Over-Representation Analysis for Metabolite Set Enrichment Analysis Considering Undetected Metabolites

Hiroyuki Yamamoto

h.yama2396@gmail.com

Japan Computational Mass Spectrometry (JCompMS) group

Abstract

Over-representation analysis (ORA) is widely used to identify significant pathways in metabolomic data. However, traditional ORA approaches, such as those implemented in MetaboAnalyst, do not account for undetected metabolites, potentially resulting in significant biases since undetected metabolites are automatically classified as non-significant. In this study, we used fasting mouse metabolomic data and developed a novel ORA method that leverages information from detected significant metabolites to estimate the possible range of p-values by considering all possible significance combinations among undetected metabolites. Furthermore, we introduced two probabilistic models—a binomial distribution to estimate the number of significant undetected metabolites and a beta distribution to model their proportion—resulting in narrower p-value ranges. Finally, a hierarchical Bayesian model utilizing shared information across all pathways was applied, with resampling from the specified distributions to calculate ORA p-values and achieve the narrowest p-value confidence intervals. This approach highlights the importance of accounting for undetected metabolites in pathway enrichment analysis to enhance the reliability of ORA results.

Keyword metabolite set enrichment analysis, missing value, metabolomics

検出されない代謝物を考慮した Metabolite Set Enrichment Analysis のための Over-Representation Analysis

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

Over-Representation Analysis（ORA）は、メタボロミクスデータにおける有意な経路を特定するために広く使用されている。しかし、従来の ORA アプローチでは、未検出の代謝物が自動的に非有意な代謝物として扱われるため、潜在的なバイアスが生じる可能性がある。本研究では、絶食マウス由来のメタボロミクスデータを用い、検出された有意代謝物の情報を活用し、未検出代謝物を含めた場合の p 値範囲を全組み合わせから推定する新たな ORA 手法を提案した。さらに、未検出代謝物の有意数を二項分布に従うと仮定するモデルと、その割合をベータ分布でモデル化する手法を導入し、狭い p 値範囲を実現した。最後に、全経路に共通の情報を用いる階層ベイズモデルを適用し、設定した分布からリサンプリングして ORA の p 値を算出し、最も狭い p 値信頼区間を達成した。この結果から、未検出代謝物を考慮することの重要性が示された。

Introduction

Over-Representation Analysis（ORA）は、メタボロミクスデータに基づく Metabolite Set Enrichment Analysis（MSEA)において、有意な経路を特定するために広く使用されている手法である[1]。ORA は、観測された代謝物セットが各経路において偶然に見られる頻度を超えているかを評価し、経路の有意性を p 値に基づいて判定するために用いられるが、従来の ORA 手法（例：MetaboAnalyst[2]）では未検出の代謝物が非有意と仮定されているため、解析結果にバイアスが生じる可能性がある。この仮定は、未検出代謝物が多いデータセットにおいて、経路の有意性を過小評価する原因となる。

未検出代謝物が生物学的に重要である可能性もあることから、検出された代謝物のみを考慮することで経路の有意性をより精度よく評価できると期待される。また、検出された代謝物に関する有意な情報を活用することで、未検出代謝物の潜在的な有意性を推定することが可能である。しかし、これまでの研究では、検出と未検出の代謝物を統合的に考慮する ORA の手法は限定的であり、未検出代謝物がもたらす不確実性を明確に扱う方法は確立されていない。

本研究では、絶食マウス由来のメタボロミクスデータを用い、未検出代謝物を考慮した新たな ORA 手法を提案する。まず、絶食マウスのデータから得られた Principal Component Analysis（PCA）の loading を用いて有意代謝物を特定し、未検出代謝物が有意か非有意かの全ての組み合わせを考慮して p 値の最小と最大の範囲を算出した。その後、二項分布とベータ分布の確率モデルを導入し、観測データに基づく確率的推定により p 値範囲をさら

に狭めた。最後に、階層ベイズモデルを用いて全経路に共通する情報を統合し、設定した分布からリサンプリングすることで p 値を算出し、最も狭い信頼区間を達成することに成功した。

Materials and Methods

本研究では、絶食状態のマウスから得られたメタボロミクスデータを用いて解析を行った。このデータは、代謝物を検出と未検出に分類し、検出された代謝物を基に有意性を評価した。具体的には、主成分分析（PCA）を実行し、第一主成分の loading 値が負で、かつ p 値が 0.05 未満の代謝物を有意代謝物として特定した。

最初に、ベースラインとして従来の ORA 手法を使用し、未検出代謝物を全て非有意とみなす解析を行った。この方法は、一般的に使用される MetaboAnalyst のアプローチに相当し、未検出代謝物がもたらす不確実性を考慮しないものである。次に、未検出代謝物の取り扱いに関する課題を解決するため、未検出代謝物が有意か非有意かの全ての組み合わせを考慮し、各経路における p 値の最小値と最大値を算出した。このアプローチにより、未検出代謝物がもたらす不確実性が経路有意性の推定に及ぼす影響を定量的に評価することが可能となった。

さらに、未検出代謝物に関する不確実性を確率的に取り扱うため、二つの確率モデルを導入した。まず、未検出代謝物における有意代謝物の数が二項分布に従うと仮定し、検出された代謝物の情報を基に、各経路での未検出有意代謝物数を推定する二項モデルを採用した。また、未検出代謝物の有意性を割合として捉え、その割合がベータ分布に従うと仮定したベータモデルも用いた。この方法は、二項モデルの代替アプローチとして、確率的に未検出代謝物の有意割合を推定するものである。

