Technical Sheet
No. 22-24

データ解析入門 13 <有機化合物のデジタル表現>

キーワード： SMILES、MOL ファイル、分子記述子、分子フィンガープリント、Tanimoto 係数

はじめに

本稿では、有機化合物の情報処理に用いられる代表的なデータ形式を紹介します。化合物を数値化することで、一般的なテーブルデータなどと同様の解析が可能になります。

SMILES とは

有機化合物のデジタル表現には様々な方法がありますが（図 1）、その中でも SMILES（Simplified molecular input line entry system）は頻繁に用いられます。SMILES は、化学構造を一行の文字列で記述するための記法です。表 1 に SMILES 文字列の具体例を示します。標準的な SMILES 記法では、水素原子は省略されるため、水は「O」と表記されます。また、二重結合は「=」、3 重結合は「#」で表現します。環構造内の原子のつながりは数字で指定し、エチルベンゼンの 2 つの「c1」は芳香族炭素原子の結合を表します。また、分岐構造は「( )」を用いて表現します。

SMILES 文字列では、実際の原子の隣接関係を正確に表現していないことがありますが、化学構造の情報を保有しています。そのため、SMILES 文字列から化合物の物性予測などを行うための自然言語処理技術が研究されています。

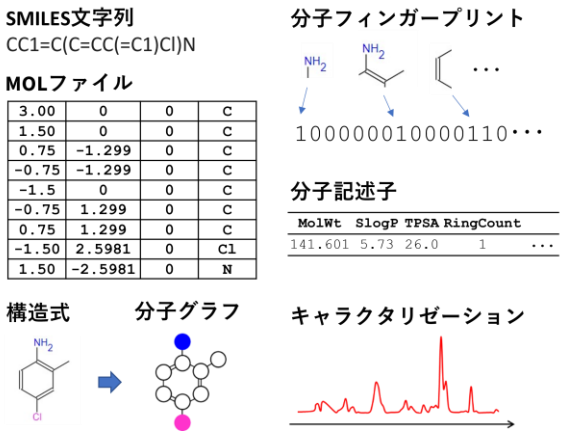

図 1 化学構造の様々な表現方法

表 1 SMILES 文字列の例

物質名	化学式	SMILES文字列
水	H_2O	O
シアン化水素	HCN	C#N
イソブタン	i-C_4H_{10}	CC(C)C
イソブテン	i-C_4H_8	CC(=C)C
エチルベンゼン	$C_6H_5CH_2CH_3$	c1ccccc1CC
アンモニウムイオン	NH_4^+	[NH4+]

SMILES にはいくつか種類があり、1 つの分子に複数の SMILES 文字列が対応することがあります。Canonical SMILES は分子構造に固有の SMILES を与えるための記法です。所定のルールに基づいた正規化処理により、統一的な SMILES 文字列を生成します。詳細な SMILES 記法の規則については文献 [1]が参考になります。

MOL ファイルとは

SMILES 文字列から化合物の物性予測などを行うことも可能ですが、SMILES 文字列を MOL ファイルと呼ばれる形式に変換後、物性予測などを実行することも一般的です。

MOL ファイルでは、原子の座標および元素記号が記載された原子ブロック、結合に関与する原子のインデックスおよびその結合の種類などが記載されている結合ブロックなどから構成されます。原子座標として 3 次元座標を指定することもできますが、2 次元座標が用いられる場合もあります。図 1 には、原子ブロックを例示しています。

また、SDF（Structure data file）と呼ばれるデータ形式は、複数の MOL ファイルを 1 つにまとめたものであり、こちらもよく用いられます。

SMILES 文字列から MOL ファイルへの変換などには、Python の RDKit[2]や R 言語の rcdk[3]などのライブラリを利用できます。上記ライブラリを用いることで、MOL ファイルを分子記述子と呼ばれる数値（ベクトル）に変換し、化合物の物性予測などのための説明変数を準備することができます。

分子フィンガープリントとは

化学構造に基づいて、分子の特徴を表現した数値を分子記述子といいます。記述子化のための手法は数多く提案されており、分子フィンガープリント（FP）と呼ばれる記述子は多用されます。FP は、分子中に存在する部分構造の有無や出現頻度をカウントすることで作成できます。官能基に代表さ

