Probabilistic 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エンリッチメント解析のための確率的 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[3]、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し、p 値の信頼区間をさらに狭めるために階層ベイズモデルを導入した。このモデルでは、各経路の成功確率（経路内の代謝物が有意となる確率）θ[p] が、全経路に共通する分布に従うと仮定し、経路間での成功確率に関する情報を統合的に扱った。具体的には、まず全体の成功確率のばらつきを示すパラメータ σ_θ を設定し、このばらつきのもとで各経路の成功確率 θ[p] が観測されるとした。次に、各経路の成功確率 θ[p]は、各経路における検出済の代謝物から観測された有意性の割合 μ_θ[p] をもとに、全体のばらつきパラメータ σ_θ により変動するものとして推定した。この観測割合 μ_θ[p]は、各経路の検出済代謝物のうち有意とされた代謝物の割合を表し、未検出代謝物を考慮した有意性の推定を行うための基礎データとして用いた。モデルのサンプリングにより、各経路ごとの成功確率の分布を得たのち、各成功確率に基づいて各経路における p 値の範囲を導出した。これにより、未検出代謝物が考慮された経路の有意性評価が可能となり、観測データのみに依存しない、信頼性の高い経路有意性の推定が実現した。最終的に、得られた成功確率のサンプルをもとに、各経路の p 値の 95% 信頼区間を算出することで、未検出代謝物の影響を含むより正確な有意性評価が達成された。

Results and Discussion

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て、従来の ORA 手法と新たに導入した確率モデルによるアプローチを適用し、未検出代謝物が経路の有意性評価に与える影響について詳細に検討した。

まず、従来の ORA 手法として MetaboAnalyst スタイルの ORA を適用し、未検出代謝物をすべて非有意と仮定して経路の有意性評価を行った（Table 1）。この手法では、Glycolysis（p = 1.00E-07）が最も有意であり、さらに Pentose phosphate pathway（p = 0.0034）や TCA cycle（p = 0.0042）といった代謝経路でも有意性が認められた。しかし、Glutamic acid and glutamine metabolism（p = 0.4552）や Valine, leucine and isoleucine metabolism（p = 0.9491）などの経路では高い p 値が得られ、有意性が低く評価される結果となった。このことから、従来の ORA 手法は観測データに強く依存し、未検出代謝物が多い経路（例：Glycolysis）では有意性が過小評価されるリスクがあることが示唆された。

次に、検出された代謝物のみを用いて経路の有意性を評価した（Table 2）。この手法では、未検出代謝物を無視し、観測データに基づく解析に限定することで、検出された代謝物のみに依存した経路評価が可能となる。Glycolysis（p = 2.33E-05）や TCA cycle（p = 0.0046）は依然として有意性を示し、さらに Pentose phosphate pathway（p = 0.034）および Polyamine metabolism（p = 0.034）が新たに有意な経路として検出された。このアプローチは、未検出代謝物に関する影響を排除し、観測されたデータのみで経路の有意性を精度高く評価する手法といえる。

これらの結果を基に考察すると、未検出代謝物をすべて非有意と仮定する場合、オッズ比が変動しやすくなり、結果として p 値に大きな影響を与える可能性がある。具体的には、Glycolysis の場合、検出された代謝物のみで評価すると、非有意な物質数が少なくなるため、オッズ比が低くなり、p 値が上昇して有意性が低く評価される傾向が見られる。一方、未検出代謝物をすべて非有意と仮定する MetaboAnalyst の方式では、非有意な物質数が増えることで、オッズ比が高くなり、p 値が小さくなって有意性が過大に評価される場合がある。このように、未検出代謝物の扱い方によってオッズ比が変動し、有意性評価に大きな影響が及ぶ点は、従来の ORA 手法における重要な課題であるといえる。また、未検出代謝物をすべて非有意と仮定する手法では、場合によっては経路の有意性が過小評価されるだけでなく、逆に過大評価されるリスクも存在する。特に、検出代謝物が特定の経路に偏っている場合や、未検出代謝物が多い場合には、メタボロミクスやプロテオミクスのデータセットにおいて観測結果に依存したバイアスが生じる可能性がある。

次に、Urea cycle についても同様に考察した。本経路では未検出代謝物が多く存在するため、従来の MetaboAnalyst スタイルの ORA 手法では未検出物質をすべて非有意と仮定することで、p 値に対する影響が特に顕著に現れる。具体的には、Urea cycle の全物質を対象にしたクロス集計表では、検出された有意な代謝物と非有意な代謝物の比率が 6/17 であ

り、検出された物質のみを用いた場合の 6/4 と比べて比率が逆転している。このため、未検出代謝物を非有意と仮定した場合、非有意な代謝物数が多くなり、オッズ比が変化して p 値が大きくなることが予測される。さらに、Urea cycle のように未検出物質が多い場合、欠損値が増えることでクロス集計表における有意/非有意のパターンが増加し、結果として従来の ORA 手法で計算される p 値がさらに変動する。このような場合、p 値が小さくなる傾向も確認され、特に従来の ORA 手法では経路の有意性が過大評価されるリスクが生じることも考えられる。従って、Urea cycle の場合は未検出物質が多いことから、未検出物質を非有意と仮定する従来手法では、特に有意性の評価が不安定になりやすいといえる。

