Technical Sheet
No. 22-26

データ解析入門 14 <ケミカルスペースの可視化>

キーワード：ケミカルスペース、ケモインフォマティクス、類似構造検索、機械学習

はじめに

前稿 [1]では有機化合物をコンピュータで処理可能な形式に変換する方法について紹介しました。本稿では低分子化合物に可視化手法などを適用し、探索的データ解析を実行します。

データセットについて

以下、臭気化合物に関するデータセット [2]のうち、重複を除いた 1098 件の SMILES と官能評価ラベル（Descriptor 1）を解析対象とします。26 種類の官能評価ラベルが存在し、各ラベル数には偏りがあります（図 1）。

上記化合物をベクトルとして表現し、UMAP（Uniform manifold approximation and projection）により2次元に可視化します。化学構造を反映した高次元空間は、ケミカルスペースと呼ばれることがあり、化合物探索などの場面で使用されます。文献 [3]には創薬分野におけるケミカルスペースの活用例などが紹介されています。

次に、RDKit により分子記述子を算出し[1]、化合物ごとにベクトルを準備します。図 2 に、UMAP の可視化結果を灰色のプロットで示し、12 種類の官能評価ラベルの分布を色付きで例示しました。図 2 を見るかぎり、化学構造と官能評価ラベルには、はっきりとした相関があるとは言えず、同じようなにおいがする化合物でも、それらの構造は必ずしも類似しないということが窺えます。ただし、化学構造の表現方法には絶対的な方法があるわけではなく、今回は 2 次元構造に基づき、解析を実行しました。したがって、どのような特徴量を用いるべきかという点には検討の余地があります。

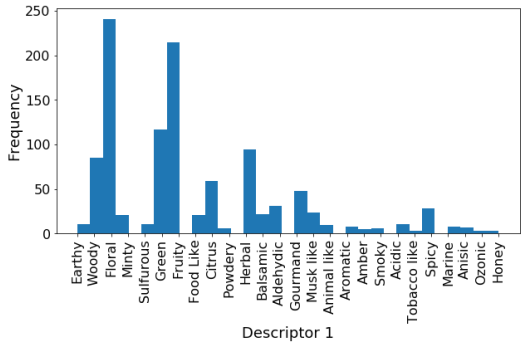

図 1　官能評価ラベル（Descriptor 1）の頻度

なお、主成分分析（PCA）の結果を確認すると、プロットが密集し、つぶれてしまっています［図 3(a)］。これは、少数の外れ値のように見えるデータ点が原因です。PCA の可視化結果をもとに、外れ値の有無を推察することがありますが、PCA では線形なデータ構造を仮定しています。データによっては上記の仮定を逸脱することもあるため、実行時には注意が必要です。

そのほかの一般的な注意点としては、解析時に変数の数があまり増えすぎないように気を付ける必

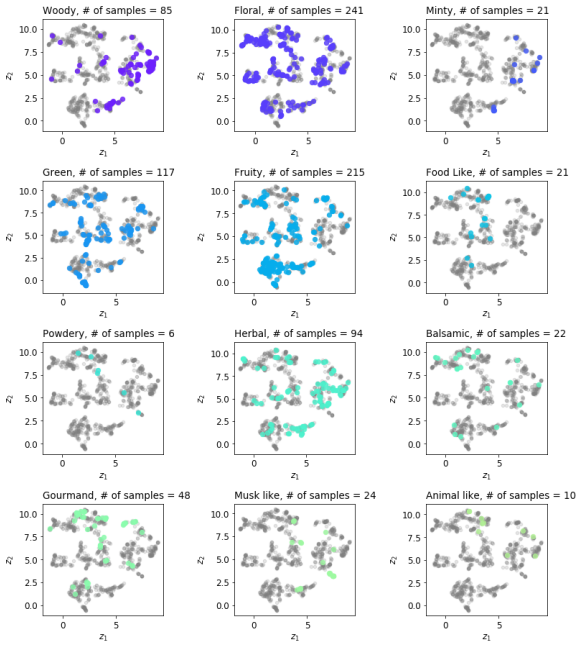

図 2　UMAP によるケミカルスペースの可視化結果（各色は官能評価ラベルに対応する）

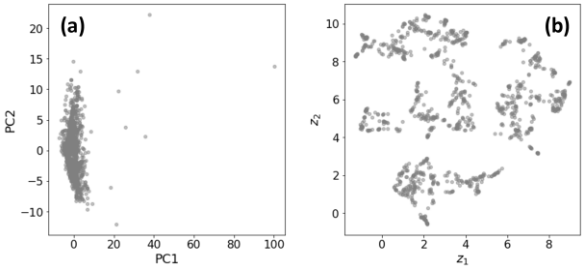

図 3　(a) PCA および(b) UMAP の可視化結果

地方独立行政法人
大阪産業技術研究所　本部・和泉センター
https://orist.jp/
〒594-1157　和泉市あゆみ野２丁目７番１号
Phone: 0725-51-2525（総合受付）