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）

れるように、分子の特性はその（部分）構造に依存します。そのため、分子中の部分構造を表現したベクトルは、物性などを支配する重要な情報として物性予測などに用いられます。これまでに様々な FP が提案されており、カウントする部分構造の種類やベクトルの長さなどは手法ごとに異なります。

ECFP（Extended connectivity fingerprint）は代表的な FP です。図 2 には注目する原子から半径 2 の範囲内の部分構造を調べる例を示しましたが、考慮する部分構造の大きさ（半径）や、生成するベクトルの長さには任意性があります。また、ECFP では解析対象は固定でなく、データごとに調査対象となる部分構造は変わります。

ECFP とは異なり、調査する部分構造があらかじめ指定されていることもあります。MACCS Keys と呼ばれるフィンガープリントでは、事前に決められた 166 種類の部分構造の有無を調べます。

ECFP や MACCS Keys は FP の一例ですが、この他にも様々な手法が提案されています。一般に、予測タスクごとに適した FP は異なります。そのため、複数の FP を併用し、変数選択を実行するということも実践的です。予測タスクによっては、分子中の局所的な部分構造だけではなく、大域的な特徴を捉えられる FP を併用することで予測精度の高いモデルを構築できるかもしれません。

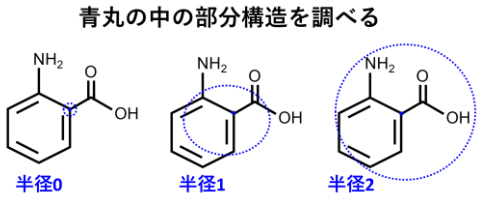

図 2 分子フィンガープリント（ECFP）の概念図

そのほかの分子記述子

FP では、部分構造の有無や出現頻度に注目することで化学構造をベクトルに変換しました。FP だけでも複数の種類がありますが、そのほかにも化学的特性や分子グラフなどに基づく記述子が数多く提案されています。

例えば、双極子モーメントや水溶性などの化学的性質に関する計算値が記述子として用いられています。より具体的には、オクタノール／水分配係数（logP）などの計算値が用いられます。化学構造に基づいて数値化される点は FP と共通しますが、連続値も特徴ベクトルの要素となることがあります。

また、分子中の原子のつながり方を表現した分子グラフに基づいた記述子も存在します。グラフ理論を化学構造に適用することで、分子のトポロジカルな情報を抽出します。なお、分子グラフに基づく深層学習モデルにより、特徴ベクトルを自動で抽出する試みもなされています。ただし、一般に深層学習では、多くの学習データが必要になります。

また、種々の分析やシミュレーション結果なども化合物の特性を反映しており、有用な説明変数になりえます。

化学構造の類似度

FP は予測モデルの入力データとしても有用ですが、Tanimoto 係数の算出などにも用いられます。Tanimoto 係数は 2 つの化学構造の類似度の指標であり、一般的には図 3 に示す式から計算されます。Tanimoto 係数が 1 に近くなるほど両分子は類似しているとみなします。

また、構造骨格中の一部の原子置換などの微細な構造変化にあまり影響されない Fraggle[4]と呼ばれる類似度評価手法も提案されています。

図 3 Tanimoto 係数について

おわりに

本稿では、化学情報処理に関する基本的なトピックスについて紹介しました。次稿では、低分子化合物を数値に変換し、これまでのテクニカルシートで取り上げた解析手法を適用します。

参考文献

1) Daylight Chemical Information Systems, Inc. Homepage, https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed on May 6th, 2022)
2) RDKit: Open-Source Cheminformatics Software, http://www.rdkit.org/ (accessed on May 23rd, 2022)
3) rcdk: Interface to the 'CDK' Libraries, https://cran.r-project.org/web/packages/rcdk/index.html (accessed on May 23rd, 2022)
4) https://github.com/rdkit/UGM_2013/blob/master/Presentations/Hussain.Fraggle.pdf (accessed on August 25th, 2022)

発行日　2022 年 12 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