また、未検出代謝物の全ての組み合わせを考慮し、各経路の p 値範囲（最小値と最大値）を算出するアプローチを導入した（Table 4）。この手法は、未検出代謝物が有意であるか非有意であるかの全パターンを考慮することで、p 値の範囲を広げる一方で、未検出代謝物の有意性によって経路評価がどれほど変わり得るかを明確に示すことができる。例えば、*Glycolysis* の p 値範囲は 8.38E-06 から 0.0001985 であり、経路が有意であることが確認される一方で、*TCA cycle* では p 値範囲が 7.27E-06 から 0.1967 と広がり、未検出代謝物の有意/非有意が経路評価に大きな影響を与えることが示唆された。このアプローチは、未検出代謝物の不確実性が経路評価に及ぼす影響を定量的に評価するため有用であるが、最小値と最大値の差が広がりすぎるため、最終的な有意性評価が曖昧になり得る。

未検出代謝物の影響をより狭い範囲で調整するため、未検出代謝物の有意数が二項分布に従うと仮定したアプローチを導入した（Table 5）。二項モデルでは、未検出代謝物の数が確率的に分布すると仮定することで、p 値範囲を絞り込み、未検出代謝物の影響を一定範囲に収束させる。この方法では、*Glycolysis* の p 値範囲が 2.55E-06 と非常に狭く、他の手法に比べ安定した有意性が得られた。また、*Tyrosine metabolism* の p 値範囲も 1.20E-17 から 1.33E-06 と狭く、未検出代謝物が存在しても有意性が確認された。この二項モデルによるアプローチは、観測データに基づく有意性評価と未検出代謝物の不確実性のバランスをとるため、経路の有意性評価において精度の高い結果が得られた。

さらに、未検出代謝物の有意割合がベータ分布に従うと仮定することで、未検出代謝物の割合に対する柔軟な評価が可能となるアプローチを導入した（Table 6）。ベータ分布は、未検出代謝物の割合に基づき、柔軟な仮定を導入するため、二項モデルよりも広範な解釈が可能である。具体的には、*Glycolysis* の p 値範囲は 2.55E-06 から 6.66E-05 に収束し、二項モデルに比べ若干広がりが見られるが、依然として高い有意性が確認された。また、*TCA cycle* の p 値範囲も 7.27E-06 から 0.0054 に調整され、二項モデルと同様に経路の有意性評価を柔軟に行えることが示された。このベータ分布モデルは、未検出代謝物の割合に応じた柔軟な推定が可能で、観測データに偏らない経路評価を実現した。

最後に、階層ベイズモデルを導入することで、全経路に共通する情報を統合し、経路間で共有される分散情報に基づく信頼区間での p 値評価を行った（Table 7）。このアプローチでは、未検出代謝物の影響が他の経路からの情報によって補完され、全体としてより狭い信頼

区間が得られた。たとえば、*Glycolysis* の p 値範囲は 2.55E-06 から 6.66E-05 で、従来の手法に比べて最も狭い範囲に収まった。また、*TCA cycle*（p 値範囲: 1.04E-04 から 0.0237）も同様に狭い範囲で評価され、信頼性の高い経路の有意性評価が可能となった。この階層ベイズモデルは、全経路間の共有情報により観測データの偏りを最小限に抑え、経路の有意性評価において最も精度の高い結果が得られることを示している。

総合すると、従来の ORA 手法では未検出代謝物の影響を考慮できず、観測データに依存するため経路の有意性が過小評価されるリスクがあった。一方で、全組み合わせによる p 値範囲の推定は、未検出代謝物の影響を定量的に把握するうえで有用であったが、p 値の範囲が広がりすぎることで、結果の解釈が困難になる可能性があった。

確率モデルの導入により、二項モデルやベータモデルを用いることで、未検出代謝物に関する不確実性を一定範囲に抑えつつ、より安定した経路評価が可能であることが確認された。特に二項モデルは未検出代謝物数の確率的推定を行うことで観測データと未検出代謝物の影響をバランス良く評価し、ベータモデルは未検出代謝物の割合に対する柔軟な仮定を導入することで、多様なシナリオに対応する解析を可能にした。

さらに、階層ベイズモデルの適用により、全経路に共通する情報を利用して未検出代謝物の影響を統合的に調整することで、観測データに過度に依存しない信頼性の高い p 値評価が実現された。このモデルは全経路間で共有される分散情報を組み入れることで、未検出代謝物に関連する不確実性が最も効果的に抑えられ、観測データの偏りを最小限に抑えた評価が可能であった。

以上の結果から、未検出代謝物を考慮した新たな確率モデルと階層ベイズアプローチは、メタボロミクスにおける経路の有意性評価の精度を高める上で有用であることが示された。本研究のアプローチは、特に未検出代謝物の影響が経路評価に大きな不確実性をもたらすケースにおいて、従来の ORA の限界を超える方法として有望である。

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.
3. Yamamoto, H., et al. (2014). Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinformatics*, 15(1), 51.