4．その他の検索様式

□　アトムバイアトム検索（完全一致および部分構造検索）
（原子単位の対応をチェックする、最もシンプルで最も厳格な検索手法）

アトムバイアトム検索は文字通り原子と結合単位で検索キー構造式と被検索化合物との完全一致を求めるアプローチである。　ある一つの原子を出発原子とし、その原子につながる結合と原子について検索キー構造と被検索化合物とでその対応関係をチェックする。このチェックが完了したならば、そのチェック済の原子から出ている結合と隣接原子について同様に検索キー構造と被検索化合物との対応をチェックする。　この手続きをキー検索構造式の総てについて行うのがアトムバイアトム検索である。

この手法はコード変換や化合物の修飾等を行わないシンプルな手法のため、最も検索精度の高い検索手法と位置付けられる。　このため、アトムバイアトム検索は検索の最終確認を行う目的で利用されることが多い。　しかしこの検索は、他の検索手法と比べると検索速度が極端に遅い事が最大の欠点となる。　このため、通常の検索ではこのアトムバイアトム検索は第1次あるいは第2次といったプレスクリーニングで被検索化合物群を充分に絞った後の少数の化合物群に対して適用される。

・アトムバイアトム検索例

アトムバイアトム検索の実行手続きと例を図　に従って順に説明する。

①　検索キー構造式の原子種に該当する原子を被検索化合物から取り出す。

検索キー構造式中の1番目の原子はケトン酸素である。　この原子と同じ特性の原子を被検索化合物から取り出す。　この場合3番目の原子がケトン酸素に該当する。

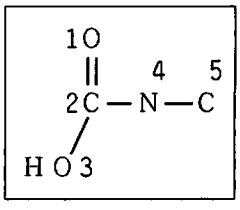
検索キー構造式原子　1　⇨　被検索化合物原子　3

②　検索キー構造式の1番目の原子につながる原子を順に取り出し、その原子に該当する原子を被検索化合物原子から探し出す。

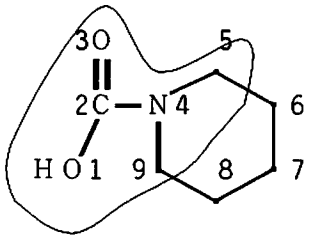
検索キー構造式の1番目の原子に隣接する原子は2番の原子である。　また、被検索化合物原子の3番に結合する原子は2番の原子であることがわかる。　この検索キー原子の2番の原子と被検索構造式の3番目の原子の特性をみると、ともにカルボニル炭素であることがわかる。　従って、検索キー原子の2番と被検索構造式の2番目の原子とは一致するという結論が得られる。

検索キー構造式原子　2　⇨　被検索化合物原子　2

検索キー構造式



被検索化合物



検索キー構造式

原子番号	原子種	隣接原子
1	＝O	2
2	＝C＜	1　3　4
3	－OH	2
4	－N－	2　5
5	－C	4

被検索化合物

原子番号	原子種	隣接原子
1	－OH	2
2	＝C＜	1　3　4
3	＝O	2
4	－N＜	2　5　9
5	－C－	4　6
6	－C－	5　7
7	－C－	6　8
8	－C－	7　9
9	－C－	4　8

図　．結合表によるアトムバイアトム検索実行例

③　②の操作を繰り返し、検索キー構造式原子の総てについて対応する原子が被検索化合物中に存在するかを原子単位でチェックする。　検索キー構造式原子の総てが被検索化合

物原子中に含まれていたならばヒット化合物となる。　②を繰り返す途中で検索キー構造式原子に対応する原子を被検索化合物原子に存在しなくなったならば、再び①にもどる。④　①の操作において検索キー構造式原子の１番目の原子に該当する原子を被検索化合物原子中に新たに見出された時は②～④の過程を繰り返す。　該当原子を見出せない時、部分構造検索はヒットしなかったこととして部分構造検索を終了する。

・アトムバイアトム検索と部分構造検索および完全一致検索との関係

アトムバイアトム検索において、検索キー構造式が被検索化合物中に含まれる時は部分構造検索が行われたことになる。　このように、アトムバイアトム検索は部分構造検索に利用されることが多いが、検索キー構造式が被検索化合物と完全に１対１で対応した時は完全一致検索が行われたことになる。　このようにアトムバイアトム検索手法を用いても完全一致検索を実行することは可能であるが、実用上検索速度が遅いため完全一致検索を目的として利用されることは少ない。

