化学構造データベースからの有効な部分構造抽出法に関する考察

田中　栄太朗† 津田　哲夫† 吉澤　有美† 稲積　宏誠†
青山学院大学理工学部†

１．はじめに

今日、大量の化学物質が世の中に存在するが、その全ての物質について化学的性質を調査することは、膨大な時間と費用がかかるため大変困難である。そこで、既存の化学データベースから共通部分構造を抽出し、性質との関係を分析することで、調査対象の絞り込みや新規化学物質合成への手がかりとすることができれば大変有意義である。

対象の各要素（ノード）間の関係（リンク）構造をグラフ構造と捉え、その中からそれぞれに共通する有効な部分グラフを抽出しようとするアプローチとして、グラフマイニングがある。本稿は、有向グラフ構造を前提としたデータマイニング手法である GBI(Graph-Based Induction)の考え方[1]を、無向グラフで表現される化学物質からの共通部分構造抽出アルゴリズムへ拡張する。さらに、抗菌活性物質データ[2]を用いた実験を通して抽出した共通部分構造の組み合わせにより各化学物質を表現し、各種知識発見手法のための基礎情報として提供することを検討する。

２．GBI 法に基づく部分構造抽出法

有向グラフ中に頻繁に現れるペア（要素対）の逐次抽出を繰り返すことにより類型的なパターン抽出を行う GBI 法に以下の改良と拡張を行い、無向グラフである化学物質構造への適用を実現する。

2.1.ペア情報文字列

化学物質構造のペア情報を、４項組の文字列として記述する。第１項・第２項は[元素記号/チャンク内結合 ID]、第３項はリンク情報（６種類の化学結合：A～F）、第４項は自己ループ・対称構造判別フラグを示す。また、各要素対を１つの要素に置き換えることをチャンクと呼ぶ。その際に付けられるラベルと他ノードに結合している初期要素の対をチャンク内結合 ID とする。その結果、この ID をソートすることで無向グラフ構造のペアを一意に決定づけることができる。図１にベンゼン環のチャンク過程を示す。

Finding Effective Substructures of Molecules from Chemical Database
† Eitaro TANAKA, Tetsuo TSUDA, Yumi YOSHIZAWA, Hiroshige INAZUMI
School of Science and Engineering, Aoyama Gakuin University

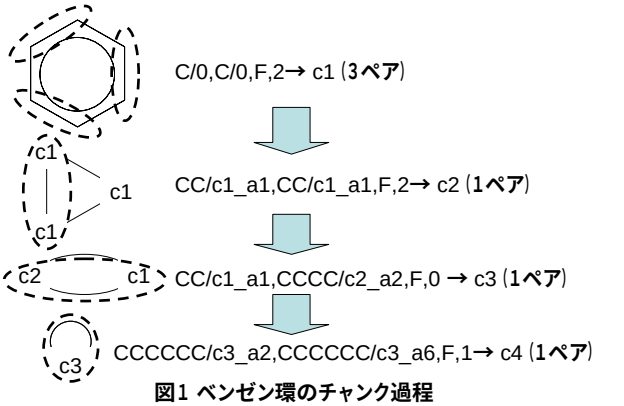

図1 ベンゼン環のチャンク過程

2.2.対称構造の識別

チャンク対象ペアの両端ノードが同一構造の場合には、チャンク後に対称な構造となってしまう。その結果、無向グラフ上では同一の構造でありながら複数のペア情報文字列が生成されるため、異なる構造として扱う場合が生じる（図２）

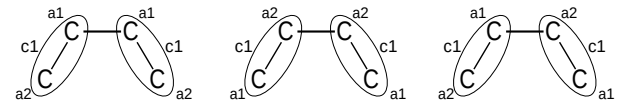

図2 3通りのペア情報文字列

そこで、同一構造同士をチャンクする際にはチャンクノード内に割り当てる ID をいったん区別するが、チャンクに影響を及ぼさないよう大文字・小文字表現を用いることとする。この場合も図３のように３通りのペア情報文字列が存在するが、ペアを数え上げる際には小文字表現に統一(CC/c1_a1,CC/c1_a1,F,2)することでこれらを同一構造とみなすことが可能となる。また、図３のペアも対称構造なので、チャンク後には図４のように ID が割り振られる。

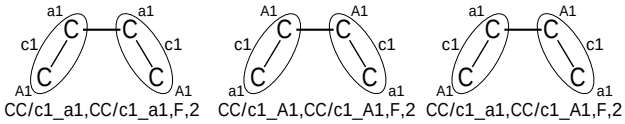

図3 大文字･小文字表現

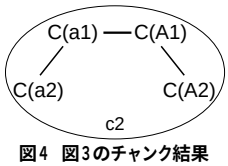

図4 図3のチャンク結果

３．グラフ同型処理

GBI 法では逐次的にペア情報を拡大していくため、異なるチャンク過程から同一の構造が抽出されることがある。ところが、GBI 処理の上では異なる部分構造と解釈されるため頻度情報が分散し、ペア数え上げにおいて不具合が生じる。

