

# 化学反応設計のための知識エディタ

Knowledge editor for chemical synthetic design

桑 原 敏

Satoshi Kuwahara

\* N T T 情報通信処理研究所

NTT Communication and Information Processing Laboratories

Recent advances in Knowledge engineering technology have resulted in a rapid progress of chemical compound design expert-system. This paper describes a design overview of the knowledge editor used in the chemical compound design expert-system which contains two design paradigms, traditional logic base design mechanism, and information oriented design method using fact data bases.

In this editor, two new technologies are implemented as the above information oriented design method. One is the chemical synthetic equivalent substitution and another is the knowledge acquisition from human experts by using analogical inference which mechanism is based on the similarity value between two characteristics of design objective compound and existing one in fact data-bases.

A graphical interface between human expert and design system is also introduced, and is very effective for synthetic knowledge acquisition from chemical experts and fact data-bases.

Several small synthetic experiments were tried under the cooperation with chemical experts. Although the kinds of chemical reaction treated in this prototype-system were limited to fragment substitution on the single aromatic ring, the essential mechanisms were confirmed.

Finally, in order to enhance this system toward supporting other complex compounds, more effort should be taken for establishing the perception algorithm such as an analyzing fine structure and characteristics of molecular, and connecting with large scale fact data-bases.

## 1. はじめに

科学技術庁では1986年から5か年計画で複数の機関を集めて、「化学物質設計のための知識ベース技術の研究」を進めており、NTTはその一環として知識ベースシステム構築の基本的ツールとしてエキスパートシステム構築支援ツールKBMS (Knowledge-Base Management System)を利用した共通基盤システムの開発を分担してきた。この共通基盤システムは各機関個別の知識ベースシステムを構築する際に、共通機能群を提供し効率的開発を可能にするだけでなく、最終的なシステム統合の時点では、システムを有機的に連携させて働かせる管理システムにもなり得ることを想定したものである。

ここではこのような共通基盤システムに必要な技術のうち、反応設計知識ベースシステムに必要な知識獲得支援技術、知識利用技術、について述べる。

## 2. 反応設計知識ベースシステムに対する 要求条件

知識ベースシステムの構築に際しては、対象業務の分析、及び整理が必要である。

反応設計の場合には、現在、専門家は膨大な量の化学文献や反応データベースを検索しながら、化学者が持っている豊富な知識や経験を屈指し、最適な反応経路の選定を行っている。化学物質の構造が複雑な場合には、考えなければならない反応経路の数は膨大になり、反応設計には多額のコストと長期の時間がかかり、最近ますます化学物質の構造が

複雑化していく傾向のため、このような業務を計算機支援する必要性は益々増大している

反応設計用の計算機支援システムは数多く開発されてきており、LHASA<sup>1)</sup>、SECS<sup>2)</sup>、SYNCHM<sup>3)</sup>等のシステムが良く知られたものである。今後開発するシステムにおいては、これらの欠点を改善し、かつ近年の計算機工学の進歩を考慮した先端的なものであるべきである。このような観点からAI (人工知能) 研究の一環として研究開発が進んでいる知識ベース技術を使った計算機支援システムの開発を本研究では最大の狙いとしている。

化学物質の反応設計を計算機で支援するためには、反応アルゴリズムが充分得られること、及び化学的データ (ファクトデータ) が充分蓄積されているが必要である。しかし化学物質は数百万種もあり、物質間の反応データはそれを組み合わせたものであるから天文学的な数になり、反応アルゴリズムを体系的に得るのは非常に困難である。一方、ファクトデータについては専門家から比較的容易に得やすいだけでなく、各方面では精力的にデータベース (DB) 化<sup>4) 5)</sup>が進められており、一部商業ベースで利用可能な段階にある

これに対応して、化学反応の設計の方法論としては、反応アルゴリズムを体系化して得た反応規則を用いた論理指向型設計方法とファクトデータから得られる知識を利用して設計する情報指向型設計方法の2種に大別できるが<sup>6)</sup>、前者は新規な反応設計には有効であるが、反応規則を正確に、網羅的に獲得する

ことは困難という欠点を持ち、後者は比較的既知の反応の組合せからなる反応設計には有効であるが、一定の枠組みからはずれた反応については無力であるという欠点を持つ。

従って、現状では上記の情報指向型設計方法と論理指向型設計方法を組み合わせハイブリッドな方法論を採用することが実用的であるといわれている。

化学者が反応設計を行う場合、まず対象となる化学物質を構造図、分子式、名称等で認識する必要がある。次のステップでは、経験的な知識、及び化学文献から得た知識に基づき、設計対象の化学物質の特徴を認識し、合成の単位に分割する。合成単位で結合を

切断し、フラグメントに分け、各フラグメント毎に、経験的な反応規則、理論的な反応規則を用いて反応前の化学物質構造を求める。

そこで求めた反応前の構造を結合して、前駆体を求める。さらにこの前駆体を対象化学物質として考え、前駆体の化学物質が既知の原料となり得る化学物質になるまで繰り返す。反応経路は複数得られ、最終的には実験により確認される。

