1H4-1

化学物質の構造特徴解析と化学クラスの自動識別

Automatic Identification of Chemical Categories of Chemical Substances Based on Substructure Feature Analysis

岩元 あすみ[*1] 高橋 由雅[*1]
Asumi IWAMOTO Yoshimasa TAKAHASHI

[*1] 豊橋技術科学大学大学院工学研究科 情報・知能工学専攻
Department of Computer Science and Engineering, Toyohashi University of Technology

Generally, chemists do various intellectual reasoning by looking at a chemical structure formula. This paper describes an approach to automatic identification of chemical categories or chemical classes of chemical compounds based on substructure feature analysis similar to that chemists do so it. In the present approach, first, a chemical structure is submitted to a chemical intelligence tool through a structure editor. Second, substructure feature analysis is carried out for the chemical structure. A knowledge file of substructures is employed for the feature analysis. It is followed by the procedure of automatic identification of the chemical category of the chemical substance. The detail of the approach is discussed with an illustrative example.

1. はじめに

一般的に、化学物質の薬理活性や毒性の予測、化学構造のプロファイリングなどをする際、化学者は化学構造式を見ながら様々な化学的類推を行っている. このことから、化学構造式を理解し、その意味を解釈する化学人工知能を実現するためには、その基盤技術として化学構造特徴を自動的に認識する技術が不可欠となる.

そこで本研究では、提示された化学構造式について、その構造特徴を自動的に解析し、知識ベースをもとに、対応する化学クラス(化合物タイプ)を自動識別するための基盤システムの開発を試みる.

2. システムの概要

本システムの入力は化学構造である. 化学構造式の入力には当研究室で別途作成した構造式エディタを用いる. 入力された化合物構造に対し、知識ベースに格納されている各々の部分構造との部分構造マッチングを行い、化学構造の構造特徴解析を行う. 本システムの基本構成概念を図 1 に示す.

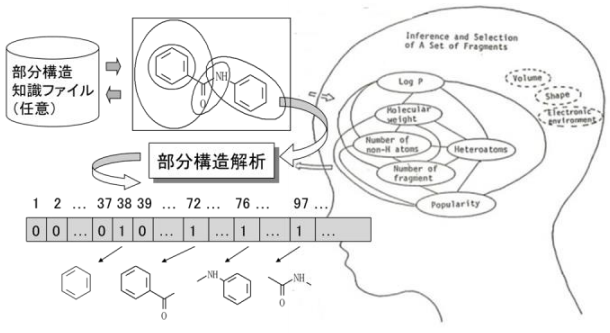

図 1 構造特徴解析システムの基本概念

連絡先:岩元 あすみ, 豊橋技術科学大学 情報・知能工学専攻, a_iwamoto@mis.tut.ac.jp

3. 構造特徴解析

構造特徴解析に際しては事前に定義された部分構造知識ファイルを参照しながら、部分構造マッチングの技法により、注目する部分構造の有無やその数を調べ、これらの結果を一時保存する. 得られた部分構造特徴解析の結果は後述の化合物クラスの分類・同定に利用される. ここでの基礎となる部分構造マッチングはグラフ間の部分同型判定の問題に帰着できる.

3.1 部分構造マッチング

本研究での部分構造マッチングには Ullmann[1]のアルゴリズムを用いた. Ullmann のアルゴリズムは、代表的な部分グラフ同形判定アルゴリズムの 1 つである. Ullmann のアルゴリズムは探索木を巡回する探索木巡回アルゴリズムと、探索空間を削減する Refinement Procedure の 2 つから構成される.

Ullmann の探索木巡回アルゴリズムは、深さ優先探索である. 部分グラフ同型判定を行う 2 つのグラフを$G_A = (V_A, E_A)$, $G_B = (V_B, E_B)$とする. また、それぞれの頂点数を n, m とする. このとき、n×m の行列M = $[m_{ij}]$を定義する. M の各要素は 0 か 1 の値を持つ. $m_{ij} = 1$のとき、頂点v_{A_i}から頂点v_{B_j}への写像において、部分グラフ同型になる可能性があることを示す. M は「各行は 1 つだけ 1 を持ち、残りは 0」、「各列は 2 つ以上の 1 を持たない」という性質を持つ. 行列 M の初期値$M^0 = [m_{ij}^0]$の各要素の値は、

$$m_{ij}^0 = \begin{cases} 1 & deg(v_{A_i}) \le deg\left(v_{B_j}\right) \\ 0 & otherwise \end{cases} \quad (1)$$

上記において、$deg(v_{A_i})$はG_A中の i 番目の頂点v_{A_i}の次数を表す. M^0を M の性質に合うように、深さ優先で探索していく.

Refinement Procedure では、同型可能性判定を行う. Refinement Procedure において部分グラフ同型になる可能性がないと判定された場合は、その接点以下の部分木は切り捨てられ、探索空間が削減される. Refinement Procedure によるオーバーヘッドは増えるが、探索空間を削減したことによって、部分グラフ同型判定全体の処理時間は大幅に軽減される.

部分グラフにおける隣接関係は、入力された部分グラフにおいてもその関係性は保たれているべきであり、G_Aの隣接行列を

A、G_Bの隣接行列を B とすると、次式の関係が成り立つ.

$$\forall k\left(A_{ik}=1 \Rightarrow \exists p\left(M_{kp}B_{pj}=1\right)\right) \quad (2)$$

この条件に従わない場合は、$M_{ij}^{d}=0$とする.