・アトムバイアトム検索の高速化のための工夫（プレスクリーニングの設定）

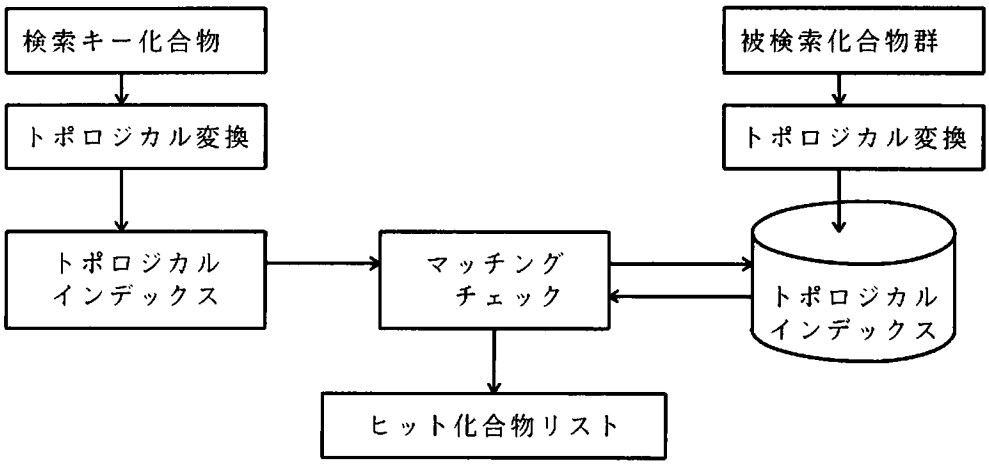
アトムバイアトム検索の手順でも分かるように、この検索の第一歩は検索キー構造式原子に該当する原子を被検索化合物原子の中に見出すことである。　従って、この最初に探す原子が特殊なものであればあるほど該当する原子が被検索構造式中に存在する確率は小さくなり（検索のヒット率が小さくなる）、従ってアトムバイアトム検索を具体的に実行する前に検索を中止し、無駄な時間をかけることがなくなる。　この結果、アトムバイアトム検索自体の検索速度は向上しなくとも、データベース全体としての検索速度を向上することは可能となる。

この特徴を利用して効率的な検索を行うならば、アトムバイアトム検索の前に総ての原子特性について検索キー構造式と被検索化合物との相関をチェックするプレスクリーニングを設定するのが良い。　部分構造検索であるならば、検索キー構造式の総ての原子に該当する原子群を被検索化合物原子群は含んでいることが必要である。　この条件を満たさない時、アトムバイアトム検索による部分構造検索を行う必要はない。

現在する第部分の化合物データベースの化合物検索（特に部分構造検索）では、その検索過程にアトムバイアトム検索を含んでいることを考えるならば、検索速度向上の為には検索キー構造式中に特徴のある原子群を多用することが検索速度向上に効果的であることがわかる。　反対に、どの化合物にも含まれる一般的な原子群だけで構成される構造式を検索キー構造式とするならば、その検索には時間がかかることになる。　この点を考慮しつつ検索キーを作成すれば、より高速な検索を実行することが可能となる。

□　トポロジカルな構造式変換技術を用いた化合物検索（完全一致検索）

化合物構造式をトポロジカルな変換式を用いて一連の数値データ（トポロジカルインデックス）に変換し、この数値データを用いて化合物検索を行う手法。　このアプローチでは化合物の完全一致検索を行うことを目標とする。



化合物の完全一致検索を行う為には構造式を基本として発生される数値データが化合物に対して一元一項対応している事が必要である。　この目的を実行するための変換手法に関する説明は既に行っている（第　章）。　ここでは化合物構造式の数値データへの変換手法としてトポロジカル変換手法を用い、この変換された数値データを用いて検索する手法についてのべる。

化合物構造式のトポロジカル変換の目的は、種々物性値と化合物構造式との対応を説明する事をメインとして開発されており、化合物検索を目的とするものではない。　従って、化合物検索で重要な一元一項対応に関する考慮は行われていない。　この一元一項対応問題をカバーするために、トポロジカル変換数値を複数用いることが必要である。

・トポロジカル変換概要

トポロジカルな変換手法としてさまざまな手法が提唱されているが、これらの詳細についての解説はここでは行わない。　これらの変換手法は化合物の原子を“点”に、結合を“線”とみなし、この点と線で構成されるグラフ（化合物）を様々なアルゴリズムに従って数値データへと変換するもので、この基本はグラフ理論に求められる。　現在、化合物を対象とするトポロジカル変換として様々な手法が開発されている。　これらの変換手法のうち、分子結合インデックス（Molecular Connectivity Indices）が様々な物性や活性との相関解析に利用され、その実績も多い。