最後に、全ての経路に共通する情報を活用する階層ベイズモデルを導入し、p 値の信頼区間をさらに狭めた。このモデルでは、各経路の成功確率 θ [p] が全経路に共通する分布に従うと仮定し、設定した事前分布からリサンプリングを行った。リサンプリングによって得られた成功確率のサンプルに基づき、各経路での Fisher の正確検定を実施して p 値を算出し、リサンプリング結果から 95% 信頼区間を算出した。この手法により、観測データに依存しすぎることなく、未検出代謝物を考慮した信頼性の高い経路有意性の推定が可能となった。

Results and Discussion

本研究では、複数のアプローチを用いて、絶食マウスのメタボロミクスデータに対する Metabolite Set Enrichment Analysis（MSEA）を実施し、未検出代謝物の取り扱い方法による経路有意性評価への影響を検証した。

まず、従来の MetaboAnalyst スタイルの MSEA を実施し、未検出代謝物を非有意と仮定して経路の p 値を算出した（Table 1）。このアプローチでは、*Glycolysis* や *Pentose*

phosphate pathway のように強い有意性が示される経路も存在したが、未検出代謝物の影響が考慮されないため、多くの経路において p 値が高くなり、有意性が低く評価される傾向が観察された。

検出された代謝物のみを用いるアプローチ（Table 2）では、未検出代謝物の情報を排除することで観測データに基づく解析が可能となり、*Glycolysis* および *TCA cycle* などが有意性を示したものの、未検出代謝物が経路の有意性に与える潜在的な影響が考慮されていない点で、過小評価のリスクが依然として存在することが明らかになった。

次に、未検出代謝物に対するシンプルな補完法を適用し（Table 3）、検出済の代謝物に基づいて未検出代謝物の有意性割合を推定した。このアプローチでは、未検出代謝物の影響が部分的に考慮され、*Tyrosine metabolism* を含む一部の経路が新たに有意性を示す結果が得られた。この補完法は全経路における p 値の範囲を絞り、未検出代謝物を無視するよりも精度の高い結果を提供するものの、補完法自体の仮定に依存する制約が依然として残されている。

また、未検出代謝物の全ての組み合わせを考慮し、p 値の最小値と最大値の範囲を算出するアプローチ（Table 4）では、未検出代謝物が有意性に与える影響が明確に示され、*Lysine metabolism* や *Histidine metabolism* などの経路で広範囲な p 値が得られた。このアプローチにより、未検出代謝物が経路評価に与える不確実性が定量化されたものの、範囲が広いため経路有意性の判断に課題が残る結果となった。

このような未検出代謝物の影響を確率的に扱うため、二項分布モデルを導入し（Table 5）、未検出代謝物の有意数が二項分布に従うと仮定して p 値範囲を推定した。このモデルでは、未検出代謝物の不確実性を考慮しつつ p 値の範囲を絞ることができ、特に *Glycolysis* や *TCA cycle* のような経路で一貫した有意性が示され、二項分布モデルが柔軟な対応手法として有効であることが示唆された。

さらに、未検出代謝物の有意割合をベータ分布でモデル化したアプローチ（Table 6）では、二項分布モデルよりも一部の経路でさらに狭い範囲の p 値が得られ、特に *Glycolysis* や *Tyrosine metabolism* のような経路で強い有意性が確認された。この結果は、ベータ分布が未検出代謝物の割合に関する不確実性をより柔軟に反映するモデルとして有効であることを示唆している。

最後に、階層ベイズモデルを適用し、全経路に共通する情報を活用して各経路の p 値を推定した（Table 7）。このモデルは、未検出代謝物の不確実性を経路間で共有される分散に基づいて調整することで最も狭い信頼区間での p 値推定を可能にし、*Glycolysis* や *TCA cycle* といった経路では、他のモデルに比べてさらに高い信頼性で有意性が示された。階層ベイズモデルは、全経路間で共有される分散情報を統合することで、未検出代謝物の影響を最小限に抑え、観測データに過度に依存しない解析手法として優れた有効性を示した。

これらの結果から、未検出代謝物の不確実性を考慮する手法として階層ベイズモデルが最も信頼性の高いアプローチであることが示され、従来の ORA 手法の限界を超えて MSEA

の信頼性を向上させる方法として有用であることが明らかとなった。本研究のアプローチは、未検出データが存在するメタボロミクス解析において、経路有意性評価の精度向上に寄与するものである。

Conclusion

本研究では、従来の ORA 手法が未検出代謝物を非有意とみなすことによって生じる限界に対処するため、絶食マウスのメタボロミクスデータを使用し、未検出代謝物を考慮した階層ベイズアプローチを提案した。全ての未検出代謝物の組み合わせを考慮した範囲推定に加え、二項モデルおよびベータモデルを用いることで、未検出代謝物の不確実性を反映しつつ p 値の範囲を狭めることができた。最終的に、全経路に共通する情報を活用する階層ベイズモデルを適用することで、最も狭い信頼区間を達成し、より精度の高い ORA 結果を得ることができた。

この結果は、未検出代謝物が存在する場合においても、MSEA の信頼性を高めるための方法として階層ベイズアプローチが有効であることを示している。提案した手法は、メタボロミクスにおける pathway enrichment analysis の精度向上に寄与し、未検出代謝物の影響を考慮した信頼性の高い解析結果を提供するものである。本研究の成果は、メタボロミクスデータの解析における ORA 手法の改善に寄与するとともに、未検出データを含むさまざまなオミクス解析に応用できる可能性を示唆している。

References

1. Chong, J., et al. (2018). MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46(W1), W486-W494. ↩
2. Xia, J., & Wishart, D. S. (2016). Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Current Protocols in Bioinformatics*, 55(1), 14-10. ↩