以上の化学者の反応設計作業モデルを図1に示し、それをシステム化する際の知識ベースシステムに要求される要求条件について以下に述べる。

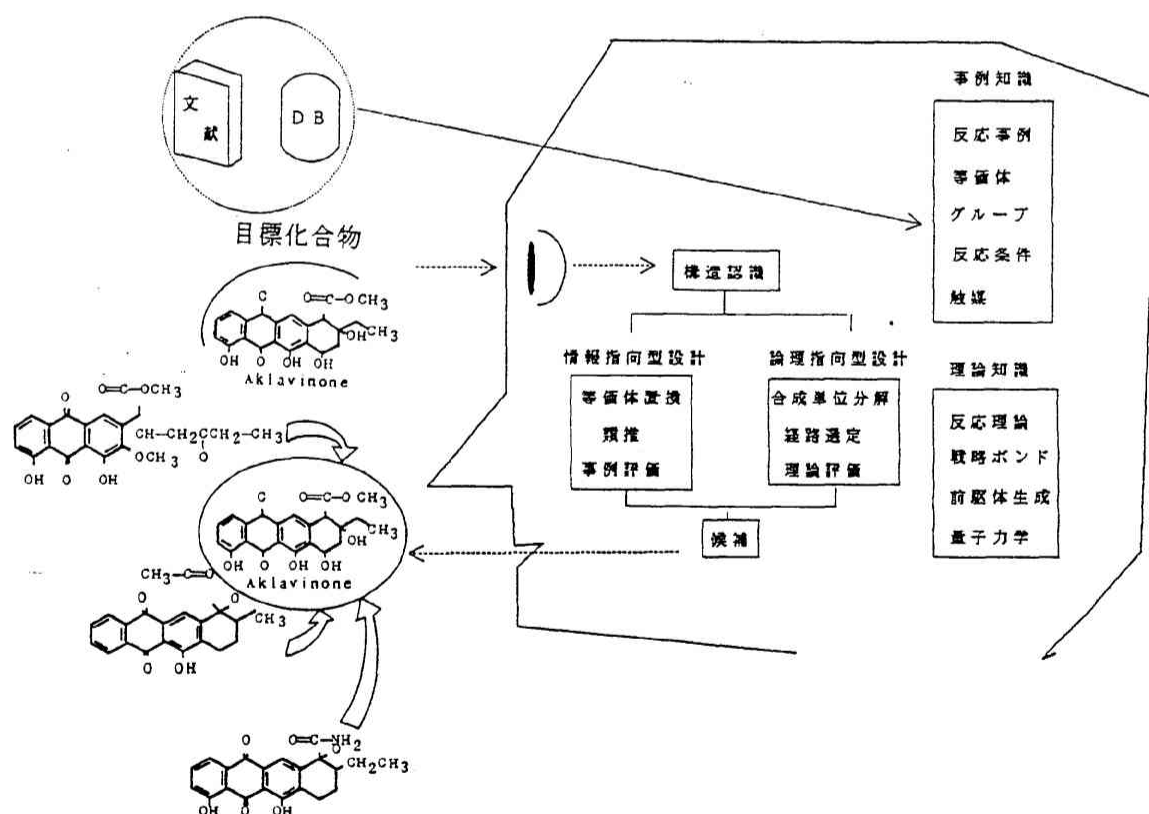


図1 反応経路選定の専門家モデル

### (1) 抽象度に応じた柔軟な知識表現

化学反応に関する知識は専門家の頭の中、あるいは化学の専門書に記述された体系的に整理された反応理論にあたる知識から、化学者の頭の中や、あるいは反応データベース等に蓄積されたファクト知識等に分類される。

計算機にこれら様々な知識を格納する場合には、その知識の量や利用法に応じて、適切な表現法を採用する必要がある。例えば、条件と実行からなるような規則的な知識は比較的ルール表現に適し、主語-述語からなるようなファクト型知識はフレーム表現に適していると考えられる。

さらに知識の格納、利用性を考慮すると知識の汎用的な表現方法も重要である。

### (2) ハイブリッドな設計手法

前述したLHASA、SECS、SYNCEM等は論理指向型の反応経路設計システムと言われるものである。また、REACCS<sup>7)</sup>やINSRED<sup>8)</sup>等は情報指向型反応設計システムとよばれるものであるが、両者の設計方法論をともに使ったシステムは余り報告が無い。新規化学物質の反応経路を設計できるためには、理論と経験を併用する必要があるが、理論については、獲得するのに莫大なコストを要する。さらに、反応理論には必ずと言って良いほど例外が付きまとうと言われ。これら知識を正確に網羅的に構築するのは難しい。従って、経験的な知識を最大限に利用して設計を進めることを原則としながら、理論が経験を補完するように構成することが最良である。経験的な知識の利用については、新規化学物質の設計を前提に

すると、そのものの経験的なデータが存在するはずは無いから、単に事例のマッチングでは設計ができず、反応部位とか反応に着目した等価性、類似性を考慮した設計を行う必要がある。

### (3) マンマシンインタフェース

利用者の利便を考え、設計対象の化学物質の構造が化学者が日常使っている化学構造図で入力できる必要がある。

このためには、利用者が入力したグラフィックイメージの構造を計算機が自動的に部分構造を抽出し、さらに特徴を認識し、計算機処理が可能なような内部表現形式に変換する必要がある。しかし、この処理は計算機が最も不得手とする所であり、かつこの認識の精度がその後のシステムの処理を決めてしまうので、最も注意してシステム構築すべきところである。

また、化学物質の構造表現については計算機が使う内部表現形式の他に、他システムとの交換情報にも使われ外部表現形式も重要である。この外部表現形式については、標準形式として取り決めておく必要があり、しかるべき機関で決定する必要がある。

具体的なマンマシンインタフェースはグラフィックディスプレイ上でアイコンやマウスあるいはライトペンを使った操作性の良いものがよく、さらに、部分構造等はユーザの指定により、任意に変更・追加できることが望ましい。また、全ての操作はオンラインで自動的に情報が管理される必要がある。

### (4) DBとの連結

反応DBからファクト知識を入手する以外

に反応DBを整理して規則知識を獲得するシステムとして、化学の膨大なDBから知識を自動的に抽出することを目的としたシステムの研究が多く進められている。<sup>9)</sup>

この場合の問題は、対象とする知識の抽出カテゴリをあらかじめ定義しておく必要があるため、カテゴリ分類の難しい化学分野の知識抽出システムの多くは、比較的簡単な反応についての知識しか抽出できないが、DBの不完全性が計算機による知識自動抽出の障害になる。そこでDBからの知識抽出システムには、なんらかの不正なデータや例外的なデータの除去を知的に行える枠組み<sup>10)</sup>が必要である。

### 3. KBMSを用いた反応設計支援システム

KBMSは知識ベースシステムを構築支援するための汎用ツールであり、各種知識ベースシステムに共通する知識蓄積/管理、推論マンマシンインタフェース等の機能群をシステム構築者に提供するものである。<sup>11)</sup>

このKBMSを用いた反応経路設計支援知識ベースシステムの全体構成を図2に示す。

化学物質の設計支援を行うために、KBMSの機能に化学物質が構造図で取り扱うことのできるマンマシンインタフェースと、専門家、及び反応DBから知識を獲得するための知識獲得エディタをかぶせた構成となる。