トポロジカル変換は化合物の結合関係、結合種、原子種等を考慮して発生される数値データであり、化合物の様々な結合状態をきめこまかに数値データへと反映することが可能である。　但し、2次元的な情報を基本として発生される為、立体不斉等3次元情報に起因する情報をのせることは不可能である。　この点での改良が必要である。

・トポロジカル変換データの特徴

化合物構造式のトポロジカルな変換で得られるデータは数値データであり、文字列とは異なる。　従って、先にのべたＭＯＲＧＡＮアルゴリズムで得られる原子対応のためのユニーク番号は必要としない。　たんに化合物の構造情報さえあれば数値データへの変換が可能である。　この点で数値データへの変換手続きは大幅に簡素化されることになる。

トポロジカル変換により創出された数値データは検索に先立ち、あらかじめその大小関係に従ってソート（並べ変え）しておくことで高速な完全一致検索を実現することが可能である。

一般的に一回のトポロジカル変換により複数の数値データが得られるが、個々の数値データが表現（内包）している化合物構造情報は限定されているのが普通である。　一方、化合物の完全一致検索では化合物全体の情報が必要であり、従って検索に利用される数値データはこれら化合物の全情報を反映していることが必要である。　このためにトポロジカルデータを用いた検索では、個々のデータによる情報もれを防ぐ目的で複数の数値データを用いた検索が必要である。

検索に利用される数値データの種類や数等については実際の検索を行いながらデータベース単位で最適なものを決定してゆく、この時出来る限り少ない数でデータベース内の

全化合物群を分類出来る数値データを選択することが重要である。

表　．トポロジカルインデックスを利用した検索とユニークナンバリングを必要とする検索手法との比較

項目	トポロジカル インデックス利用	ユニーク ナンバリング利用
構造式表示数値データ	数個の数値データの 同時利用	数値データ１個 （長い数値データとなる）
変換手続き	変換アルゴリズム利用	ユニークナンバリング及び ルールに従った構造式変換
記憶領域サイズ	＜	

・トポロジカルデータを用いた化合物検索手続き

トポロジカル変換技術を用いた化合物の完全一致検索システムの作業フローを図　に示す。　検索もれに対する保証の目的で、検索に用いられるトポロジカル変換データは複数（ｎ個）用いている。

