分子構造符号化と対称な部分構造の検出

RA 小市俊悟

情報理工学系研究科数理情報学専攻

概 要

分子構造符号化とは，分子の平面構造を線形表記 (文字列) によって表現することである．これによりデータベース容量を抑えたり，よく似た分子をデータベースから検索したりすることが，文字列の二分探索により可能になるなどの利点がある．そのような利点を最大限に生かすためには平面構造と 1 対 1 に対応する線形表記の生成法が必要である．そのような線形表記を規範的線形表記と呼ぶ．本研究では多項式性はないものの，実用的には十分高速に規範的線形表記を生成するアルゴリズムを開発した．さらに，そのアルゴリズムを応用し，分子中の対称構造を検出するアルゴリズムを開発した．

1 はじめに

本研究で開発した分子構造符号化法は (改良) CANOST 表記法として，^{13}C-NMR 化学シフト値予測システム CAST/CNMR[1, 2] で使用される．CANOST 表記法は分子の平面構造を元に分子グラフを作り，その分子グラフを表す規範的線形表記を出力する．対応する原子を含む官能基の種類によって，分子グラフの各頂点には CANOST コードがふられる．例えば，図 1-(a) の分子であれば，その CANOST 表記は図 1-(b) のようになる．C1,Q,H などが CANOST コードである．線形表記を求めることは，分子グラフの頂点に番号を付けることに等しく，特に規範的線形表記を求めることは，グラフの問題として，グラフの頂点の標準的番号付けを求めることに等しい．この問題はグラフ同型性判定問題を含むために，一般のグラフに対して多項式時間アルゴリズムは知られていない．そこで，CANOST 表記法は対象が分子グラフという点に着目し，実用的に高速なアルゴリズムであることを目指して設計された．

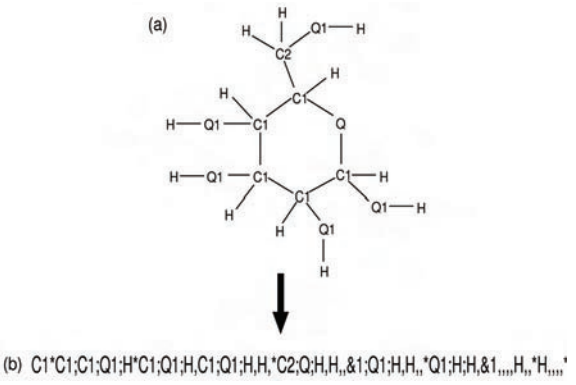

図 1: 分子グラフ (a) とその CANOST 表記 (b)

2 CANOST 表記法

CANOST 表記法では，以下の処理を分子グラフ中の各頂点を始点として行う．まず，始点に指定された頂点から幅優先探索を行い，探索されるまでの始点からの枝数が等しいものを一つの層とする層別ネットワークを構築する．

層別ネットワークの各頂点が属す層は，その定義から一意的に決まる．したがって，各層に関して，頂点の番号が決まればよい．あらかじめ決められた CANOST コードの優先順位を元に，頂点の番号を頂点分類型アルゴリズムと呼ばれる枠組に入る手法を層別ネットワークに対し，繰り返し適用することで，同じ番号であるような頂点の多い粗い番号付けから，より細かい番号付けを求める．一般には，頂点分類型アルゴリズムの繰り返しの後も，同じ番号であるような頂点が残る．と

ころが，それでは線形表記を生成することが出来ないので，同じ番号である頂点に相異なる番号を付けることが必要となる．注意すべきは，その番号の付け方によって生成される線形表記が異なることである．したがって，任意に一つ番号付けを決めるのでは，等しい分子であっても異なる線形表記になることが起こり得て，分子の検索において網羅性を欠いてしまう．そこで，同じ番号である頂点に対する番号付けを列挙する．列挙された番号付けは，それぞれ一つの線形表記と対応するので，列挙された番号付けの数だけ線形表記を得る．列挙する番号付けの個数は一般に多項式オーダとならないので，計算量を多項式オーダに抑えることができない．

列挙された線形表記の中で辞書式最小のものを選べば，それは始点となる原子を指定されたときに一意的に決まる線形表記である．各原子を始点としてそれぞれ線形表記を求め，さらにその中から辞書式最小のものを選べば，それが規範的線形表記となる．

列挙する番号付けの個数を減らすことが計算量の削減となる．新たに開発したアルゴリズムの性質から，同じ番号である頂点集合のうち，ある条件を満たす部分構造に含まれているものは，全ての番号付けを列挙しなくても，一つの適切な番号付けにより，列挙したときに生成され得るものが網羅される．そのような部分構造は容易に見付けることが可能であって，これにより計算量を大幅に削減することが出来る．言い換えれば，そのような列挙しなくてもよい構造というのは，対称な部分構造を含んでいる．そのような対称な部分構造を検出することは容易である．

3　対称な部分構造の検出

始点となる原子が指定されたとき，一意的な線形表記を得るために，候補となる線形表記，すなわち，番号付けを列挙するが，その列挙された番号付けを利用することで対称な部分構造を検出することが可能である．辞書式最小の線形表記は一意的に決まるが，同じ線形表記を与える番号付けは複数あり得る．このとき，辞書式最小な線形表記を導く番号付けの中に，ある二つの番号付け L_1, L_2 があって，異なる二つの頂点 a, b が a は L_1 で，b は L_2 である同じ番号になるのならば，その二つの頂点を同じ組に分類するということを行う．すると，これは同値類分解であって，各同値類の頂点に対して，それらをそれぞれ含むような対称構造が存在することがわかる．このようにして対称な部分構造を検出する．図 2 はこのようにして得られる対称な部分構造である．

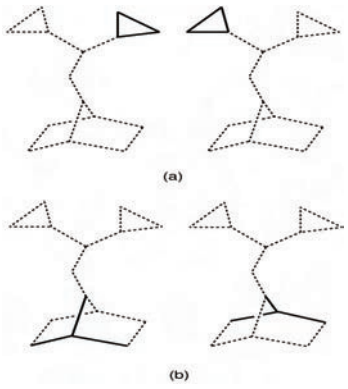

図 2: 検出される対称な部分構造 (a),(b)(実線部)

4　おわりに

CAST/CNMR システムの予測法では，分子検索の正確性が必要不可欠であり，本研究の成果はそれを支援する．また，NMR 化学シフト値を用いた解析では，対称構造の検出は非常に重要であり，本研究の成果はそれを支援する．

参考文献

[1] H. Satoh, H. Koshino, J. Uzawa, and T. Nakata: CAST/CNMR: Highly Accurate ^{13}C NMR Chemical Shift Prediction System Considering Stereochemistry. *Tetrahedron*, 59/25 (2003), pp. 4539–4547.

[2] H. Satoh, H. Koshino, T. Uno, S. Koichi, S. Iwata, and T. Nakata: Effective Consideration of Ring Structures in CAST/CNMR for Highly Accurate ^{13}C NMR Chemical Shift Prediction. *Tetrahedron*, 61(2005), pp. 7431–7437.