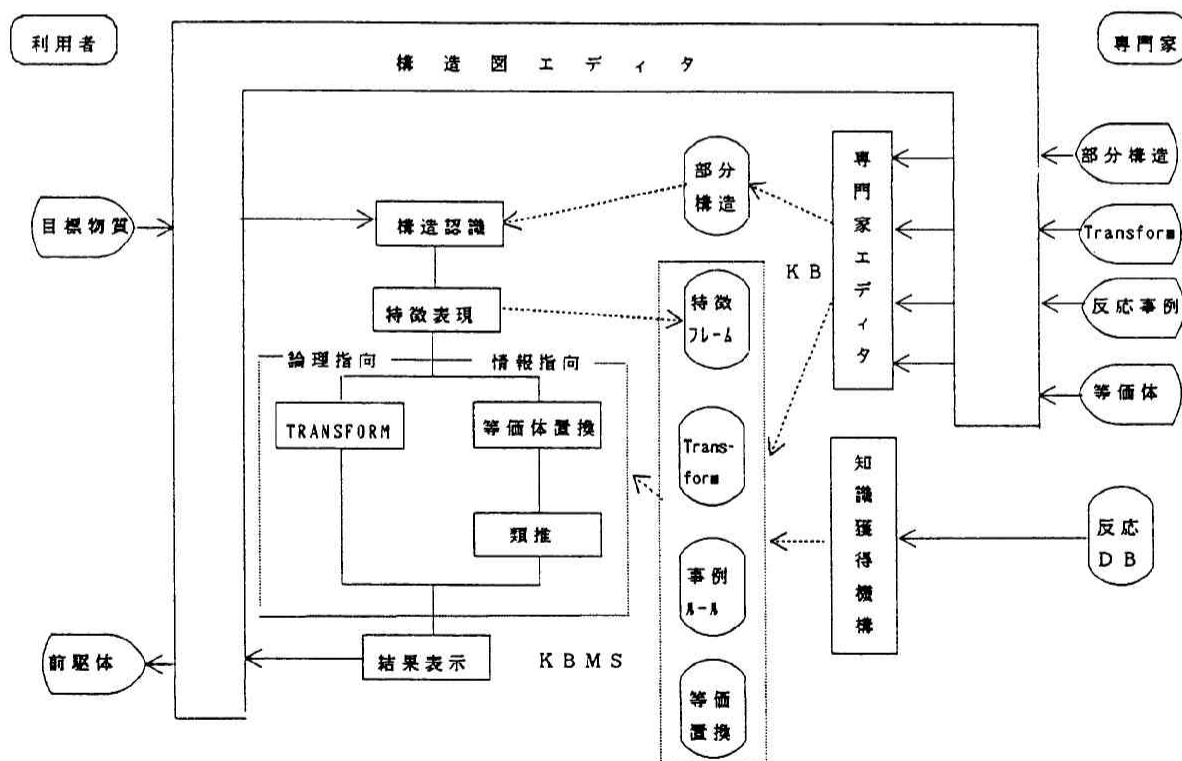


図2 システム構成

### 3. 1 K B M S の概要

N T T が開発した知識ベースシステム構築支援ツール K B M S の概要について図 3 に示す。

#### (1) 知識表現

知識をフレーム、ルール、及びワーキングメモリ等の組合せで表現し、特に、フレーム間に利用者独自の関係を定義することで、知識の表現や体系化が容易にできる。またルールはルールセットにより、モジュール化、及び階層化でき、ルールの体系化を可能としている。

#### (2) パフォーマンス

推論では多くの可能性を知識の中から探索するため、大きなマシンパワーが必要であり、大規模な知識ベースからなるシステムでは実行の高速化が要求される。K B M S では予めルールをコンパイルして最適化する方式を用いることにより、推論を高速化でき、リアルタイムのシステムにも使用できる。

またパソコン版 K B M S の場合には、15 M B というメモリ空間を使用でき、さらに、二次記憶の併用により、知識ベースの規模を事実上制限無しで使用可能である。

#### (3) マンマシンインタフェース

知識システムでは多様な知識を利用するために、マシンとそとの知識源との表現のギャップを極力短縮する必要がある。化学者の

日常使われる表現方式が化学構造図であれば計算機がこれを直接扱えることが必要である

さらに、スペクトルデータ等のようにアナログ信号等を扱う必要があれば、K B M S のような記号を処理するシステムにアナログ処理部を付加する必要がある。K B M S では汎用的な知識ベースエディタを持っており、マルチウィンドウ、メニューやマウス等を利用できる他、K B M S ツールキットやスクリーンデザイナーによるグラフィック作成支援ツールを利用してユーザ独自のインタフェースの容易な構築が可能である。

#### (4) 知識ベースの流通性

K B M S は P C 9 8 0 1 から大型マシンまで多様なマシンで動作できるが、知識ベースの表現とマンマシンインタフェースを統一しているため、異機種間のシステム移行が容易である。利用者の開発／実行環境に合わせて、各種形態がとれることは効率的システム構築に不可欠である。

#### (5) その他

以上の他に、K B M S は標準的な Common-LISP 言語で記述されているが、システムの C 言語への変換ツールや D B を知識源として利用できるように、D B - K B 変換ツールが用意されている。



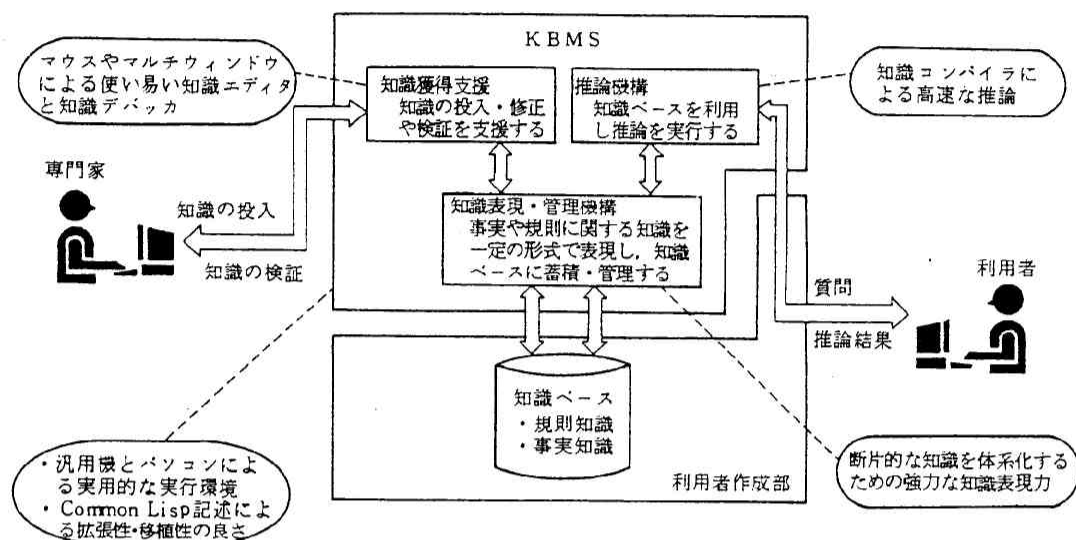


図3 エキスパートシステム構築支援ツールKBMSの特徴

### 3.2 各種機能概要

#### (1) 化学構造図入力エディタ

本エディタはLISPを用いて作成され、設計対象や知識獲得の際に、化学物質の化学構造をシステムに入力するため、また処理途中や設計結果の出力に使用される。操作性を良くするために、化学物質を構成する部分構造等の各パーツや操作コマンドをアイコン

で用意し、利用者はマウスでそれを選択するように構成した。さらに、名称等の投入にはキーボードを併用した。

本エディタが持つ主なメニューを表1に示す。図4はこれを操作している画面の1例である。

表1 構造図エディタの操作メニュー

メニュー	サブメニュー	メニュー	サブメニュー	メニュー	サブメニュー
拡大・縮小	定率拡大 範囲指定 定率縮小 全体表示 復帰	移動	要素移動 部分移動	ファイル	ロード マージ セーブ 一覧 検索
		部分構造	入力 名称一覧		
		印刷	開始 停止	付加情報	分子式 分子量 登録者名
スクロール		検索	名称検索 要素検索		
原子レベル			要素置換	グリッド	設定 解除
回転		置換 削除	要素置換 要素削除 部分削除 全削除		
消去	要素消去 部分消去 全消去	検査		終了	
複写	要素複写 部分複写				