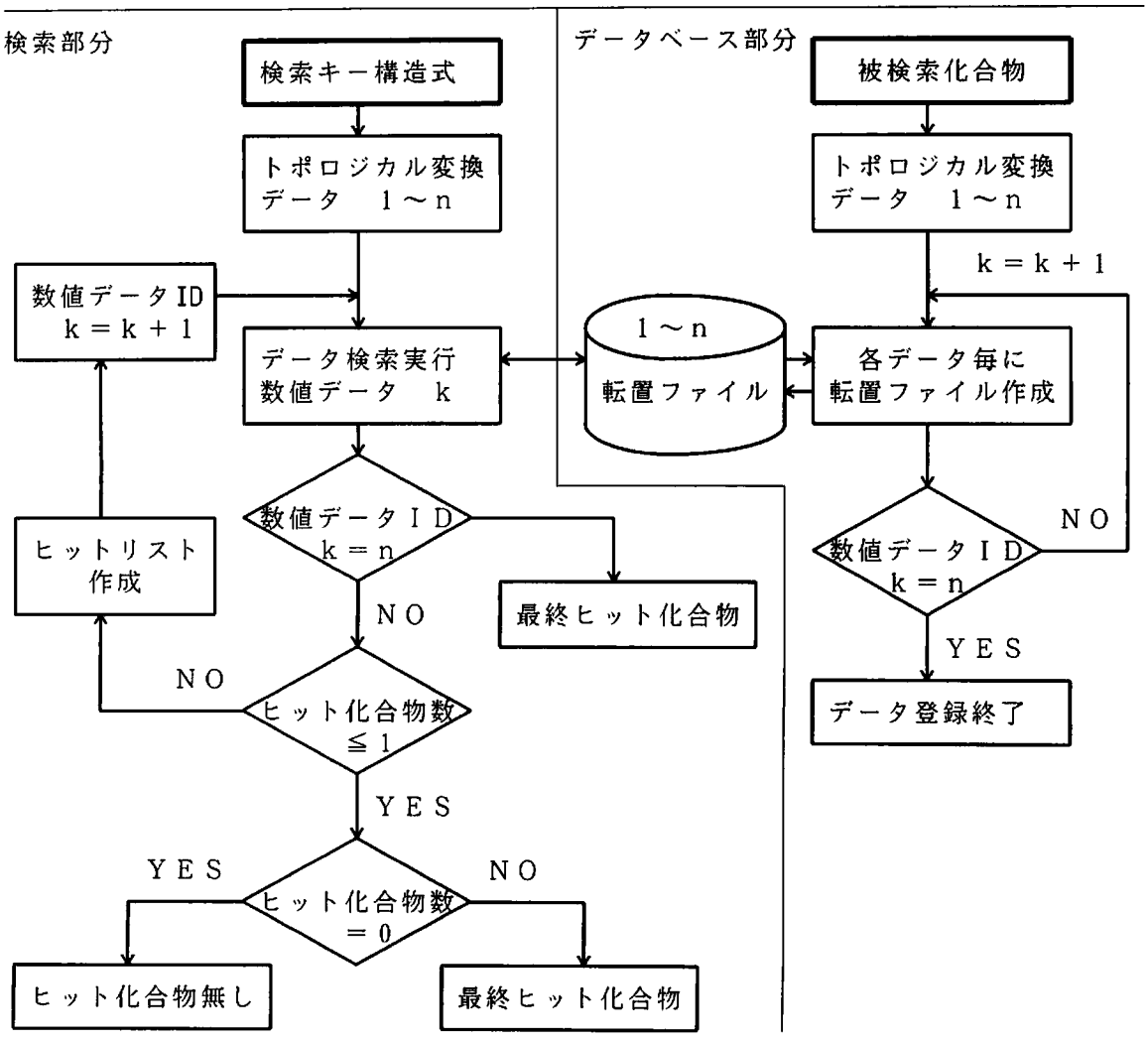


図　．トポロジカル変換データを用いた化合物検索（完全一致検索）システム流れ図

〔データベース部分〕　検索に用いる被検索化合物は、トポロジカル変換により得られた個々の数値データ単位で作成されたｎ個の転置ファイルとしてデータベース内にストアされる。　この転置ファイルは被検索化合物をデータベースに登録する度に新たに更新される。　この更新過程では新たな数値データ（化合物）を加えると同時に、その値の大きさ

の順にソーティングされて再ストアされる。

〔検索部分〕　最初に、検索キー構造式はトポロジカル変換によりｎ個の数値データへと変換される。　この数値データの一個を取り出し、そのデータと同じ特性を持つ数値データを用いて作成された転置ファイルを用いてデータの一致をチェックする。　一致する番号が存在しない時、該当化合物は存在しないとして検索を終了する。　また、一致する番号を見出したならば、その番号に対応する化合物のインデックス番号を取り出せばヒット化合物を見出したことになる。　この時、ヒット化合物が一個の時は検索を終了し、その化合物を最終ヒット化合物として出力する。　しかしヒット化合物が複数存在する時、検索は終了せずに次のステップに進む。

先の検索で取り出されたヒット化合物群のリストを作成し、そのリストを新たな被検索化合物群とする。　このリストを用いて、検索キー構造式のトポロジカル変換により生成された２番目の数値データと２番目の数値データに基づいて作成された転置ファイルとを用いて一致する番号を探し出す。　この手続きをトポロジカル変換データを用いてヒット化合物が１個になるまで繰り返す。　ｎ個の変換データ総てを用いても被検索化合物群が１個とならない時は、その複数の化合物を最終ヒット化合物として出力する。

