

MSplusR: An R Package for Visualization of MS/MS Spectral Data Using UMAP

Hiroyuki Yamamoto

h.yama2396@gmail.com

Japan Computational Mass Spectrometry (JCompMS) group

Abstract

Mass spectrometry-based omics research requires efficient tools for processing and visualizing MS/MS spectral data. To address this, we developed **MSplusR**, an R package that integrates preprocessing, similarity computation, dimensionality reduction, and interactive visualization. MSplusR processes raw mzML files into binned and normalized spectral matrices, computes similarity matrices for clustering, and employs UMAP and PCA for dimensionality reduction. An interactive Shiny viewer enables visualization of UMAP projections and MS/MS spectra. Applicable to metabolomics, proteomics, and lipidomics, MSplusR facilitates biomarker discovery and sample clustering. The package is open-source and available on GitHub, with documentation to ensure reproducibility and broad adoption.

Keywords: mass spectrometry, metabolomics, dimensionality reduction

MSplusR: UMAP を用いた MS/MS スペクトルデータ可視化のための R パッケージ

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

質量分析を用いたオミクス研究では、MS/MS スペクトルデータを効率的に処理・可視化するツールが必要である。この課題に対応するため、MSplusR を開発した。本パッケージは、mzML ファイルの前処理、類似度行列の計算、次元削減、インタラクティブな可視化を統合的に提供する。UMAP や PCA を用いた次元削減と、Shiny を活用した可視化ツールにより、UMAP 結果と MS/MS スペクトルを容易に探索可能である。MSplusR は、メタボロミクス、プロテオミクス、リポドミクスに応用可能であり、バイオマーカーの発見やサンプルのクラスタリングを支援する。GitHub でオープンソースとして公開されており、詳細なドキュメントが再現性を支援している。

Introduction

メタボロミクスやプロテオミクスをはじめとする質量分析を用いたオミクス研究では、測定された生データを解析し、サンプルと代謝物もしくはタンパク質の行列データを作成するまでに、いくつものステップが必要である[1, 2]。その中でも、生データの可視化は、データを正しく理解し、解析を効果的に進めるために不可欠な工程である。このニーズに応えるため、質量分析メーカー各社が可視化を支援するソフトウェアを提供しており、例えば Mass++ は、MS スペクトルを可視化するための代表的なツールとして広く使用されている[3]。

Data Dependent Acquisition (DDA) などの手法では、数千もの MS/MS スペクトルが一括して取得され、これらは代謝物やペプチドのアノテーションに利用されている[4,5]。しかし、取得されたスペクトルを一つずつ目視で確認するのは現実的ではなく、効率的な解析手法が求められている。MS/MS スペクトルネットワークは、データを視覚的に理解するための有効な方法の一つである[6]。ネットワークの構造を利用することで、特定の MS/MS スペクトルに着目し、類似性の高いスペクトルを探索することが可能であるが、ネットワークが複雑化するため、数千に及ぶスペクトル全体を直感的に把握するのは難しい。

そこで本研究では、DDA で取得した MS/MS スペクトルデータ全体を効率的に理解するために、MS/MS スペクトルを UMAP で可視化する「MSplusR」パッケージを開発した。このパッケージは、まず MS/MS スペクトルをビンニングしてデータ行列に変換し、次に主成分分析 (PCA) またはグラフ隣接行列の特異値分解 (SVD) を用いて次元削減を行う[7,8]。その後、UMAP を用いてデータを可視化することができる。また、UMAP

スコアの各データポイントをクリックすると、対応する MS/MS スペクトルを表示する機能も含まれており、データ全体を俯瞰するだけでなく、個々のデータについて詳細な解析を行うことも可能である。

MSplusR

MSplusR は、質量分析データ解析を効率化するための多機能な R パッケージであり、mzML データの前処理、スペクトル間の類似度計算、次元削減、そしてインタラクティブな可視化を統合的に提供している。本パッケージは、質量分析研究における幅広いニーズに応える柔軟性を持つフレームワークである。