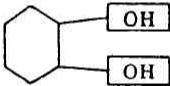
	官能基 (CH)
	官能基 (COH)
	官能基 (CN)
	官能基 (NH)
	官能基 (S)
	官能基 (NO)
	官能基 (C)
	官能基 (N)
	その他の基
	環
原子	
結合	
<p>&lt;入力モード&gt;  ロード マージ セーブ ファイル一覧 原子レベル 展開  要素消去 全消去 ハードコピー 終了</p>	

図4 化学構造図入力例



## (2) 特徴認識

化学物質の構造認識は図5で示す手順により行われる。

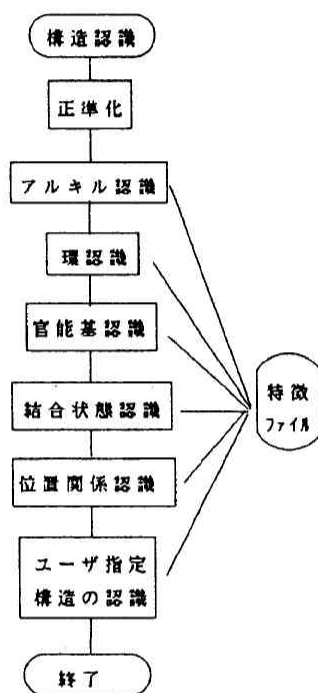


図5 構造認識の手順

### ①部分構造の認識

入力された化学物質は一旦原子結合表の形に表現され、正準化された後、LIST表現に変換する。正準化は代表的なモルガンアルゴリズム<sup>12)</sup>を使用した。次に、環やアルキルを認識する。環の認識にはSSSRのアルゴリズム(The Smallest Set of Smallest Ring)を利用した。環、骨格以外は部分構造として登録されている構造が認識される。

最後に、これらを特徴フレームに表現する。

### ②部分構造間の結合状態の認識

部分構造間のボンドの状態、即ち単結合、二重結合、共役結合等の特徴フレームに表現する。但し、現在は立体は認識できない。

### ③各部分構造の位置関係の認識

部分構造相互の位置関係を認識する。位置関係は予めシステム構築者が定義した関係を探査し、あれば特徴フレームに表現する。

位置関係は化学反応の分野に依存する知識であり、場合によっては、遠く離れた2つの部分構造間の位置を把握する必要がある。

これら位置関係を探索する固有の手続きについてはシステム構築者が作成する必要がある。

### ④各部分構造の性質の認識

各部分構造の電気的性質等を認識し、特徴フレームに表現する。例えば、官能基の電子供与性や電子吸引性等である。

この性質の認識システムについても、化学反応に依存するので、システム構築者自身が予め作成する必要がある。

以上のうち、かなりの認識手続きが化学者自身の作成を要求される。但し、これらの多くは化学者自身が既に利用しているものであり、今後はしかるべき機関で標準的な認識項目、及びフォーマットが審議され、システム間の交換情報としても利用されることが、望ましい。

本システムでは、実験的システムなので比較的単純な反応しかサポートしていないが、実用システム構築には、今後多くの化学的な認識プログラムが追加せさなければならない

### (3) 知識表現

#### ① フレーム

設計対象の化学物質はフレームで表現される。

フレームには知識の構造を定義するクラスと知識データそのもののインスタンスから構成される。クラスはスーパークラスと通常

のクラスに階層化され、クラス、スーパークラスで定義された情報はインスタンスに引き継がれる。図6はこれらの関係を使って化学物質の特徴フレームを表現した例を示す。

フレームのインスタンスの各スロットには特徴認識の結果が記述される。

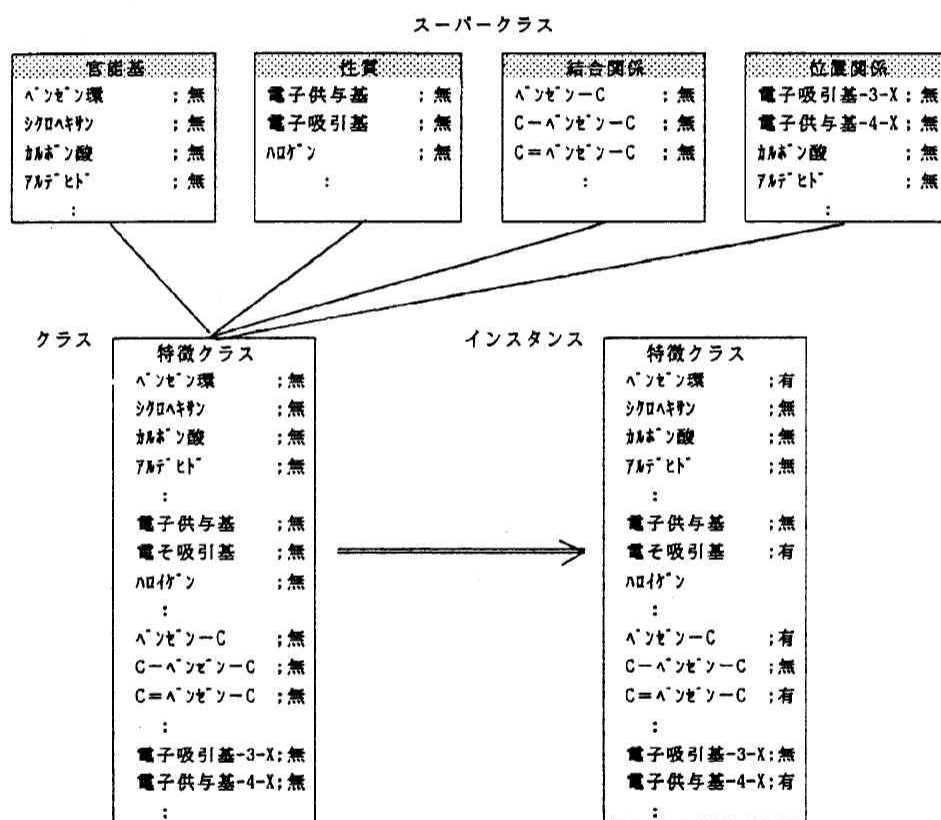


図6 特徴フレームの記述例

#### ② ルール

ルール知識には論理指向型設計を行うための反応理論知識と情報指向型設計のための反

応事例に基づくファクト知識の2種類がある  
反応理論知識は各反応のプリミティブなオペレーションを化学物質の特徴を示す各種シ

ンボルを使用して条件と実行を表現したもので、1つの反応オペレーションは原則として、複数のプリミティブなルールの連結により達成される。図7は本システムの実験で用いたルール例である。ルールの条件部は特徴フレームに記述されている化学物質の構造特徴が記述され、対象化学物質にこの特徴が