このようにトポロジカル変換技術を利用することで、化合物検索は単なる数値データの比較問題へとすりかえられるため、高速な化合物検索が可能となる。

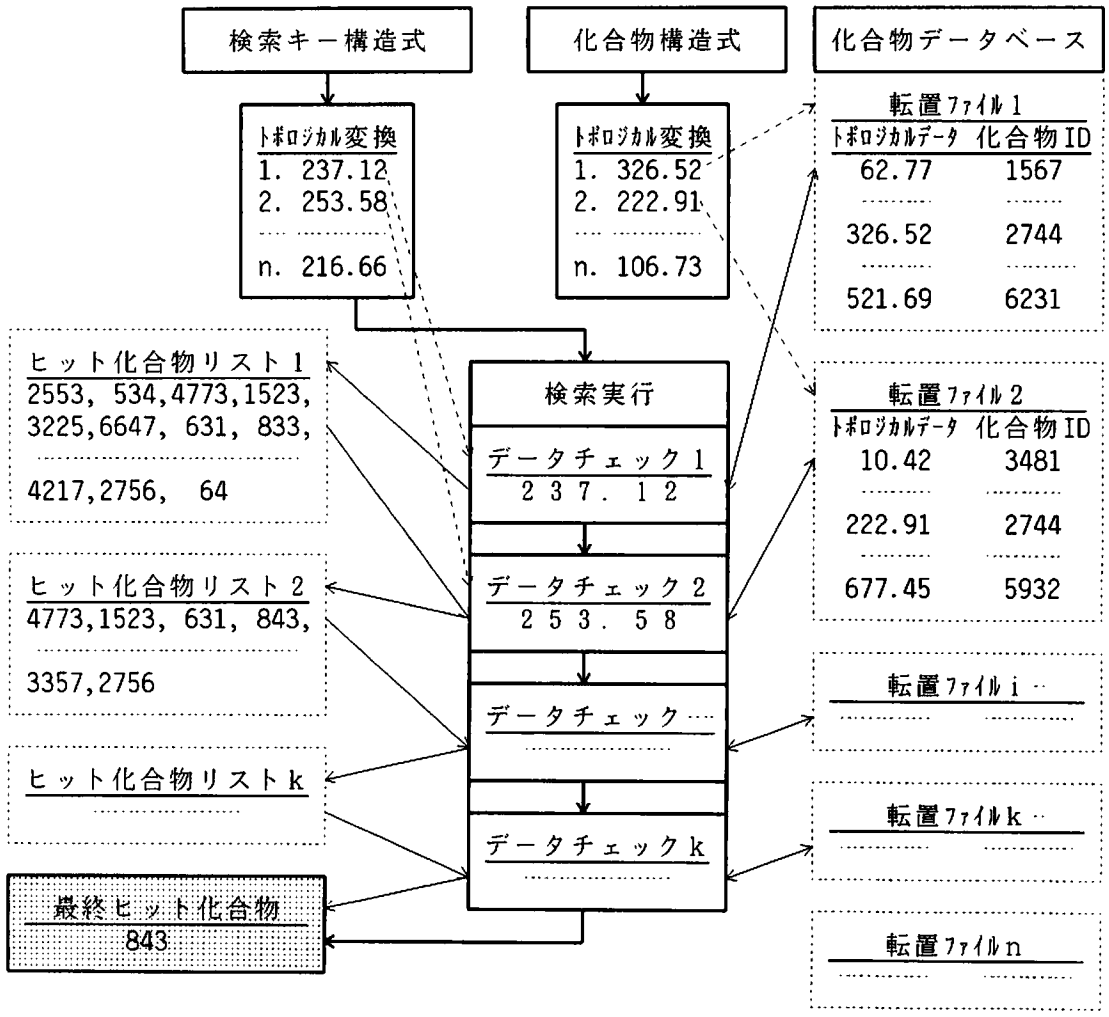


図　．トポロジカルデータを用いた化合検索（完全一致検索）実行例

図　にはトポロジカル検索の流れに従って、数値データと検索ヒット化合物リストを中心として表示したものである。

被検索化合物がトポロジカル変換によりｎ個の数値データへと変換され、その個々のデータについて転置ファイルが作成される。　検索構造式が入力されると、トポロジカル変換によりｎ個の数値データへと変換され、これらの数値データが順にデータベース内の転置ファイルと比較されながら検索が実行され、最終的に１個の化合物がヒット化合物として出力されてゆく過程がしめされている。

□　ＷＬＮコードで構築されたデータベースの部分構造検索

ALBERT V. TOMEA, PETER F. SORTER; On-Line Substructure Searching Utilizing Wiswesser Line Notations, J.C.I.C.S., 16,223～227(1976)

化合物の線型表記法であるＷＬＮコードを用いて化合物の部分構造検索を行うことも試みられている。　ＷＬＮコードは本来、計算機上での利用を目的とした化合物のネーミング手法として開発されたもので、化合物のデータベースへのストアやシステムへの入力の容易さ等を特徴としたものである。　このＷＬＮコードに関する説明は既に詳しく行っているのでここでは省略し、ＷＬＮコードで構築された化合物データベースの検索という観点からまとめる。

・ＷＬＮコードを用いた検索について

ＷＬＮコードは文字及び数字データのみで構築されている。　従って、この特徴を最大限に利用するならば化合物検索は文字列及び数字列の検索と置き換えられ、従来の検索技術を用いて化合物検索を行うことができる。

このように、完全一致検索を行う限りにおいてはＷＬＮコードは大変有効なものであるが、部分構造検索を行う時にはこのＷＬＮコードをそのまま用いることは不可能である。なぜならば、ＷＬＮ規約により、化合物の表現は化合物全体を基本として（原子の番号付け、原子の相対的位置、その他）おり、従って部分構造を基本としたＷＬＮコードとはコード自体が全く異なるものとなることである。　部分構造検索をＷＬＮコードを用いて行うとすれば、文字として表現することの出来る最小単位としての官能基レベルでの検索が限界であり、より高度な情報を含む部分構造の検索は不可能である。