要があります。特に、RDKit などから算出される数値には欠損値や相関の高い変数が含まれることもあり、数多くの変数が一挙に出力されることもあります。多くの変数の中から冗長な情報を削除することで解析結果が改善されるかもしれません。

クラスタリング結果の可視化

次元削減と同様に、化合物にクラスタリングを適用することも可能であり、クラスタリングにより化学構造が類似する化合物群を推察することができます。化合物のクラスタリングの適用例としては、代替材料のスクリーニングが挙げられます。例えば、材料性能の観点からは好まれる化合物（原料）であっても、悪臭や原料コストといった機能とは別の理由で使用を断念せざるを得ないことも考えれます。そのような場合、代替物質候補をスクリーニングする上でクラスタリングは有用です。つまり、化学構造が類似する化合物同士では、両者の特性も類似するという「Similar-Structure, Similar Property Principle」が成り立つことを期待し、クラスタの抽出を図ります。

図 4 上に、階層的クラスタリングにより得られたデンドログラムを示します。なお、クラスタリング手法は階層的クラスタリングである必要はなく、そのほかの手法を用いることも可能です。ただし、手法によってはデンドログラムのような図は得られず、クラスタラベルのみが出力されることがあります。そのような場合であっても、可視化結果にクラスタラベルを反映することで、クラスタの分布を視覚的に確認することができます（図 4 下）。

Tanimoto 係数を用いた類似化合物の検索

保有する化合物データセットの中で、特定の化合物（クエリ分子）と化学構造が類似する分子を知りたいという場面があるかもしれません。ここでは、

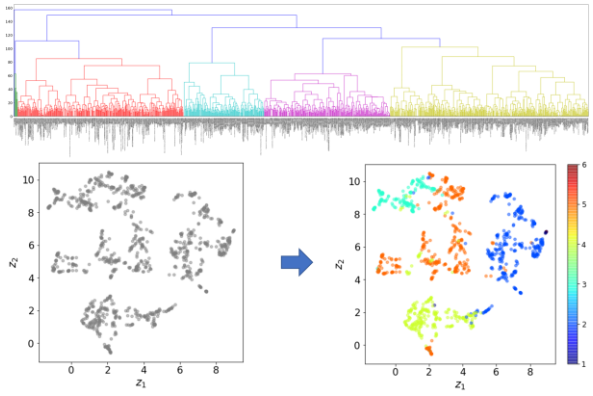

図 4 階層的クラスタリングおよび UMAP の実行結果

分子フィンガープリントにより算出した Tanimoto 係数 [1]を化学構造の類似度指標とし、その指標が大きくなる化合物を検索します。また、ここでのクエリ分子は、アリルプロピルジスルフィドであるとします（図 5 右上）。なお、類似化合物の検索において、計算に用いる分子フィンガープリントにより、検索結果が左右される点には留意すべきです。図 5 に類似度指標の上位 3 番目までの化合物を示していますが、2、3 番目の化合物はそれぞれ異なります。どのような化合物が類似度上位にランクするか確認しながら実行すると良いかもしれません。

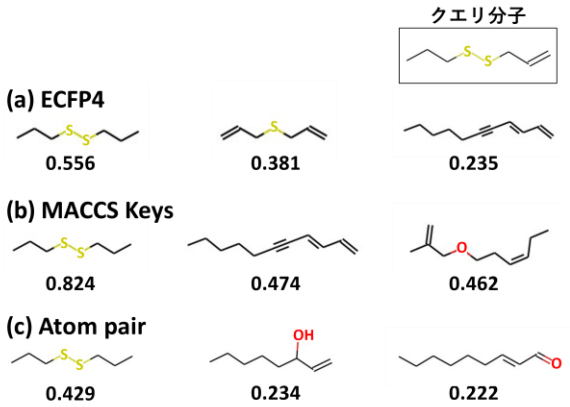

図 5 異なる分子フィンガープリントを用いた類似化合物検索結果（Tanimoto 係数を付記）

おわりに

本稿では、低分子化合物として臭気成分を取り上げましたが、匂いの感じ方は臭気成分の濃度や官能評価者の属する文化圏などにも影響されるため、解析時には注意が必要です。なお、本シート「データ解析入門 14」までで、主に可視化やクラスタリングなどの教師なし学習と呼ばれる解析手法について紹介してきました。次稿から材料開発などにおいても大きな役割を担っている回帰分析について技術紹介を行います。

参考文献

1) 永廣卓哉：データ解析入門 13 <化合物のデジタル表現>、ORIST テクニカルシート、No. 22-24 (2022)
2) https://github.com/pyrfume/pyrfume-data/tree/main/ifra_2019 (accessed on June 29th, 2022)
3) AI 創薬のための ケモインフォマティクス入門 https://github.com/Mishima-syk/py4chemoinformatics/blob/master/py4c.asciidoc (accessed on July 20th, 2022)

発行日　2023 年 1 月 1 日
作成者　高分子機能材料研究部　生活環境材料研究室　永廣卓哉
Phone: 0725-51-2611　　E-mail: ehirot@orist.jp