**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しますが、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必要です。LargeScaleMSPyは、スペクトルデータの前処理、フィルタリング、PCAを用いた次元削減、UMAPによる可視化、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ライブラリです。本稿では、LargeScaleMSPyの主な機能である大規模スペクトルデータのスパース形式での処理、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について説明します。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柔軟でスケーラブルなソリューションを提供します。

**Introduction**

質量分析（MS）は、オミクス科学において重要な分析技術であり、分子構造や存在量に関する詳細な情報を提供します。特にタンデム質量分析（MS/MS）で生成されるスペクトルは、化合物の特定と定量に不可欠な情報をもたらします。しかし、大規模なMS/MSデータセットを扱う際には、高次元性やデータサイズの問題から多くの課題が生じます。これらの課題に対応するには、効率的な前処理、次元削減、可視化のための計算ツールが必要です。

LargeScaleMSPyは、これらのニーズに応えるために開発されたPythonライブラリであり、スペクトルデータの高度な処理能力、PCAおよびUMAPによる次元削減、インタラクティブな可視化を統合しています。このライブラリのモジュール設計により、シームレスなデータ解析ワークフローが可能となり、メタボロミクス、プロテオミクス、リピドミクスの研究者にとって貴重なリソースとなります。

**Method**

1. 前処理とデータ変換

LargeScaleMSPyは、MSP形式のスペクトルデータを効率的なHDF5ファイルに変換し、メモリ使用量を効果的に管理するためにスパース行列を利用します。前処理パイプラインには、m/z範囲のビニング、ノイズフィルタリング、強度の正規化が含まれ、データが後続の解析に適した状態に整えられます。これらの操作はmsp2hdf5.pyスクリプトで実行され、前駆体m/z値やスペクトル固有の属性などのメタデータを保持します。

2. データの読み込みとスパース形式の処理

このライブラリは、load\_hdf5.pyスクリプトを介してHDF5ファイルからスペクトルデータを取得する機能を提供します。個々のスペクトルをロードし、それらのスパース行列表現と関連するメタデータ（前駆体m/z値など）を簡単に操作できます。この効率的なデータ処理により、大規模データセットに対するスケーラビリティが保証されます。

3. PCAとUMAPによる次元削減

LargeScaleMSPyは、初期次元削減のためのPCAと、さらに2次元空間への埋め込みを可能にするUMAPを統合しています。このワークフローにより、高次元データセット内のクラスタリングや関係性を視覚化しやすくなります。これらの操作は、pca\_umap\_analysis.pyスクリプトを使用して自動化され、解釈可能な可視化を生成します。

4. インタラクティブな可視化

UMAPプロジェクションをインタラクティブに探索するために、DashベースのWebアプリケーションがcreate\_umap\_app.pyで提供されます。このアプリケーションでは、UMAPデータポイントをクリックして対応するMS/MSスペクトルを表示できます。直感的なインターフェースと最近閲覧したスペクトルの履歴管理により、データの分析と解釈が向上します。

**Results**

LargeScaleMSPyは、ハイスループットMS/MSデータ解析に特化しており、多様な応用が可能です。

1. メタボロミクスおよびプロテオミクス: クラスタリングパターンや化合物間の関係、バイオマーカーを特定するための解析を支援します。
2. 品質評価: インタラクティブにMS/MSスペクトルを可視化することで、データポイントの品質を評価し、解析を最適化できます。
3. 次元削減による可視化: PCAおよびUMAPのワークフローにより、複雑なデータセットを解釈しやすくし、隠れたパターンを発見できます。

**Conclusion**

LargeScaleMSPyは、MS/MSスペクトルデータの前処理、次元削減、および可視化のための包括的なフレームワークを提供します。スパースデータ処理、PCA、UMAP、およびインタラクティブな可視化を活用することで、大規模スペクトルデータ解析の課題に対応します。今後の開発では、機械学習モデルとの統合による自動スペクトル解釈を含む機能拡張を目指しています。

公開情報

LargeScaleMSPyは、オープンソースのPythonライブラリとしてGitHubで公開されています：<https://github.com/hiroyukiyamamoto/LargeScaleMSPy>。詳細なドキュメントとサンプルスクリプトが提供され、多様な研究ワークフローでの採用を支援します。