正規表現拡張SMILESが示す化合物集合に対する部分構造探索

藤芳 明生^{*}, 坂本 雅志

茨城大学大学院理工学研究科情報工学専攻, 〒316-8511 茨城県日立市中成沢町 4-12-1

^{*}E-mail: akio.fujiyoshi.cs@vc.ibaraki.ac.jp

(Received: December 4, 2018; Accepted for publication: December 25, 2018; Online publication: February 16, 2019)

This paper proposes a substructure search method for finding an element of a set of chemical structures represented by a SMILES string with regular expression extension. It is known that substructure search is an NP-hard problem in general. However, by taking advantage of a graph-theoretical characterization of chemical compounds, the proposed method runs fast enough. Software for substructure search using the proposed method was implemented and released.

キーワード: Substructure search, Regular expression, SMILES, Graph automaton, Treewidth

1 はじめに

単一の化合物を簡潔に表現する方法には、IUPAC名、SMILES、InChI等、様々な物が存在するが、化合物の集合を表現するには困難が伴う。有限個の化合物の集合であれば、すべての要素を列挙することも可能であるが、無限個の化合物の集合を表現するにはどうすればよいのか。特許の請求項では、化合物の集合を表現するためにマルクーシュ構造(Markush structure)が広く用いられている。しかし、これには2次元的に描かれた図と自然言語による説明文が必要であり、コンピュータで処理するために適したデータ構造とはなっていない。グラフ理論では、単項二階論理(Monadic second ordered logic)が用いられることもあるが、基本的な集合でも非常に複雑な記述が必要となる。

本研究では、簡潔な文字列によって化合物の集合を表現する方法として正規表現拡張SMILESを考案するとともに、その化合物集合に対する部分構造探索手法を提案する。部分構造探索はNP困難な問題であり、一般的には高速な手法は存在しないことが予想されているが、化合物に共通するグラフ理論的特徴を捉えることにより、提案手法は十分高速に動作するものとなっている。正規表現の登場によって文字列検索の利便性が飛躍的に向上したように、正規表現拡張SMILESを利用することで化学構造検索の利便性を大幅に向上することが期待できる。また、情報秘匿のため、本当に検索したい化学構造

を曖昧にしたまま部分構造検索することも可能となる。

2 方法

正規表現拡張SMILESは、既存の分子構造表記法SMILES [1]に対し、波括弧「{」及び「}」による範囲指定、縦棒「|」による選択、アスタリスク「*」による0回以上の繰り返し、プラス「+」による1回以上の繰り返しを表現するように拡張した文字列である。例えば、「O=C (C) O, O=C (C) Cl, O=C (N) O, O=C (N) Cl」の4つの化合物の集合は、「O=C ({C|N}){O|Cl}」の1つの文字列で表現でき、シクロアルカンの集合は、「C1C{C}*C1」と表現できる。

提案手法による正規表現拡張SMILESが示す化合物集合に対する部分構造探索は、以下の手順で実現される。

1. 正規表現拡張SMILESをグラフオートマトンに変換する。
2. 部分構造探索をしようとするそれぞれの化合物の内部において、グラフオートマトンが受理する部分構造を探索する。

手順1で使用するグラフオートマトンは、この部分構造探索手法を実現するために新たに考案したものである。オートマトン理論において、グラフオートマトンの研究は乏しく、本研究に適したものは存在しなかった。一方、木構造を受理する木オートマトン [2]は古くから研究されており、標準的な概念として定着している。本研究では、この木オートマトンを拡張し、シンプルで扱

Figure 1. Compounds of a large treewidth

しやすいグラフオートマトンを得ることに成功した。前出のSMILES表記法は、化学構造中の環を成す結合を切り開き、グラフ理論でいう全域木を生成して、これを表現したものである。この方式に習い、化学構造の全域木を受理するようなグラフオートマトンを考案した。

手順2では、グラフオートマトンが受理する部分構造を探索するが、知られている通り、部分構造探索はNP困難な問題であり、一般的には高速な手法は存在しないことが予想されている。そこで本研究では、化合物に共通するグラフ理論の特徴として、木幅(Treewidth)に注目した。木幅とは、グラフの構造がどれくらい木構造に近いかということを表すグラフの特徴量である。木幅が小さいグラフに対しては、動的計画法を用いることで、部分構造探索をはじめとする多くのNP困難な問題も高速に解決できることが知られている。医薬品及び医薬品候補化合物データベース ChEMBL version 8 に登録されている 635,933 化合物の木幅を調べたところ、木幅1(木構造)が11,753個(1.8%)、木幅2(単純な環構造)が603,806個(94.9%)、木幅3が20,096個(3.2%)、木幅4が64個、木幅5が0個、木幅6が18個、木幅7が109個、木幅8が0個、木幅9が2個、木幅10が2個、木幅11以上が83個であった。木幅が9以上であった化合物は、Figure 1に示すような、バックミンスターフラーレン等、数種類の木幅が非常に大きな構造を部分構造に持つものに限られることが分かった。よって、これらを例外とすれば、木幅7以下の化合物に対して高速に動作する部分構造探索手法が得られれば十分であると考えた。

提案手法がグラフを処理する様子をFigure 2に示す。まず、グラフの全ての辺に、ハイパーエッジを割り当てる。各ハイパーエッジには、グラフオートマトンの可能

Figure 2. Movement of the proposed method

Figure 3. Compounds containing a tetracene as substructure

な状態遷移の集合を対応づける。そして、グラフの頂点を1つずつ消しながら、ハイパーエッジを結合していく。この時、結合前の「可能な状態遷移の集合」を組み合わせ、無矛盾なものだけが残るように結合していく。適切な順序で頂点を消すことにより、ハイパーエッジの最大サイズは元のグラフの木幅を超えないようにすることができる。最後に、ハイパーエッジは1つだけ残り、それに対応づけられた「可能な状態遷移の集合」が空でなければ、求める部分構造が見つかったことを意味する。

3 結果

提案手法に基づき、正規表現拡張SMILESをグラフオートマトンに変換するソフト及びグラフオートマトンが受理する部分構造を探索するソフトの実装を行った。このソフトウェアは、研究室のWebページ <http://apricot.cis.ibaraki.ac.jp/RegExSMILES> で公開している。

ソフトウェアの実行速度を検証するため、ChEMBL中の木幅が7以下の635,846化合物に対し、テトラセンとそれより大きなアセン類からなる化合物の集合を表現するグラフオートマトンを用いて、部分構造探索を行った。使用PCのCPUは、Intel core i5-5200U (2.20GHz)である。結果、605.3秒(平均0.95ミリ秒/個、最大0.46秒/個)でFigure 3に示す化合物を検索した。

4 まとめ

正規表現拡張SMILESが示す化合物集合に対する部分構造探索手法を提案し、実装を行った。医薬品及び医薬品候補化合物に共通するグラフ理論的特徴として、木幅 (Treewidth) に注目し、木幅が小さい化合物に対し高速に動作する部分構造探索手法を得ることに成功した。

参考文献

- [1] D. Weininger, *J. Chem. Inf. Model.*, **28**, 31 (1988).
DOI:[10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005)
- [2] H. Comon, et al., Tree automata techniques and applications, <http://tata.gforge.inria.fr/> (2007).