

5 E - 5

筑波大学電子・情報工学系

3 有機合成情報の構造化

有機合成情報は化合物についての情報と化学反応についての情報からなると考えられる。今まで開発されたシステムには反応規則を記述するエキスパートシステムと反応事例を記述するデータベースシステムがある。しかし、エキスパートシステムにおいては反応規則の収集、体系化が困難であり、データベースシステムにおいては収録されていない反応に対処できないという問題点がある。本研究では、大量の反応事例を集めて情報の構造化手法で化合物の知識及び反応規則を獲得し、それに基づいて類推などを利用して新しい問題を解決するというアプローチである。

原情報： 本研究の原情報は5万件以上の反応例が収録されている反応データベースで、主な内容としては有機反応の出発物と生成物の構造を表す結合表と反応条件（試薬、触媒、収率など）を表す要約データである。

情報の構造方式は意味のレベルに従って次の三つの構造から構成される [2]。

- 情報構造化のプロセスはまさに機械学習のプロセスである。ここでは情報の自己組織化 (Self-organization) 手法を用いる。自己組織化とは本来生物の持っている機能であるが、ここでいう情報の自己組織化とは、情報資源に内在する意味的関係を構造化情報として抽出し、それを用いて情報資源全体を自動的に組織化するということを指す。情報の構造化とは繰り返して行なわれるプロセスでもある。即ち構造化された情報を利用して、新しい情報の構造化或いは更に高いレベルの構造化を行なうことである。構造化の結果はある領域の知識をカバーできるネットワークである。

Quit					
Distance	500062	500041	500042	500067	500003
C_Broad					
C_Narrow	500006	500075	500028	5000451	500098
Weight					
	500171	5002513	5000120	5000459	500020
					
	5008303	5008304	5008306	5008308	500
					

圖一 1 化合物概念構造

化合物概念構造の構築：化合物の性質は基本的には分子構造によって決まる^[3]ので化合物の類似性は分子構造或いは原子結合の類似性に依存する。つまり、化合物が似た構造を持てば似た性質を持ち、その逆も言える。一般的にいうと類似は概念間の重なりを示す関係なので類似の検出と測度は当然重なりを検出と測度に依存する。これで、本研究では化合物の分子構造間の包含関係（グラフ間のサブグラフ関係）によって概念構造を構築した。

構築された概念空間の一部を図-1で示す。この構造が化合物を表現するノードとノード間の包含関係を表現するラベル付きリンクから構成される。ラベルにはノード間分子構造の違いが明記されている。

有機合成は新しい化合物の合成を目標とするため本質的には閉じた世界におけることではない。化合物の概念構造は今まで集めてきた化合物だけではなく、仮想化合物までの拡張が要求される。これは研究者の思考活動を支援できるシステムの基本機能である。そのために、本研究の概念構造システムでは仮想化合物の埋め立て機能が実現された。当然、実化合物と仮想化合物を区別する必要がある。

化学反応概念構造の構築：化学反応の概念構造は反応の類似関係を表現する構造である。これは構造の変化（出発物対生成物）及び反応条件などによって構築される。構造の変化からは反応パターン（ $\text{CH}_3 \rightarrow \text{CHO}$ 、 $\text{NO}_2 \rightarrow \text{NH}_2$ など）を抽出して、さらに反応タイプ（酸化反応、還元反応、置換反応など）に分類できる。反応条件からは触媒、試薬などによって別の視点から反応の分類がされて概念構造が構築される。

論理構造の構築：論理構造は化合物間の反応関係を表現する構造である。論理構造の構築は二つの段階に分けている。まず原情報中の論理関係 出発物 $\Rightarrow$ 生成物 から化合物の同定によって構造を構築する。次に類推（次の節に述べる）によってこの構造を拡張していく。概念構造と同様に類推によって作られた部分は仮想構造なので区別する必要がある。尚、類推の根拠に基づいてトリガを設置しておくべきである。これで新しい情報による仮想構造を自動的に見直すことが可能になる。

4 構造化された情報における類推の実現

問い合わせに対して、まず論理構造において、入力された出発物と生成物の間に経路があるかどうかをチェックする。あればこの経路及び反応条件を出力し、無い場合には類推を行なう。

類推の手順：

(1) 入力された出発物と生成物の分子構造の変化から反応パターンを認識する；

(2) 反応の概念構造から類似反応パターンの反応をアクセスする（類似反応候補）；

(3) 類似反応候補と目標反応（類推しようとする反応）の出発物と出発物、生成物と生成物間の類似度を計算する；

前述したように化合物の類似度が分子構造間の重なり依存するので化合物間の類似度は基本的には化合物の概念構造における距離（ノード間パスの長さ）の関数とする。つまり、パスの長さが短いほど重なりが多く類似度が高い。但し、パスが同じであっても、着目する反応の種類によって化合物間の類似度が変化するので概念構造のトポロジカルな距離だけでは不十分である。つまり、

構造の異なりの意味と程度を考慮する必要がある。具体的に言えば、反応パターンに対応して概念構造のラベルに重みを付けて化合物間パスの長さとラベルの内容との組合せによってより詳細な類似度を測ることである。

(4) 類推結果の評価：適切な類推を行なうために領域の対象間で成立する事実や規則が対応している点を視点として取らなければならない。化学反応においては反応パターンの類似度が最も重要な評価基準である。次は反応条件（試薬、触媒など）についての考査も必要である。推論の段階と違って、評価する時には目標と候補間の類似性を考査するのみならず、候補間の類似性も考査するのが特徴である。

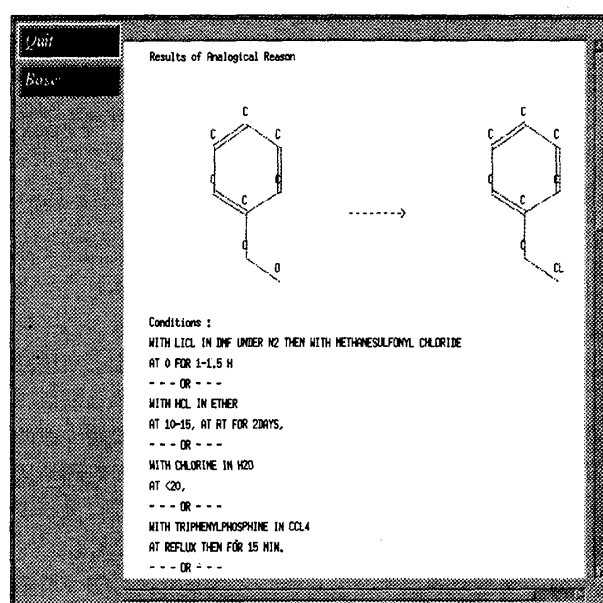


図-2 類推の例

5 むすび

自己組織化により情報の構造化アプローチについて述べた。有機合成反応を例として構造化された情報空間における類推の手順を示した。類推の評価については別の機会に報告する。

参考文献

- [1] Patric H. Winston: *Learning and Reasoning by Analogy*, Communications of the ACM, Vol.23, No.12, 1980, 689-703.
- [2] Z.Q. Wang, Y. Fujiwara et al: *Learning and Reasoning in the Information-Base System for Organic Synthesis Research*, Journal of Japan Society of Information and Knowledge, Vol.2, No.1, 1991, 71-82.
- [3] John M. Barnard: *Structure Searching Methods: Old and New*, J. Chem. Inf. Comput. Sci. Vol.33, No.4, 1993, 532-538.