**LargeScaleMSPy: A Python Library for Visualizing**

**Large-Scale MS/MS Spectral Data**

**Hiroyuki Yamamoto**

[h.yama2396@gmail.com](mailto:h.yama2396@gmail.com)

Japan Computational Mass Spectrometry (JCompMS) group

**Abstract**

Mass spectrometry (MS) generates vast amounts of spectral data that require efficient preprocessing, dimensionality reduction, and visualization. LargeScaleMSPy is a Python library designed to streamline these tasks by providing tools for spectral data preprocessing, filtering, PCA-based dimensionality reduction, UMAP visualization, and interactive exploration of MS/MS spectra. This paper introduces the key components of LargeScaleMSPy, including its ability to handle large-scale spectral data in sparse formats, perform robust PCA and UMAP analysis, and visualize MS/MS spectra through an interactive web-based application. The library offers a flexible and scalable solution for researchers working with high-throughput MS/MS data.

**Keywords:** Mass spectrometry, MS/MS, UMAP, PCA, Python library, spectral data analysis.

LargeScaleMSPy: 大規模なMS/MSスペクトルデータ可視化のためのPythonライブラリ

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（MS）は膨大なスペクトルデータを生成するが、その前処理、次元削減、および可視化には効率的なツールが求められる。本研究では、スペクトルデータの前処理、フィルタリング、主成分分析（PCA）による次元削減、UMAPを用いた可視化、さらにインタラクティブなMS/MSスペクトルの探索を可能にするPythonライブラリLargeScaleMSPyを紹介する。特に、大規模スペクトルデータのスパース形式での効率的な処理、PCAとUMAPによる次元削減、インタラクティブな可視化ツールの機能について説明する。このライブラリは、ハイスループットMS/MSデータを扱う研究者に対し、柔軟かつスケーラブルなソリューションを提供する。

**Introduction**

質量分析は、プロテオミクスやメタボロミクスをはじめとするオミクス研究において必要不可欠な技術である。特に、タンデム質量分析（MS/MS）で得られるプロダクトイオンスペクトルは、化学構造の同定やアノテーションにおいて重要な役割を果たしている。これらのスペクトルを視覚的に表現することで、スペクトルデータと化合物情報を効率的に紐付けることが可能となる。

我々が以前開発したMSplusR [1]は、Data Dependent Acquisition (DDA)で取得されたプロダクトイオンスペクトルを対象に、主成分分析（PCA）を用いて次元削減を行い、その後UMAPを適用して可視化するツールである。このツールでは、同一クラスター内のスペクトルをインタラクティブに確認することで、類似度の高いスペクトルを抽出できる。DDAで得られるスペクトル数は1サンプルあたり数千程度であり、そのため計算時間やメモリ使用量に関しては問題が生じることはほとんどない。しかし、アノテーションのために利用される大規模スペクトルライブラリ、例えばCompMSが公開する脂質の理論スペクトルライブラリやMassBank Humanには、数十万のスペクトルが含まれている。このような大規模データに対してPCAを適用することは、従来の方法では容易ではなかった。

そこで本研究では、シングルセルトランスクリプトーム解析などの大規模データ解析に利用されている、逐次的にデータをメモリに読み込みつつ主成分分析を行い、最終的に統合するout-of-coreな手法であるIncremental PCA [2,3]を採用した。この手法を用いることで、得られた主成分スコアに対してUMAPを適用し、大規模なスペクトルデータの可視化を実現した。さらに、Incremental PCAを効率的に行うため、スペクトルデータをビニングして大規模な行列データに変換した。このデータは0を多数含むスパースデータであるため、スパースなHDF5形式で保存することでファイルサイズを大幅に削減した。

加えて、これらの一連の処理を効率的に実行可能なPythonライブラリとしてLargeScaleMSPyを開発した。このライブラリにより、大規模スペクトルデータ解析の柔軟性とスケーラビリティを向上させた。

**Implementation**

本研究では、大規模なスペクトルデータを効率的に処理・解析するための包括的なワークフローを構築した。このワークフローは、データの前処理、次元削減、インタラクティブな可視化までを含んでいる。