・ＷＬＮコードで構成されたデータベースでの部分構造検索

このＷＬＮコードで書かれたデータベースを用いた部分構造検索も行われている。　然し、この場合ＷＬＮコードそのものを用いた検索ではなく、一旦ＷＬＮコードを結合表へと置き換え、その置き換えられた結合表を用いて部分構造検索が実行されている。

このようにＷＬＮコードを一旦結合表へと変換するのは、ＷＬＮコードでは表記の単純化を目的として表記上の最小単位が官能基レベルで統一されているからである。　このために、部分構造検索で必要となる原子単位での認識が不可能となっている。　この欠点をカバーする目的で、一旦ＷＬＮコードを結合表へと変換することが必要となる。

・ＷＬＮコードの結合表への変換

結合表も部分構造検索だけを目的とするものなので、この部分構造検索を実行するのに必要な最小限度の情報を持つ結合表であれば良いことになる。　この目的で作成された結合表の構成を以下に示す。

結合表に必要な要素として、①原子番号、②原子シンボル、③隣接原子ＩＤ、④環情報の４種類が定義されている。　検索対象とする化合物構造式を前記４種類の情報へと変換する。　また検索キー構造式に関しても同様に４種類の情報へと変換する。　この変換により化合物の部分構造検索が実現されることになる。

表　に実際に用いられた原子シンボルの例を示す。　化合物の水素以外の１原子単位で該当するアルファベットをシンボルとしている。　表　には環情報の定義がしめされている。

・結合表を用いた部分構造検索の実行

部分構造検索は結合表を用いたセットリダクション法を基本として行う。　ここでのセットリダクションは先に展開したアプローチとは異なり、より単純化されたアプローチでをとっており、検索対象としてセットを構築しながら検索を行うという意味でのセットリダクションである。　このセットリダクションを行い、最終的な検索ヒットの確認は検索キー原子と被検索化合物原子との対応をチェックするアトムバイアトム検索で行う。この２段階検索によりＷＬＮデータベースでの部分構造検索を実現している。

図　に検索構造式の結合表が示されている。　この結合表中、②の原子シンボルが複数存在するのは、ＧＥＮＥＲＩＣ（総称）検索を意図しているためである。

図　には図　の検索キーを用いて検索された結果の化合物がしめされている。

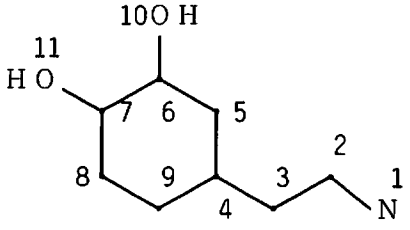
表　．原子シンボル及びその構造情報

原子シンボル	該当構造式情報
D	－ＣＨ＝
E	－Ｂｒ
J	Generic halogen
N	－Ｎ＜
M	－ＮＨ－，＝ＮＨ
O	－Ｏ－
Q	－ＯＨ
S	－Ｓ－，－Ｓ－，－Ｓ－
....	.....
....	.....
Z	－ＮＨ2
1	－ＣＨ3

表　．環情報及びその定義

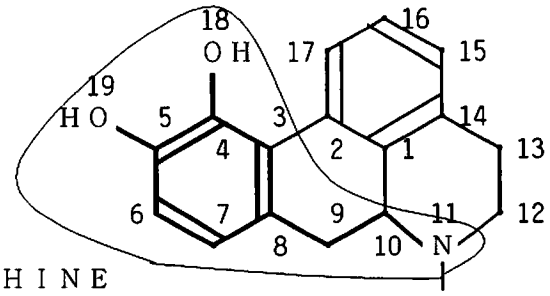
環情報ＩＤ	環構造情報
0	非環状
1	総てのタイプ
2	総ての環（ベンゼン環を含む）
3	非ベンゼン環
4	ベンゼン環

検索キー構造式



原子番号	原子シンボル	隣接原子情報	環情報
1	N，M	2	1
2	C，D，L，……	1　3	1
3	.....	2　4	1
4	.....	3　5　9	2
5	.....	4　6	2
6	.....	5　7　10	2
7	.....	6　8　11	2
8	.....	7　9	2
9	.....	4　8	2
10	Q	6	0
11	Q	7	0

図　．部分構造検索キー構造式及びその結合表



APOMORPHINE

原子番号	原子シンボル	隣接原子情報	環情報
1	T	2 10 14	2
2	T	1 3 17	2
3	T	2 4 8	2
4	T	3 5 18	2
5	T	4 6 19	2
6	D	5 7	2
7	D	6 8	2
8	T	3 7 9	2
9	L	8 10	2
10	Y	1 9 11	2
11	N	10 12 20	2
12	L	11 13	2
13	L	12 14	2
14	T	1 13 15	2
15	D	14 16	2
16	D	15 17	2
17	D	2 16	2
18	Q	4	0
19	Q	5	0
20	1	11	0

