメタボロームデータに本手法を適用し、従来法[4]と比較した結果について紹介する。

Methods

まず初めに、代謝パスウェイなどの代謝物セットに対して、サンプル毎にエンリッチメント解析を行う。具体的には、代謝物が検出されたかどうか、代謝物セットに含まれるかどうか、の 2 つの項目に対して該当する代謝物の数をカウントし、2×2 のクロス集計表を作成

し、フィッシャーの正確確率検定を行う。例えば、解糖系の代謝物群が多く検出され、かつ解糖系以外の代謝物群について検出された代謝物が少ない場合は、p 値が小さくなる。本研究では、p 値の対数を取り、マイナスの符号を付けたものを ssE スコアとする。これにより、特定の代謝物セットに含まれている代謝物が多く検出されていればスコアの値は高くなり、逆に特定の代謝物セットに含まれる代謝物のうち検出されたものが少なければ、スコアの値が小さくなる統計量となる。このスコアをサンプル毎に計算し、この得られたスコアを統計的仮説検定、多変量解析、機械学習などの各種解析に利用することが出来る。

本手法は、検出された代謝物と検出されなかった代謝物を選定する基準に影響を受ける。本研究で用いたデータは信号強度が S/N=3 未満の場合に検出されなかったと定義しているが、S/N の情報を利用できる場合には、閾値として S/N を他の値に設定することも出来る。また S/N が利用できない場合でも、代謝物毎に信号強度の閾値を設定し、閾値以下の値を検出出来なかった代謝物として設定して解析を行うことも出来る。

実際の計算は全て R を用いて行い、mseapca パッケージとして公開している（ https://github.com/hiroyukiyamamoto/mseapca ）。主成分分析には loadings パッケージ（ https://cran.r-project.org/web/packages/loadings/ ）を用いた。

Results and Discussion

通常状態で飼育したマウスと、12 時間絶食条件下で飼育したマウスの肝臓のメタボロームデータ[ref]を用いてサンプル毎に ssE スコアを計算し、ssE スコアを用いて主成分分析を行った結果を図 1 に示す。

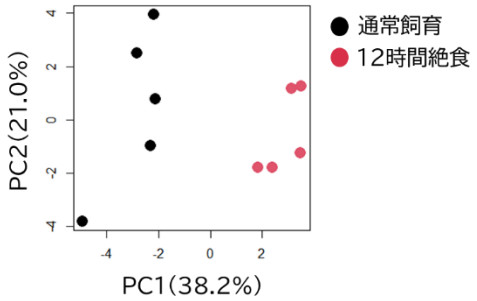

図 1. 絶食マウスでの ssE スコアに対する主成分分析

第 1 主成分スコアで 2 群間が分かれていることから、第 1 主成分負荷量から 2 群で差のある代謝パスウェイを確認した。第 1 主成分負荷量の p 値に対して Benjamini-Hochberg 法による補正を行って得られた q 値が 0.05 未満の条件を満たす代謝パスウェイは、Glycolysis (R=-0.910、p=0.000257、q=0.00334)、Pentose phosphate pathway (R=-0.786、p=0.00707、q=0.0368)、Alanine, aspartic acid and asparagine metabolism (R=-0.850、p=0.00186、q=0.0121)、Valine, leucine and isoleucine metabolism (R=-0.862、p=0.00134、q=0.0117)、Taurine metabolism (R=-0.917、p=0.000187、q=0.00334)となった。従来の方法である、

統計的に有意な代謝物を用いて ORA によるエンリッチメント解析を行った結果と比較しても、12 時間の絶食によって、Glycolysis、Pentose phosphate pathway は統計的に有意に低下しており、従来の解析と部分的に同様の結果が得られることが確認された。

Conclusion

本研究では、シングルサンプルエンリッチメント解析を絶食マウスのメタボロームデータに適用した。解析結果より、従来の解析と部分的に同様の結果が得られることが確認された。また本手法は、検出された代謝物と検出されなかった代謝物の情報を利用している点で、従来とは異なるエンリッチメント解析手法であり、全ての代謝物を完全に網羅することが難しく欠損値が多い質量分析を用いたオミックスデータに対して、特に有用であると考えられる。

References

[1] Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).

[2] Wieder, C., Lai, R.P.J. & Ebbels, T.M.D. Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics* 23, 481 (2022).

[3] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA, Global function profiling of gene expression. *Genomics*. 2003, 81: 98-104.

[4] Yamamoto, H., Fujimori, T., Sato, H. *et al.* Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinformatics* **15**, 51 (2014).