まず、`process_mzML` 関数は、mzML 形式のスペクトルデータを読み込み、ピーク検出やビンニング、正規化などの前処理を行う [9]。これにより、データは統一された形式で整理され、後続の解析に適した形に変換される。また、スパース行列表現を活用することで、処理効率を向上させている。

次に、`process_similarity_matrix` 関数は、共有スペクトルピークに基づいて類似度行列を計算する。この類似度行列は、クラスタリングやネットワーク解析において重要な要素となり、特にデータ内の関係性を把握する際に役立つ [6]。このように、類似度行列を効率的に生成する機能は、メタボロミクスやプロテオミクスといった多くのオミクス分野での応用が期待される。

次元削減手法としては、`process_graph_umap` 関数と `pca_umap_analysis` 関数が提供されている。これらの関数は、UMAP や PCA を利用して高次元データを低次元空間に変換することで、データ内のパターンや関係性を視覚的に理解することを可能にしている [7,8]。また、`create_shiny_umap_viewer` 関数を用いることで、研究者は UMAP の結果をインタラクティブに探索し、対応する MS/MS スペクトルを簡単に確認することができる。このインタラクティブビューアには、選択されたスペクトルの履歴を管理する機能が含まれており、スペクトル間の比較を容易にしている。

MSplusR は、メタボロミクス、プロテオミクス、リポドミクスといった多様なオミクス分野で応用可能である。メタボロミクスでは、mzML データの処理や類似度行列の計算を通じて、代謝パターンの特特定や類似サンプルのクラスタリングを支援する。プロテオミクスやリポドミクスでは、複雑なスペクトルデータを解析し、バイオマーカーの特特定やサンプル間の関係性の可視化を効果的に行うことができる。また、インタラクティブなビューアを活用することで、データ品質の検証や特定のスペクトル特徴の詳細な調査を行うことも可能である。

このように、MSplusR は、質量分析データの前処理、類似度計算、次元削減、可視化を統合的に提供する堅牢で使いやすいフレームワークであり、メタボロミクスやプロテオミクス研究における重要な課題に対応している。本研究では、MSplusR を用いて MS/MS スペクトルデータ全体を効率的に処理し、データを可視化した。MSplusR は、GitHub 上でオープンソース R パッケージとして公開されており

(<https://github.com/hiroyukiyamamoto/MSplusR>)、詳細なドキュメントとサンプルワ

ークフローが提供されている[10]。これにより、研究での採用と再現性の確保が容易になっている。

最後に、本研究の解析結果を図 1 に示す。UMAP を用いてデータ全体を可視化した結果、いくつかのクラスターが形成されていることが確認された。さらに、特定のクラスターに注目して表示をズームすると、クラスター内のより詳細な情報を確認することができる。また、UMAP プロット上の各データ点をクリックすることで、対応する MS/MS スペクトルを表示することが可能である。実際に、UMAP 上で非常に距離が近いスペクトルを同一クラスター内で確認したところ、同じ m/z にピークを持つ類似したスペクトルが確認された。この結果は、MSplusR によるデータの効率的な解析と可視化が、スペクトル間の類似性を明確に示すのに有用であることを示している。