図　．検索ヒット化合物構造式およびその結合表

③　関連特殊検索

化合物データベースは単に化合物をストアし、化合物を取り出すだけの目的以外にも様々な目的を持った検索に利用される。　ここではこのような様々な目的を持つ検索手法、あるいは化学特有の問題に起因する特殊な検索手法についてのべる。

1．類似／最近隣検索手法の特徴

類似／最近隣検索手法は検索キーと被検索体との類似関係（類似関係を距離的な関係に置き換えた時は距離関係となる）を基本として検索する手法である。　このようにこの検索手法はいままでのべた完全一致検索や部分構造検索と異なる目的／形態を持つ第3の検索手法である。　類似／最近隣検索では従来の検索手法で議論されていた検索キーと被検索化合物間の包含関係は大きな議論の対象とはならない。　むしろ、検索キー構造式と被検索化合物間の類似関係だけが議論の対象となる。　従って、類似関係の評価手法等によっては検索キー構造式と被検索化合物とになんの包含関係がなくとも類似度は高いという関係も存在する。　このような特殊な関係を持つ化合物は、従来の完全一致検索や部分構造検索では決してヒットしなかった。

従来手法の完全一致検索は類似度が1（１００％）となる特殊なケースであると考えられる。　また部分構造検索でのヒット化合物は類似度の高い化合物となりうる類似候補化合物となる。　このように類似／最近隣検索手法は従来の検索手法と比べると特異な検索形態を持つ。　またこの検索手法は従来の検索手法では見出し得ない高度な情報を提供しうる第3の検索手法として高い存在意義を持つ。

2．類似／最近隣検索手法における類似度の決定手法

類似／最近隣検索手法でもっとも大事なキーポイントは検索キーと被検索体との類似関係を求めるところにある。　この類似関係の把握手法に関しては現在までに様々なアプローチが提唱されてきた。

この類似度を求めるのにはある一定の計算式を定義し、この計算式の値を比較することで類似度を求めるのが最も一般的かつ原始的なアプローチである。

・対象化合物群を限定し、カストマイズされた計算式を用いた類似／最近隣検索

比較対象とする化合物群が同系列の化合物群に限定されている時、この化合物群の類似度を計るための計算式は一般的な関数を用いるよりも、対象とする化合物群の特徴を強く意識した計算式を用いた方がより良い結果を得られることが多い。

このように対象とする化合物群が強く限定されている時、その類似度算出に利用される計算式は、対象とする化合物群を特徴付ける代表的な原子種、官能基、環その他の必須部分構造の存在情報等を利用することが多い。　従って、このような計算式は対象とする化合物群が限定されている時は極めて高い信頼度をもつが、対象とする化合物群が広範な化合物群である時への適用は困難である。　従って、一般的なデータベースを扱うという目的がある時は類似度計算式として、より一般的な計算式を利用することが多い。

$$SI(\%) = Wb \cdot \sum \frac{Pi}{5} + Wm \cdot \sum \frac{Pi \cdot \dfrac{1}{e^{Mi}}}{N} \qquad (\ )$$

基本部分　Ｐｉ：	可変部分　Ｍｉ：
1.縮合４／５員環	1.　各種ヘテロ原子
2.ベータラクタム	2.　各種ハロゲン原子
3.カルボン酸	……
4.Ｎ／Ｓ	……
5.アミド	Ｎ.　Ｘ員環の存在

Ｐｉ：ｉ番目の項目を満たす時１、満たさない時０
Ｍｉ：ｉ番目の項目（基準化合物の値－比較対象化合物の値）
Ｗｂ：基本部分の重み（種々の化合物を扱う時７０、誘導体のみを扱う時１０）
Ｗｍ：可変部分の重み（種々の化合物を扱う時３０、誘導体のみを扱う時９０）

この類似度計算式の特徴は対象化合物群を基本部分と可変部分とに分割し、それぞれについて値を求めていることである。　この算出式において、基本部分とはその化合物群を

特徴付ける基本構造部分であり、可変部分は各化合物に変化をつける部分を意味する。特に基本部分には必須要素を満たす時１、満たさない時０の係数がかかっているので、基本要素の存在情報は類似度算出に大きく影響することになる。　また、基本部分と可変部分とで、類似度計算部分にかけるウエイトが変化している。　このウエイトは、基本部分と可変部分による類似度に関する貢献度を規定する。　変化に富んだ種々化合物群を扱う時は基本部分の貢献度を大きくし、可変部分の貢献度を低くする。　一方、同族体を扱う時は、既に同族体ということで基本部分の情報を満たしているので、基本部分の貢献度を低くし、可変部分の貢献度を高くする。　このウエイトの操作により、全く構造の異なる化合物から誘導体まで扱うことが可能となる。　しかし、この類似度計算式では化合物群を種々化合物群と同族体化合物群とに分けて扱うことが必要で、両系統の化合物が混在する場合への適用は不可能である。

