

化学物質の性質を決定する特徴的な部分構造発見の試み

○ 田中栄太郎、速水亜希子、稲積宏誠

青山学院大学理工学部 (〒229-8558 相模原市淵野辺 5-10-1)

【目的】

既存の化学データベースから共通部分構造を抽出し、物質の持つ性質との関係を分析することは、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なる。一方、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。本稿では、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の拡張方法を示す。さらに、共通部分構造の組み合わせにより各化学物質を表現し、抗菌活性物質データに対して基本骨格に基づいた決定木による解析実験[2]を進展させ、決定木解析の応用についても検討する。

【概要】

GBI 法そのものの改良は、さまざまな角度から取り組まれている。本発表では無向グラフで表現された対象に対して適用可能なアルゴリズムと、異なる抽出過程で同一構造が生じた際にそれを認識するための簡易手法を示し、これにより探索精度の向上が図られることを確認する。無向グラフにおける解析では、対称構造や鏡像体などに対する処理が必要となるが、これについてはノード情報とリンク情報の記号表現の改良により誤認識の確率を大幅に減少させた。同一構造であることを確認するためには、グラフ同型問題を解決しなければならないが、ノードとリンクの分布情報を用いてグラフ同型処理を実現することによって、比較的短時間で近似的にグラフ同型を判断することができ、探索の補正を実現した。さらに、部分構造の関係や組合せがその物質の特徴をどのように決定づけられるのかについて検討し、GBI 法の部分構造選択基準のなかに特徴を示すクラス分類能力を組み込むアプローチと、機械学習における代表的な手法である決定木分析によるアプローチを提案した。特に、単独の部分構造ではなく、それらの組合せがどのように物質の特徴に反映するかについて検討するために、決定木の属性情報に対する前処理方法を提案した。

以上の結果、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。さまざまな部分構造と、部分構造間の包含関係により各物質を表現し、部分構造の組合せかたによって、高活性を示す要因となる部分構造に関する有効な規則を導くことを試みた。表 1 は、改良 GBI 法により得られる部分構造の組合せ形式を示している。また図 1 と 2 は 2 種類の決定木解析で得られたルールの例を示している。

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
各チャックノード(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

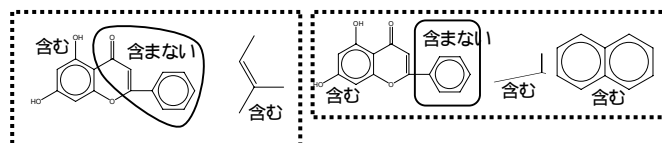


図 1 C5.0 系列情報によるルール

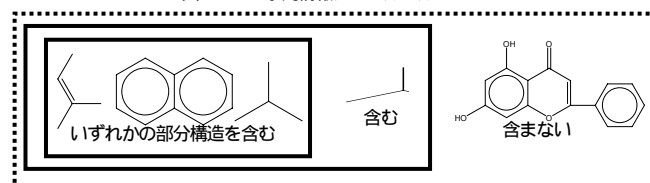


図 2 DTMACC 構造情報によるルール

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, pp.363-374(2001)
- [2] 稲積宏誠, 吉澤有美, 野中健一, 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)