あれば、実行としてTransform実行関数がコールされる。Transform実行関数は標準的な関数と化学反応を想定して、システム構築者が予め作成しておく関数とがあり、切断、付加、置き換え等のプリミティブなオペレーションについてはシステムが予め用意した。

```
(defrule (trans2 r4)
  (frame (部分構造 入力構造 (ベンゼン-単結合-電子吸引基 有)))
  (frame (部分構造 入力構造 (ベンゼン-単結合-x 基 有)))
  (frame (部分構造 入力構造 (電子吸引基-4-x 基 有)))
  -->
  (call (change 'x 基 'H)) .....taransform
)
```

図7 taransformルールの記述例

他方のファクト知識については、専門家と対話により取得する方法と、反応DBから自動獲得する方法がある。前者は反応を前駆体、生成体と反応条件からなる反応方程式と

みた時、対話的にそれらを専門家に入力させ計算機内でそれをルールの形に表現するものである。図8に事例ルールの例を示す。

```
(defrule trl r12 (:priority 50)
  (frame (入力データ 入力構造 (環種類 環のモル名)
    (位置ハタツ (0 2 3))
    (反応種類 求電子)
    (置換基名 (CH3 - - - CHO -000
    (性質 (R - - - - Z -))
    (試薬 CHO)))
  -->
  (CALL (change 'CHO 'H))
)
```

図8 反応事例ルールの記述例

本事例ルールは反応が芳香族性単環の置換反応の例であり、化学物質の構造特徴は置換基の種類と位置、性質に限定した簡単な表記を用いた。さらに、反応条件としては求電子、あるいは求核置換反応の別、試薬を入力できるようにした。さらに、ルールの競合解決や類推に使用するプライオリティを各ルールに付与した。プライオリティはルールの重要度を示すもので、例えば、工業的に著名な反応の値を高くすることによって、この反応が優先的に使用されるよう制御するものである。

DBからのファクトルールの獲得ではDB内に記述されている原子結合表、あるいは特徴テーブルから前駆体と生成体の構造知識を得、また反応条件についてはDB中のテキストから抽出するようにした。これらの手続きは対象とする反応DBによって異なる。

化学ではこの反応DBの作成は古くから行われているが、各々形式が異なるため、相互流通がなされていない、化学物質の種類は今後も増大すると思われ、反応DBの標準化が行われれば、このような知識獲得は容易にできるものと考えられる。

### ③等価体

本システムでは反応知識の他に、教科書的なファクト知識として、合成等価体知識を専門家から取得する機構を設けた。後でも述べるが、合成等価体知識はルールを用いた設計を進めて行く過程で部分構造を等価体に置き換えて、あらためてルール適用を行うもので、これを行うことによってより現実的な反応経路を設計することができるようになる。

### ④ その他

知識表現はしていないが、グループファイルを有し、専門家にグループを入力させることができる。これは、電子供与基、電子求引基、X基、M基といった官能基の性質を定義するもので同一の性質を持つ部分構造は同一のグループに所属することを登録する。

グループ表示子には各々シンボルを与え、必要なら専門家はグループシンボルを使ってルール入力が可能である。このようにグループシンボルを使えば、より汎用的なルールの記述ができ、ルールの適用可能性を増大するとともに、ルール量の圧縮も可能である。

## 4. 推論

知識に対応して論理指向型設計のための推論と情報指向型設計の推論から構成される。

ここでは情報指向型設計で使用される等価体置換、及び類推について述べる。

### 4. 1 等価体置換

化学合成的にみて、極めて容易に相互に置き換え可能な複数の部分構造の組合せを等価体と呼び、これを実際の設計時に対象となる化学物質の一部を等価な別の部分構造に置き換える。設計はこの置き換えた化学物質に基づいて行い、これらを合成等価体置換という。このような等価体置換は日常専門家が多用しているもので、これだけでかなりの反応設計が可能である。

等価体置換には相互に可逆性のあるものと単方向のものが考えられるが、これは次の制約リストによって制御される。

化学反応が化学物質の一部の反応ではなく、他の部分構造に影響される場合が多いが、等価体の考えにも、このような制限を知識化した制約リストが必要になる。なお、本システムでは設計結果に対して、単位反応評価を行うシステムが付加され、その中で等価体置換の制約チェックを行うことを前提としているので、制約リストをインプリメントしなかった。しかし、候補を予め絞り込むために制約リストの表現方法、利用法について検討を要すると考える。

#### 4. 2 類推

新規化学物質の設計を仮定すると、それに対応する事例知識は獲得されていないが、過去の事例を利用して対象の化学物質に対する設計結果を推定するなんらかの知識を生成する方法が考えられる。その代表的な方法として類推という方法がある。化学反応事例にこれを適用するとすれば、既に知識として獲得されている反応事例の中から、設計対象の化学物質に類似した化学物質の設計事例を探索し、これをテンプレートにして、設計対象に適合するように反応事例を変形し、新規な事例知識を獲得、及び推論を行うものを類推と呼ぶことができる。

類推により生成された知識は必ずしも、正しいとは限らない。化学反応の場合、正しくても、コスト、収率、安全性等の要因で実用にならない知識もあり、これらは正しくない知識と判断すべきものも存在するはずである。類推の精度を高めるには、単一の事例のみに基づいて、変形類推を行い、新規知識

を生成する方法では、精度を高めることはできない。精度向上策としては、複数の似かよった過去の事例を学習することにより、新規の知識を生成する方法がとられるべきである。つまり、類推と学習は組み合わせて使用することにより、類推の効果を高めることが可能になる。その他に、より本質的に、類推の精度を左右するファクタが存在する。

1つは、設計対象の化学物質と事例中の化学物質との類似性判定方法であり、他方は探索された事例をもとに、設計対象化学物質に対応した新規知識を生成するための変形方法である。いずれも、類推システムのトータル精度を決めるベースとなるもので、この正当性が類推システム構築上のキーポイントになる。

##### (1) 類推原理

反応設計とは、新規化学物質の化学構造  $C_{tj}$  について  $S_j$  を決定することである。つまり、 $S_j = \{C_{pj} \ C_{tj} \ K_j\}$  において、前駆化学物質の構造式  $C_{pj}$  と反応条件式  $K_j$  を求める問題と定義できる。そこで化学反応における類推とは何かについての考え方を述べる。

図9は類推の考え方を図示したものである

$S$  は全反応知識の集合であるが、これを分類すると、化学反応として意味のある Positive (正) な  $S$  と化学反応としては意味が無いか、あるいは間違っている Negative (負) な  $S$  に分けられる。設計する場合には、Positive- $S$  のみで設計するのが順当であろうが、場合によっては、Negative- $S$  を制約として、設計を進めることも考えられる。