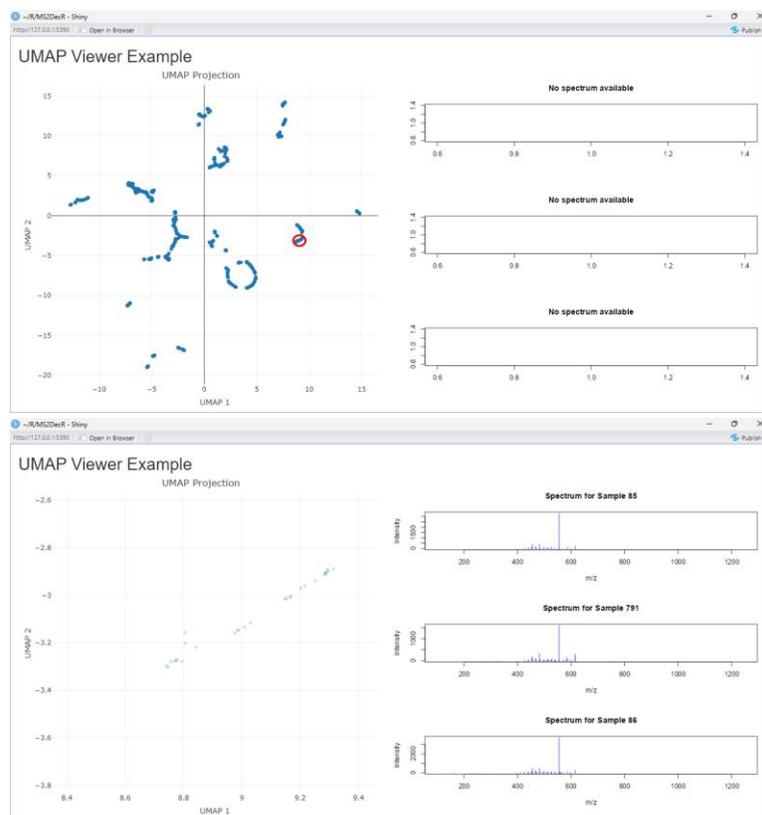


図 1 MSplusR でスペクトルデータを可視化した結果

Reference

1. Dunn W.B., Erban A., Weber R.J., Creek D.J., Brown M., Breitling R., Hankemeier T., Goodacre R., Neumann S., Kopka J., Viant M.R. (2013). "Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics." *Metabolomics*, 9(1), 44–66. doi:10.1007/s11306-012-0434-4.
2. Aebersold R., Mann M. (2016). "Mass-spectrometry-based proteomics." *Nature*, 537(7620), 347–355. doi:10.1038/nature19949.
3. Tanaka S., Fujita Y., Parry H.E., Yoshizawa A.C., Morimoto K., Murase M., Yamada Y., Yao J., Utsunomiya S., Kajihara S., Fukuda M., Ikawa M., Tabata T., Takahashi K., Aoshima

- K., Nihei Y., Nishioka T., Oda Y., Tanaka K. (2014). "Mass++: A visualization and analysis tool for mass spectrometry." *Journal of Proteome Research*, 13(8), 3846–3853. doi:10.1021/pr5006636.
4. Kind T., Fiehn O. (2010). "Advances in structure elucidation of small molecules using mass spectrometry." *Bioanalytical Reviews*, 2(1–4), 23–60. doi:10.1007/s12566-010-0015-6.
 5. Smith C.A., Want E.J., O'Maille G., Abagyan R., Siuzdak G. (2006). "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification." *Analytical Chemistry*, 78(3), 779–787. doi:10.1021/ac051437y.
 6. Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. (2003). "Cytoscape: A software environment for integrated models of biomolecular interaction networks." *Genome Research*, 13(11), 2498–2504. doi:10.1101/gr.1239303.
 7. Becht E., McInnes L., Healy J., Dutertre C.A., Kwok I.W., Ng L.G., Ginhoux F., Newell E.W. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP." *Nature Biotechnology*, 37(1), 38–44. doi:10.1038/nbt.4314.
 8. McInnes L., Healy J., Melville J. (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv preprint arXiv:1802.03426*. Retrieved from <https://arxiv.org/abs/1802.03426>.
 9. Couvillion S.P., Nagana Gowda G.A., Raftery D., Suslick K.S. (2015). "Metabolomics." *Annual Review of Analytical Chemistry*, 8, 17–37. doi:10.1146/annurev-anchem-071114-040237.
 10. Tyanova S., Temu T., Cox J. (2016). "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics." *Nature Protocols*, 11(12), 2301–2319. doi:10.1038/nprot.2016.136.