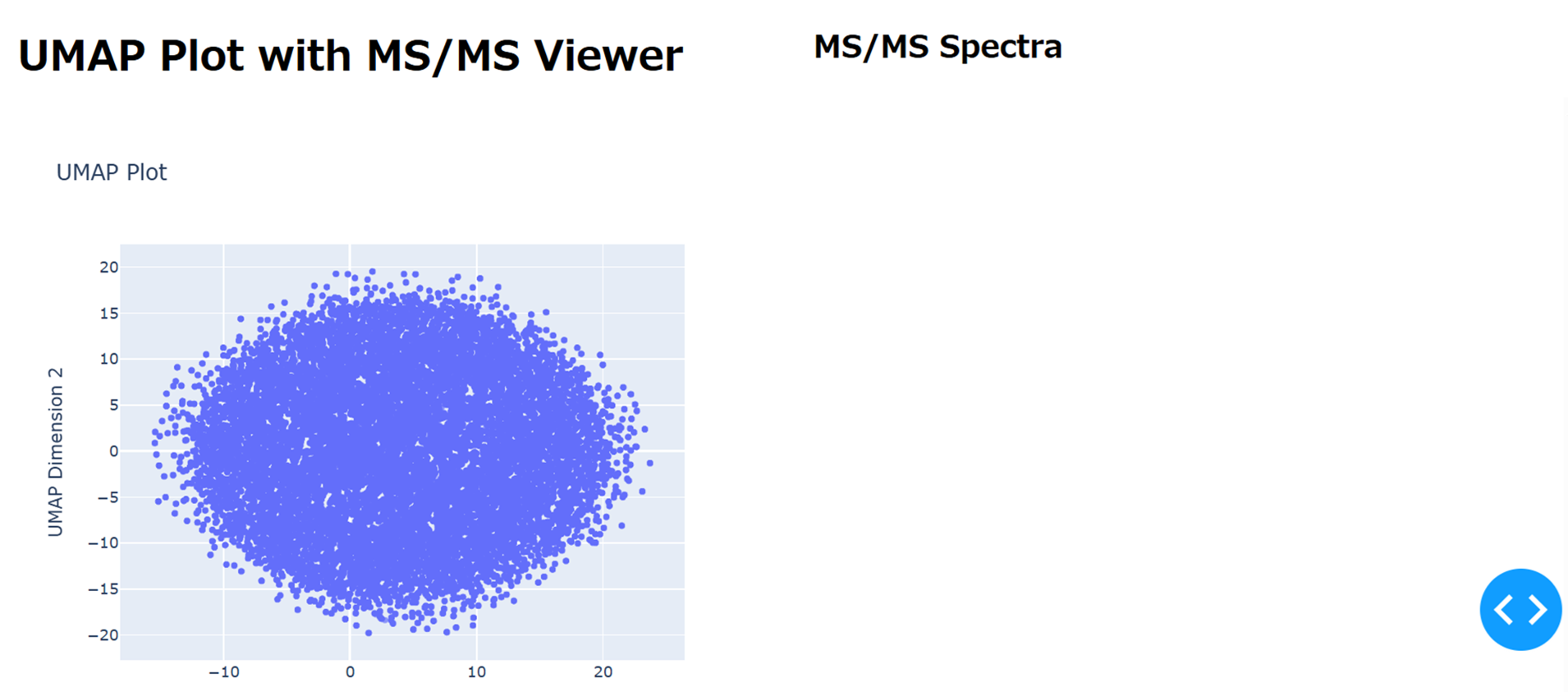
まず、スペクトルデータの前処理と変換を行った。msp2hdf5.pyスクリプトを用いて、MSP形式のスペクトルデータをHDF5形式に変換し、スパース行列を活用してメモリ使用量を最適化した。この前処理では、m/z範囲のビニング、ノイズフィルタリング、強度の正規化を実施し、データを後続の解析に適したスパース行列形式に整えた。

次に、次元削減には主成分分析（PCA）とUMAPを組み合わせたワークフローを構築した。pca\_umap\_analysis.pyスクリプトを使用してHDF5ファイルを読み込み、まずPCAを適用して高次元データを低次元に圧縮し、その後UMAPを用いてデータを二次元空間に埋め込んだ。このプロセスにより、高次元データのクラスタリングや関係性を視覚的に把握しやすくした。

最後に、UMAPによる可視化結果をインタラクティブに探索できるWebアプリケーションを開発した。このアプリケーションはcreate\_umap\_app.pyスクリプトを基にDashフレームワークを活用して構築したものである。UMAPプロジェクション上のデータポイント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機能を提供し、さらに直感的なインターフェースと閲覧履歴管理機能を実装することで、データの分析と解釈の効率を向上させた。

**Results**

　本研究では、CompMS [4]で公開されている脂質の理論スペクトルライブラリMSDIAL-TandemMassSpectralAtlas-VS69-Pos.mspをデータとして用いて、可視化を行った。mspファイルを読み込み、スパースなHDF5ファイル形式に変換した後、主成分分析（PCA）を用いて次元削減を行い、その後UMAPを適用した。得られたUMAPの計算結果を可視化したものを図1に示す。