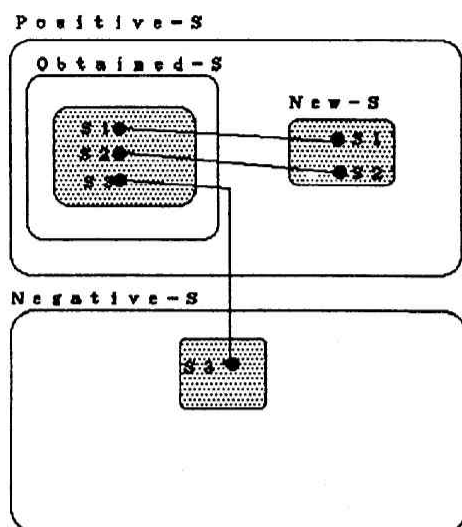


図9 類推の考え方

全てのSはPositive-SとNegative-Sの足し合わせた集合であるが、類推が必要な場合はこのSが得られていない場合、即ちSの部分集合である既に獲得されているObtained-Sのみが得られている場合である。

類推はこのObtained-Sを利用して行われる。類推する場合、先ずObtained-Sの中から、設計結果のNew-Sに類似性のあるSを探索することを行わなければならない。図9の場合、黒丸が1つの反応知識を示しており、Obtained-SのなかのS1、S2、及びS3が探索された類似性のあるSである。

これらのSからは、New-Sが対応しており、それぞれS1'、S2'、S3'がNew-Sになり得る。ここで注意しなければならないのは、全てがNew-Sになる保証は無く、図9の場合にはS3はNew-Sにならず、Negative-Sに登録されることを示している。

最悪の場合には、全てNegative-Sになることも有り得る。類推を充分の精度で行えるためには、既に獲得されているObtained-Sが充分であること、及び類似性有りとしたSの探索が確実であることが必要である。

## (2) 類似性

Sの探索に必要な類似性判定方法について述べる。上述したように、類似性はNew-SとObtained-S間について定義される。しかし、類似性判定段階では、New-Sの全てが判明していないので、両者間の部分照合(Partial-Matching)により判断することになる。部分照合の考え方について述べる。

今、SoをObtained-Sとし、SnをNew-Sとすると、反応知識Soと反応知識Snは、それぞれ以下のリストで構成される。

$$S_o = \{C_{po} \ C_{to} \ K_o\} \leftarrow \text{Obtained-S}$$

$$S_n = \{C_{pn} \ C_{tn} \ K_n\} \leftarrow \text{New-S}$$

Soは全て判明しているが、Sn中のCpn、及びKnは判明していない。Ctnは設計対象の化学物質構造式であるから判明している。従って、照合は同じカテゴリ間(化学構造式というカテゴリ)で行わなければならないから、Cpo、Ctoと及びCtn間の部分照合を行うことになる。

反応は構造式の変化とみる。つまりCpo→Cto、及びCpn→Ctnの構造変化である。

部分マッチングは原子レベルのマッチング(atom-by-atom matching)を行い、類似性の判定には極大一致度の計算を行うことを基本とする。しかし、原子レベルのマッチン



グはかなりの計算量であるため通常の計算機では、実用的な性能を確保できないことが予想される。従って、実用性のためには、計算量の削減を図った方法の導入が必要である。

そこで、近似的な方法として、既知化学物質、及び設計対象化学物質双方とも、いくつかのインデックスに分類して、インデックス間の一致度を求め、これらを合計することによって、類似性を判定する方法を用いる。

このインデックスは構造の特徴と電気的な性質の特徴から構成される。構造の特徴としては、極めて安定性の強い、即ちそれ自体が変形しない部分構造記述子と部分構造間の結合を示すなんらかの結合パターンの2種から構成される。また、性質には各部分構造の結合点における電気的性質、及び部分構造そのものの電気的性質等に分類される。例えば、電子吸引性、あるいは電子供与性等である。

例えば、芳香族単環の置換反応においては、構造、及び性質のインデックスを図8の事例ルール例に対応して以下のように捉えることができる。

#### <構造のインデックス>

- ①環の種類
- ②置換基の位置パターン
- ③置換基の種類

#### <性質のインデックス>

- ④置換基性質

以上の部分マッチングを全ての知識について試験する必要はない。化学反応が類似し

ているとは、化学構造が類似しているだけでなく、化学反応それ自体の反応メカニズムも類似、あるいは同種であることが必要である

従って、この場合の部分マッチングは、化学反応中心から近傍の類似性を判定することで充分である。しかしながら設計対象のCtnを見て、どこが反応中心であるかは決定できない。全ての既知反応には反応中心点があり、類似性判定の前に、その反応中心が設計対象の化学物質に持ち得るかどうか検査する。持ち得ない場合には該当の既知反応と設計対象の化学物質に関する新規反応は同じ反応中心にならないから、この既知反応は類推に利用できない。反応中心を同一とすることができる既知化学反応が見つかれば、この反応中心の近傍について部分マッチングを行う。

#### (3) 類似度計算

類推は、反応部位を同じくする全ての反応について、CtoとCtnの上記のインデックス毎の一致度が計算される。

- ・環の種類については、一致する場合をa点とする。
- ・位置パターンの場合には、一致する場合をb点とする。
- ・置換基種類については、一致する置換基個数をn個とすると、類似度はc n点とする。
- ・置換基性質についても、一致する置換基個数をm個とすると、類似度はd m点とする。



全体の類似度は、

$$\text{類似度} = a + b + c n + d m$$

と計算される。

類似度が大きいほど、反応が似ていると判断され、類推は類似度が大きいほうから行われることを原則とするが、反応には、構造や性質の類似性の他に、重要度といったファクタがある。例えば、収率、経済性、及び反応の安全性といった点は、実用的な反応かどうかを評価する重要なパラメータであり、類似なルールを類似度で判定する際にも、この要素を加味する必要がある。

この方法としては、各反応について上記の a から d の値をこの観点から幅をもたせる必要がある。これらの反応の重要度は各反応の反応条件に記述される他、各反応に付与可能なプライオリティの値がこれを反映することとした。

従って、具体的な a から d はプライオリティを p とした時、下表のように設定した。

表 2 類似度計算例

	1 < p < 50	51 < p < 100
a	6	10
b	6	10
c	8	15
d	5	5

#### (4) 類推反応の生成

生成は類似な反応をテンプレートに、これを設計ターゲットの化学物質に適合するようにアップデートすることによって行われる。

その基本は So の Cto を Sn の Ctn に合致させ、かつ So の Cpo を変形し、Sn の Cpn を生成することである。その他の Kn は Ko に合致させる。

芳香族単環における置換反応に限定したプロトタイプでの類推反応の例では、Sn のの生成は以下の方法で行われる。

項目	参照する反応
・環の種類	; Ctn
・位置パターン	; Ctn
・置換前置換基種類	; Ctn
・置換前置換基性質	; Ctn
・置換後置換基種類	; Cto
・置換後置換基性質	; Cto
・反応種別	; Ko
・反応条件	; ko

#### (5) 類似性判定の高度化

反応事例を利用した推論方式は、比較的事例知識が専門家、ないしは反応事例 DB から入手しやすいことから、有効な方式であり、その一実現例として事例知識の類推による方法を示した。

