**未検出代謝物を考慮したMetabolite Set EnrichmentのためのOver-Representation Analysis**

**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。

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

**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.

**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**

**Data Preparation**

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使用した。このデータは、代謝物の検出と未検出に分類し、検出された代謝物から有意性を評価した。具体的には、PCA を実行し、第一主成分における loading 値が負であり、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。

**Traditional ORA Approach (Baseline)**

まず、従来の ORA 方法をベースラインとして使用し、未検出代謝物を考慮せずに解析を行った。この方法は、MetaboAnalyst において採用されているアプローチと同様に、未検出代謝物を全て非有意とみなすものである。

**ORA with Full Range Estimation for Undetected Metabolites**

従来の ORA 方法の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 値の最小と最大の範囲を算出した。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の影響を定量化することが可能となった。

**Probabilistic Models for ORA**

1. **Binomial Model for Undetected Metabolites**:

未検出代謝物の中に含まれる有意代謝物の数を二項分布に従うものとしてモデル化した。これにより、検出済の代謝物の情報を基準に、各経路での未検出有意代謝物数を推定することができた。

1. **Beta Model for the Proportion of Significant Undetected Metabolites**:  
   未検出代謝物の中に含まれる有意代謝物の割合をベータ分布に従うものとしてモデル化した。この方法は、確率的に有意性の割合を見積もるものであり、二項モデルの代替手法として機能する。

**Hierarchical Bayesian Model**

最後に、全ての経路に共通の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ものとして仮定し、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 の正確検定を実行し、各パスウェイの p 値を算出した。リサンプリング結果から 95% 信頼区間を得ることで、観測データに依存しすぎない p 値の推定が可能となった。

**Results and Discussion**

**Baseline ORA Results**

従来の ORA 方法では、未検出代謝物が全て非有意とみなされてしまうことから、未検出代謝物の影響を過小評価するリスクがあることが示された。特に、未検出代謝物が多い経路において、この過小評価の影響が顕著であることが確認された。

**Range Estimation for ORA with Undetected Metabolites**

未検出代謝物が有意か非有意かの全組み合わせを考慮することにより、p 値範囲が大幅に広がり、未検出代謝物による不確実性の影響が明確に反映された。特に、代謝物が未検出である場合に生じる不確実性が、従来の ORA 方法による結果とは異なり、過小評価されるリスクを軽減することができた。しかし、このアプローチでは、全ての組み合わせを考慮するため、得られた p 値範囲は非常に広く、経路の有意性を判断するには精度に課題が残ることがわかった。

**Probabilistic Models**

次に、未検出代謝物の不確実性を二項モデルおよびベータモデルを用いて確率的に取り扱った。二項モデルでは、未検出代謝物の中で有意とされる代謝物数が二項分布に従うものと仮定し、検出された有意代謝物の割合に基づいて推定を行った。これにより、p 値範囲を従来の ORA よりも狭めることができ、未検出代謝物の有意性に関する不確実性を反映する結果が得られた。同様に、ベータモデルでは、未検出代謝物の有意割合がベータ分布に従うと仮定し、p 値範囲の精度をさらに向上させた。これら二つの確率モデルは、いずれも類似した結果を示し、未検出代謝物に関連する不確実性を確率的に取り扱うことで、より信頼性の高い p 値範囲を得られることが示された。

**Hierarchical Bayesian Model**

最後に、階層ベイズモデルを用いることで、全経路に共通の情報を活用し、さらに狭い p 値信頼区間を達成することができた。階層ベイズモデルでは、各経路の成功確率 θ[p]\theta[p]θ[p] が全経路にわたって共通の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サンプルに基づいて Fisher の正確検定を実行し、各経路の p 値を算出した。こうすることで、観測データに依存しすぎることなく、事前分布の影響を適切に反映した p 値範囲が得られ、従来の ORA に比べて最も狭い信頼区間を実現することができた。

これにより、絶食マウスデータにおいて、階層ベイズアプローチを用いることが未検出代謝物を考慮した場合の pathway enrichment analysis において有効であることが示唆された。

**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. [↩](#user-content-fnref-1)
2. Xia, J., & Wishart, D. S. (2016). Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Current Protocols in Bioinformatics*, 55(1), 14-10. [↩](#user-content-fnref-2)