グラフ

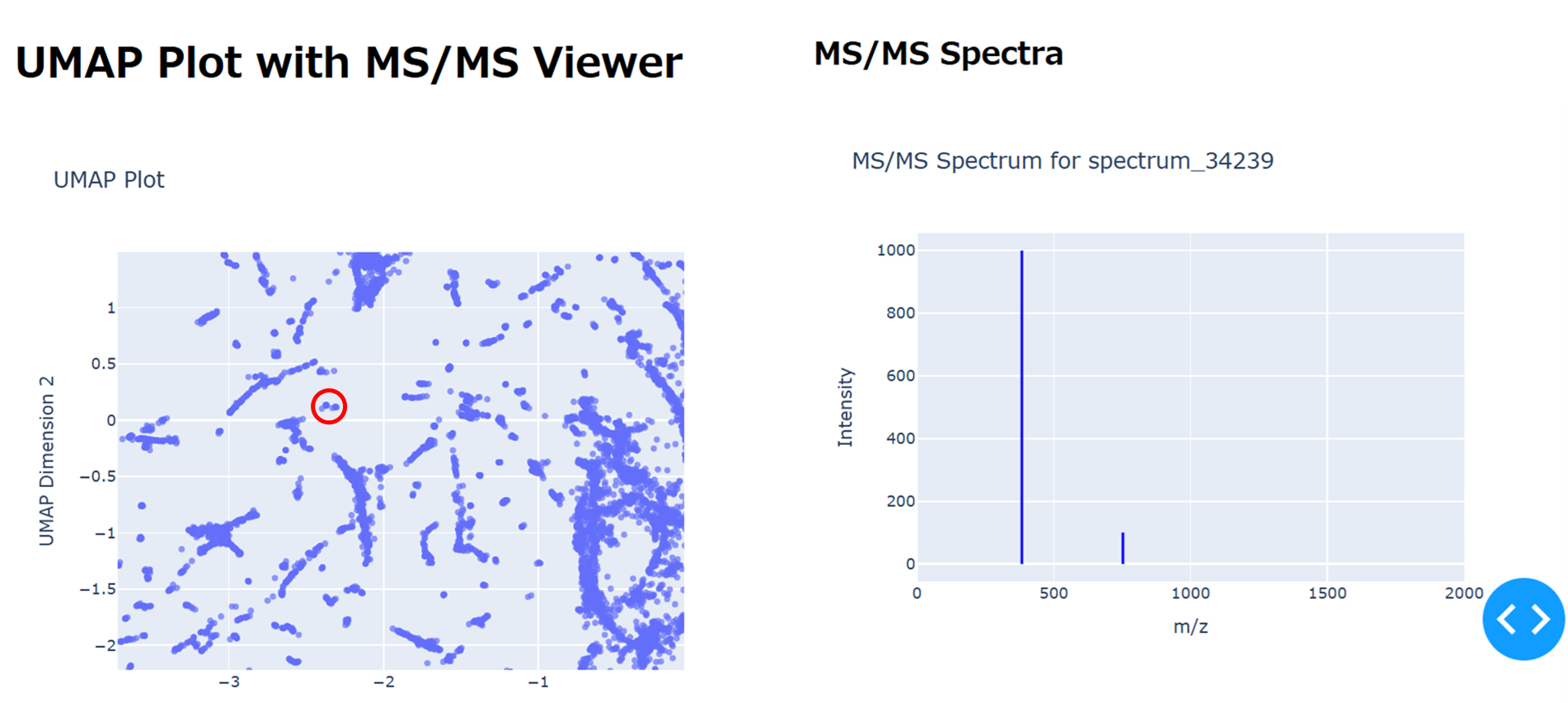
中程度の精度で自動的に生成された説明

図1 UMAPとスペクトルの可視化

図1の上部に示されたUMAPプロットの特定領域を拡大し、クラスター状に分布している点を確認したところ、同じm/zにピークを持つスペクトルが含まれていることがわかった。この結果は、UMAPによる可視化が、MSplusRで得られる結果と同様に、類似したスペクトルがある程度集中してクラスターを形成する傾向を示していることを示唆している。

**Conclusion**

本研究で開発したLargeScaleMSPyは、MS/MSスペクトルデータの前処理、次元削減、および可視化を一貫して実行できる包括的なフレームワークを提供する。このライブラリは、スパースデータ処理、主成分分析（PCA）、UMAP、およびインタラクティブな可視化を活用し、大規模スペクトルデータ解析の課題に対応する柔軟でスケーラブルなソリューションを実現した。LargeScaleMSPyは、オープンソースのPythonライブラリとしてGitHubで公開している[5]。

**References**

[1] MSplusR : https://github.com/hiroyukiyamamoto/MSplusR

[2] Tsuyuzaki, K., Sato, H., Sato, K. *et al.* Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biol* **21**, 9 (2020). https://doi.org/10.1186/s13059-019-1900-3

[3] D. Ross, J. Lim, R. Lin, M. Yang, Incremental Learning for Robust Visual Tracking, International Journal of Computer Vision, Volume 77, Issue 1-3, pp. 125-141, May 2008.

[4] <https://systemsomicslab.github.io/compms/msdial/main.html>

[5] <https://github.com/hiroyukiyamamoto/LargeScaleMSPy>