類推をシステム化する場合の最大のポイントは、設計対象の化学物質に類似する化学物質の類似性の判定であり、さらに構造の類似性だけでなく、反応メカニズムの類似性をも考慮すべきであることを示した。

類似度の計算では、原子レベルの部分マッチングを行うことは、計算量の膨大性から採用できず、構造、及び性質をインデックス分類した各要素毎の一致度の合計点で判定することを提案した。また、最終的に複数の新

規な反応を提示する際には、各々の事例に付与された優先順位を利用して、ある程度の順位付を行うことができることを示した。

ここでは、反応事例を利用した推論方式について、基本的な枠組みを提案したが、対象とする化学物質、あるいは専門家にデペンドしたよりきめ細かな推論方式の実現に向けては、この枠組みに厳密な化学アルゴリズムの付加が必要である。これら化学により密着したプログラム群の開発については、化学専門家の今後の努力に期待するものである。

## 5. 候補の順位付

1つの化学物質の合成経路を求めると複数の経路を出力することが考えられる。専門家はこのうち、専門家が持っている固有の判断知識で優先順位を付けて経路を設計することになる。ここで、優先順位が必要な理由は、一般に新規化学物質の場合には、複数段の経路を設計することが必要であるが、優先順位なしに経路設計を全組合せについて行うと膨大な経路を設計してしまうことが容易に想定される。このなかの多くは合成化学的に意味の無い経路を選定することが多く、システムが段階毎に優先順位付をして、最終的にある程度の優先順位付をして経路を提示できるようにすることが考えられる。

優先順位付の判断は理想的には反応が専門家に意味があることを認識するメカニズムをシステム化する必要があるが、これら認識の知識は専門家固有の知識であり、獲得することが極めて困難であること、また専門家によってその知識の質が全く異なることが予想さ

れることなどから、経路選定段階では厳密な優先順位付は行わず、経路評価システムを別途、経路選定後に使用することを前提としておおまかな優先順位付をシステム化することで充分と考えた。

### (1) 優先順位付の手法

優先順位付の代表的な手法として、統計的な知識学習方法の導入が考えられる。

全ての反応事例知識にはプライオリティ  $p$  が付与され、この  $p$  は知識構築段階で、あらかじめ初期設定しておくべきものである。

例えば、 $p$  が1から100まで付与可能であるとして、知識構築時には専門家の責任で  $p$  を設定する。専門家によっては反応の収率を判断ポイントとする場合もあろうし、あるいはコストを判断ポイントとする場合も考えられる。いろんな要素を組み合わせで  $p$  を設定するのであるが、厳密な数値を設定するのではなく、例えば、 $p$  を高、中、低程度に分け、それぞれ、100、50、10のように設定する。

### (2) 自動生成する優先順位

類推によって新規に得られる事例知識については、システムが  $p$  を生成する必要がある

類推の場合には、知識を生成するためのテンプレート知識があり、その知識の  $p$  を新規知識に適用する。

### (3) 優先順位の変動

知識構築段階でおおまかに与えられた  $p$  に対して、システムの運用段階で運用状況をシステムが判定して、 $p$  を変動する機能を付加し、より適切な事例知識が反応経路設計に利用されるようにする。

変動方法として、システムの利用者の事例知識の使用頻度に応じた統計的な方法を採用する。システムでは各事例知識の使用頻度を記録しておき、 $p$  の変動幅を計算する。

通常、 $p$  は数値であり、正の整数で与えられるので、再計算の際には、 $p$  が上限、あるいは下限を越えないようにする必要がある。

この為には、再計算結果に  $p$  を 50 を中点に 1 から 100 の間に値を分散させる必要がある。

$p$  の再付与に対しては、以下の機能をさらに必要とする。

- ①利用者が再付与を指示できる機能。
- ②特殊な反応事例については再計算の対象外とする機能。
- ③  $p$  決定アルゴリズムをユーザ定義できる可能性

#### (4) 事例の参照回数

$p$  の再計算の基になるのが事例の参照回数である。反応事例が参照されるケースを以下に示す。

- ①実際の設計に使用される反応事例が参照された時。
- ②類推に使用され、かつ新規ルールとして登録された場合の反応事例が参照された時。

#### (5) 高度な推論制御の検討

反応事例の個々に  $p$  を付与し、この  $p$  に従って、反応事例知識の利用順番を制御するメカニズムについて示した。

化学合成設計の分野であっても、対象とする化学物質によって、その手法は様々なものがある。専門家によては、利用すべき反応事例の順番はシステムによる画一的な制御だけでは不十分である。

得られた結果の反応が正当なものかどうかは、さらに厳密な評価システムを利用する必要がある。以上の制御方式については、反応事例の参照回数に応じた統計的な方式を採用してあるが、システムとしては、専門家固有の優先制御アルゴリズムを組み込める手段も考慮するべきと考えられる。

さらに高度な制御方式を推論システムに求めるとしたら、メタ知識制御の基本的機能を具備する必要がある。知識ベース構築支援ツールの中にはメタ知識を記述できるものがあるが、KBMS には該機能が組織化された仕様でインプリメントされておらず、かつ該知識は専門家が明確には持っていない知識であるために、本システムでは今後の検討課題とした。

#### 6. その他

知識獲得を効率良く行わせるためには、本システムには次のような機能を実験的にインプリメントしてみた。

##### 6. 1 負知識ベース

専門家からの知識投入時、及び類推による新規知識登録時、システムは専門家と対話的に登録の是非を問う方式を採用してい。この時点で登録する知識が正しい場合には、そのまま知識ベースに登録するが、正しくない

場合でも、同じ誤りを生じた時に、判断を再び専門家に求めることのないように、誤り知識を知識ベース化している。これを負知識ベースと呼んでいる。

## 6. 2 部分構造名称の内部表現方法

本システムで認識している部分構造は反応知識獲得時に新規な部分構造が出現したら、システム内の部分構造ファイルに自動的に追加されるようになっている。そのために、システムは自動的に部分構造の名称を発生する方式をとっている。その方法としてはモルガンの規範化による線形表記を用いた。これにより、専門家はグラフィックスで自由に構造入力しさえすれば良く、知識投入時の煩わしさが減少する。

## 7. 経路選定実験

本システムの基本構想と具体的インプリメントの仕様については、化学の専門家と協議し取り決めた。プロトタイプは開発の費用と期間の制限で基本的枠組みを確認する範囲にとどめた。

確認実験では化学の専門家、及び反応DBから約200種の反応知識を入手した。

非化学専門家なので、実験の良否判定は充分にはできないが、図10に規則知識を使った論理型設計手法によるイソフィトールの合成経路選定を行った例を示す。さらに、事例知識の類推を使った情報指向型設計手法による芳香族性単環の置換反応の場合では、合成の教科書程度の約40種程度の反応経路を推定できた。いずれの場合にも、意味の無