実験で用いた抗菌活性物質におけるチャンク過程の一部を例にあげて、グラフ同型処理によりチャンク処理が変化するようすを示す。図５において、チャンクノード 14 と 23 が同一構造であることがわかる。そこで、チャンク処理ごとにそれ以前に抽出した全ての部分構造とグラフ同型比較を行うこととし､同一構造への統合処理を実現する。

図６の左側が同型処理を行わなかった場合のチャンク過程、右側が同型処理を行った場合のチャンク過程である。チャンクノード 14 と 23 はグラフ同型であるが、統合処理を行うことにより、

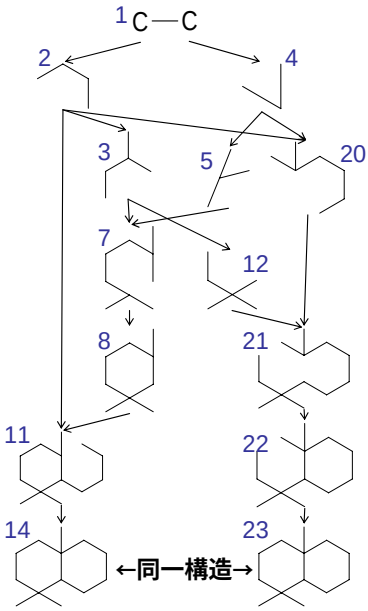

図5　異なる過程から抽出される同型部分構造

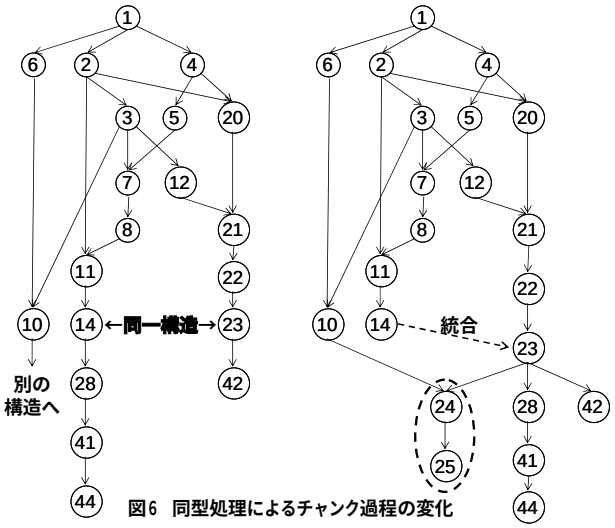

図6　同型処理によるチャンク過程の変化

チャンクノード 23 と他のノード（この場合はチャンクノード 10）とのペアの頻度が上がり、同型処理を行わなかった場合には現れなかった新たな部分構造（チャンクノード 24,25）が抽出されることが示される。

４．部分構造の組合せによる化学物質の表現

本稿では、GBI 法の逐次的探索の結果得られる処理過程に加えて、求められた部分構造間の包含関係を抽出することを提案する。その結果、構造情報をもつデータベースから一定頻度以上の部分構造と、その部分構造間の包含関係を抽出することができる。

以上の改良や拡張により抽出された共通部分構造の有無により、各化学物質を表現することが可能となった。抗菌活性物質群（フラボノイド類 325 種・ジテルペン類 117 種）を XML 仕様である CML(Chemical Markup Language)データベースとして用意し、提案手法を適用した。その結果、部分構造の組み合わせとデータベースに存在する抗菌活性値などの非構造情報を付加することで、決定木などの各種知識発見手法の入力データとして使用可能な出力形式を実現した。

表1　抽出部分構造の包含情報と非構造情報

化学物質名	c1	・・・・	c11	c12	c13	・・・・	c32	mic
3 ', 4 '-Dihydroxyflavone	1	・・・・	1	0	1	・・・・	0	64
6, 7- Dihydroxyflavone	1	・・・・	1	0	1	・・・・	0	64
7, 8- Dihydroxyflavone	1	・・・・	1	0	1	・・・・	0	32
Sikonin	1	・・・・	0	0	1	・・・・	0	8
Alkannin	1	・・・・	0	0	1	・・・・	0	4
Lawsone	1	・・・・	0	0	0	・・・・	0	128

構造情報
各チャンクノード(部分構造)が
含まれている場合:1
含まれていない場合:0

非構造情報
ex.抗菌活性値

５．結論

本稿で提案したペア情報文字列表現と対称構造判別により、無向グラフに対しても擬似有向グラフ化することなしに GBI 法を適用することができた。またグラフ同型処理により、より正確な部分構造の抽出が可能となった。さらに、部分構造間の包含関係を抽出することによって知識発見手法への基礎情報を提供することができた。

今後の課題として、逐次探索の欠点を補うためにチャンクの並列処理などのさらなる改良方法を検討することがあげられる。

参考文献

[1] 松田喬、元田浩、鷲尾隆：”一般グラフ構造データに対する Graph-Based Induction とその応用”,人工知能学会論文誌,Vol.16 No.4 A (2001)
[2] 稲積宏誠、野中健一、吉澤有美、木村純二：“抗菌活性物質におけるデータマイニング”, 日本化学会第 81 回春季年会, 1 PC169, p.572 (2002)