このとき、得られた写像 M についえて次式の条件を満たすとき、2 つのグラフは部分グラフ同型であると判定される.

$$\forall_{i,j},\ A_{ij}=1 \ \Rightarrow\ C_{ij}=1 \qquad (3)$$

ただし、$C=M(MB)^{T}$

3.2 部分構造知識ベース

本システムに化学構造を入力し、解析した結果を図 3 に示す. なお、今回使用した知識ベースには、100 種以上の代表的な官能基や特徴的な部分構造情報が収録されている. それらの一部を表 1 に示す. 化学構造情報の計算機表現にはすべて独自の仕様にもとづく結合表を用いている.

表1 部分構造知識ファイルに官能基や部分構造(例)

番号	部分構造	基名
0001	−OH	hydroxy
0002	$-OCH_3$	methoxy
0003	−COOH	carboxy
0004	−CHO	fornyl
0005	−SH	mercapto
0006	−CHS	thiofornyl
0007	−COSH	mercaptocarbonyl
0008	−CSOH	hydroxythiocarbonyl
0009	$-SO_2H$	sulfino
:	:	:

番号	部分構造	基名
0151		pyridine
0160		oxazole
0269		biphenyl
0271		benzimidazole
0372		phenanthrene
:	:	:

4. 化学クラスの分類・同定

部分構造特徴解析の結果と構造分類に関する化学知識をもとに、入力された化合物の化学クラスを同定する. ここでは、構造分類に関する化学的知識は前述の部分構造知識とは別の知識ベースとして実装している. 構造分類のための分類規則の例を下に示す.

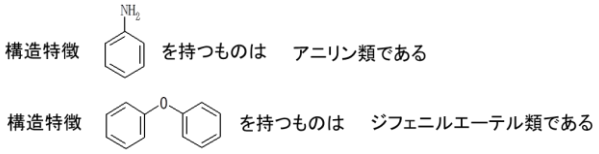

図 2 構造特徴にもとづく化学クラスの分類知識(例)

現在、ケーススタディとして、多様な化学構造を有する一般化学物質の生態環境毒性の予測問題の中で米国環境保護庁が提案している 112 種類の化学クラスに対する構造分類の自動化を進めているところである.

5. システムの実装と実行例

本研究で開発した構造特徴解析にもとづく化合物クラスの自動同定のための試作システムは、化学構造エディタ、部分構造特徴解析モジュール、部分構造特徴からの化学クラス(化合物クラス)の分類・同定モジュールの3つの主要モジュールから構成される. 試作システムの主な処理の流れを図 3 に示す. 試作システムは PC(mouse computer、Windows7.0 COREi7)上でプログラミング言語 C#を用いて開発を行った. 作成したシステムの実行画面例を図 4 に示す.

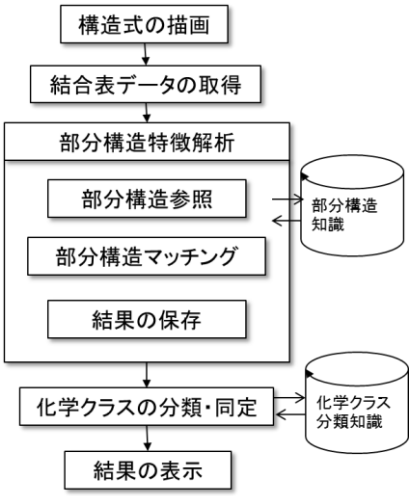

図 3 化学クラスの分類・同定処理の流れ

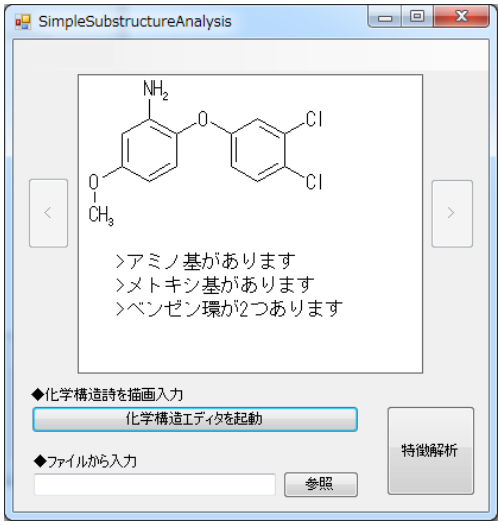

図 4 試作システムの実行画面例

化学構造式を化学構造エディタで作画、もしくはファイルから入力し、特徴解析ボタンをクリックすることで、部分構造知識ファイルにもとづく構造特徴解析が開始される. 解析の結果として、知識ベースに格納されている官能基のうち、部分グラフ同形であると判定された官能基に関する情報が出力・表示される.

6. まとめ

化学人工知能の実現のために必要となる基盤ツールの一つとして、知識ベースをもとに、入力された化学構造の部分構造特徴の自動解析プログラムを開発するとともに、同機能を利用した化合物クラスの自動分類・識別のためのシステムを試作した. 今後の課題としては、まず第一に、知識ベースの充実が挙げられる. また、知識ベースの充実を図るとともに、構造プロファイリングや類似性解析、さらには別途開発を進めている毒性予測システムの高度化などへの応用についても検討していきたい.

参考文献

[1] J. R. Ullmann: *An Algorithm for Subgraph Isomorphism*. *Journal of the Association for Computing Machinery*, Vol.23, pp.31-42 (1976).