い、あるいは誤った反応が多く出力されたがこれは主として、本システムを構成している構築プログラムの制限によるものである。

本システムは現在化学専門家に提供し、機能の一部高度化と知識の投入等を進めており近々にその結果が報告できる予定である。

なお、今後の実用的な化学反応設計支援知識ベースシステムの開発に際しては、前節までに述べた種々の高度化と検証するための必要十分な知識の確保が必要である。これら多くの作業は専門家自身が寄与すべき部分であり、化学専門家の努力に期待するものである。

## 8. おわりに

化学物質設計用知識ベースシステムを構築するにあたっての共通基盤システムのプロトタイプについて述べた。本システムの構想及びそれに具備した各種機能については、本プロジェクトに参加して頂いている化学専門家と協議して決定したものである。また、プロトタイプの作成にあたっては、費用と期間の制限の中で大部分はNTTが作成したものであるが、一部分は実際に化学専門家が作成したプログラムを利用している。本プロトタイプは化学物質、及び化学反応がまだ限られたものにしか対応できない状況であり、最も重要なシステムに蓄積されるべき化学の知識も小量であるために、実用に供するにはいまだ以上に多くの専門家の協力が必要である。さらに、反応DBについては、反応事例知識の獲得と規則知識の獲得に有効ではあるが、本システムでは試みとして、専門家

が実験的に作成した小規模なDBを利用したために、得られた知識量は少なく、今後はさらに大規模なDBの連結を検討する必要がある。

本システムの高度化は、実際に化学専門家に提供し、化学専門家自身が高度化を図れるための機能増強できるように、提供者としてはツール類の整備等の各種支援を行うつもりである。科学技術庁の全体プロジェクトは残り2か年を実用的システムの作成とシステム統合がメインの研究課題となっており、これまで以上に、化学専門家と情報工学関係者の協力体制が重要になりつつある。

design program (SECS), 11回情報化学討論会予稿集, 6117, 72-75, 1988

(10)山崎毅文、桑原敏、服部文夫；例外事例を含むDBからの化学知識の自動抽出, 知識工学と人工知能研究会, 65-6, 1989

(11)NTTソフトウェア編；エキスパートシステム構築支援ツール KBMS/PCマニアル, 1989

(12)H. L. Morgan ; The generation of unique machine description for chemical structure

#### (参考文献)

- (1)E. J. Corey, W. T. Wipke : Science, 166, 178-192, 1969
- (2)W. T. Wipke, H. Braun, G. Smith, F. Chopin, W. Sieben : ACS Symposium 61, 97, 1977
- (3)K. K. Agarwal, D. L. Larson, H. L. Gelernter : Comput. Chem. 2, 75-84, 1977
- (4)CAS (Chemical Abstract Service)
- (5)CIS (Chemical Information System)
- (6)国際科学振興財団編：「反応設計データの収集、整理、分析、評価」調査研究報告書, 1985. 3
- (7)REACCS (Molecular Design Ltd.)
- (8)松浦育敏：データベース型反応設計支援システム, Molecular Design, vol 7, no. 5
- (9)谷中幹郎、中村和彦、棚沢あずさ、W. T. Wipke : Automatic knowledge base building for the organic synthesis

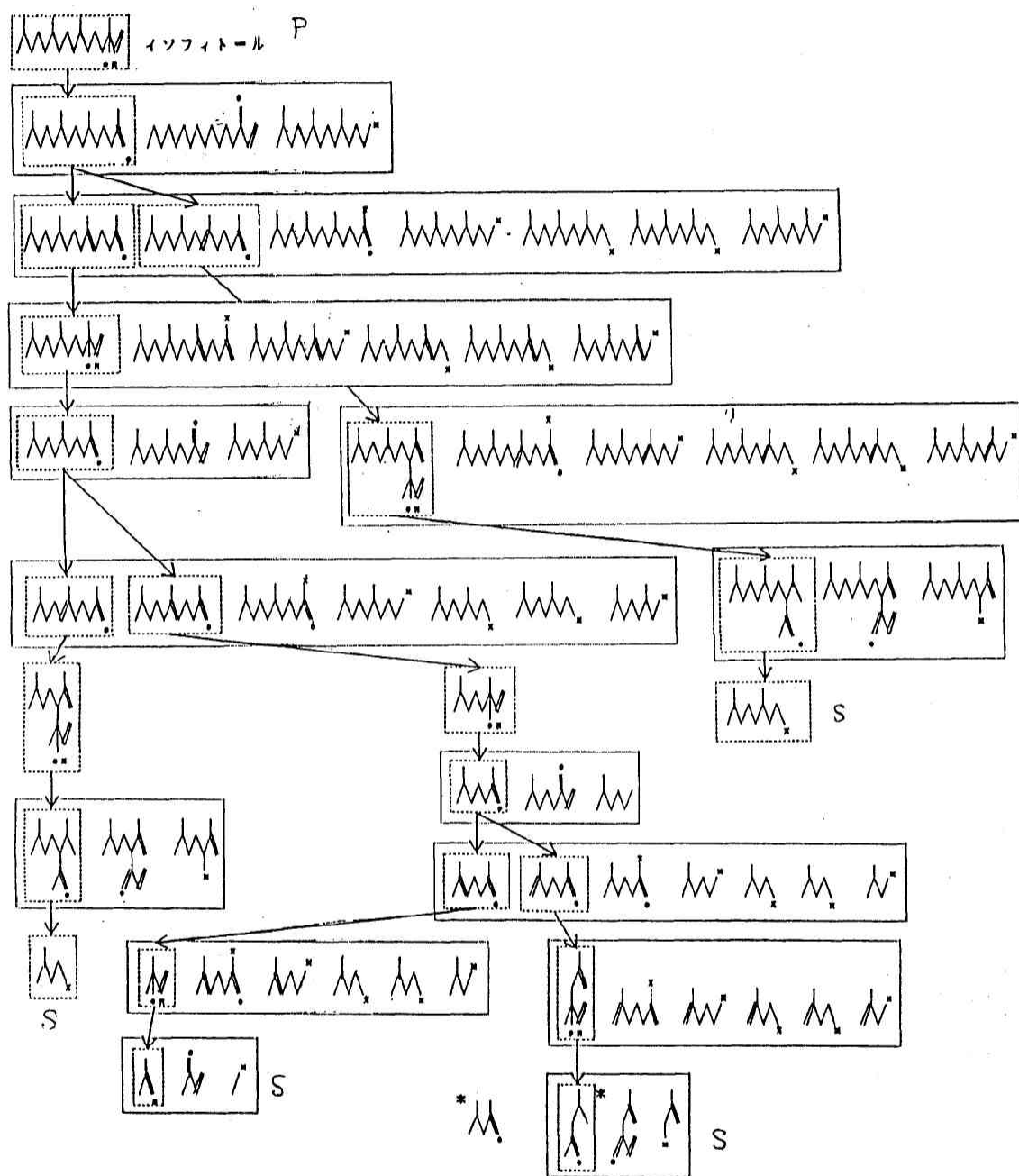


図 10 反応経路選定実験 (Transformルール使用例)