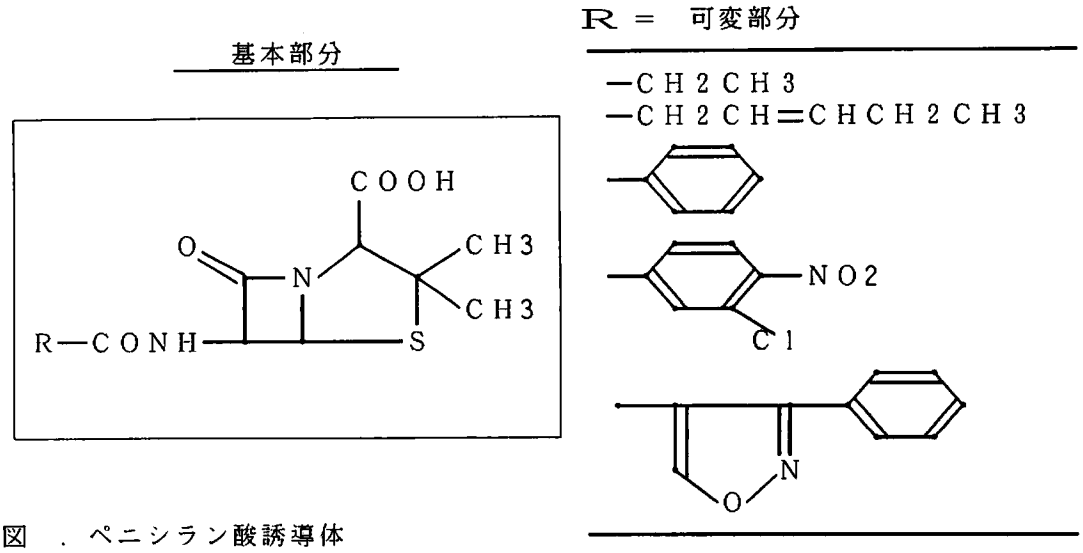


図　．ペニシラン酸誘導体

・一般的な計算式を用いた類似／最近隣検索

特定な関数を定義する他に、様々な分野で利用されて実績のある関数を利用することも考えられる。　現在、固体間の類似度（距離で代表することが多い）を計算する関数として様々なものが提唱されている。　その一部を以下に列挙する。

ＣＯＳＩＮＥ係数：

$$SI_{A,B} = \frac{C}{(Q^2 S^2)^{1/2}}$$

谷本係数：

$$SI_{A,B} = \frac{C}{Q+S-C}$$

Dice Coefficient：

$$SI_{A,B} = 2C/(Q+S)$$

Overlap Coefficient ：

$$SI_{A,B} = C/\mathrm{Min}(Q, S)$$

Hamming Distance：

$$SI_{A,B} = Q+S-2C$$

Ｃ：サンプルＡに対して１、サンプルＢに対して１の場合の数。
Ｑ：サンプルＡに対して１、サンプルＢに対して０の場合の数。
Ｓ：サンプルＡに対して０、サンプルＢに対して１の場合の数。
Min （Ｑ，Ｓ）：ＱとＳのうち、小さい方の値を取る。

これらの他にも様々な類似度を求めるための関数が用意されている。　解析の目的やデータの種類等により自分の目的に最も相応しいものを選ぶことが必要である。　また、ここで示された関数はサンプルＡとサンプルＢ、そしてサンプルＡ、ＢのＡＮＤを取った時のデータを用いたものである。　これ以外にも、類似度をもとめる為のデータが多次元になる時はパターン認識の手法の利用（例えば“最近隣法（Ｋ－ＮＮ法）”等）も考えられる。　これは、類似度というものを、多次元空間におけるパターン間の距離と考えることで可能となるアプローチである。

□　部分構造を基本とした類似度算出のためのデータ

類似度を算出する為の様々な計算式は前項にて示された。　この計算式を用いて類似度を求めるためには計算を行うためのデータが必要である。　しかもこのデータは検索キー構造式と被検索化合物との類似関係を的確に表すものであることが必要である。

現在利用されている類似度算出のためのデータは部分構造を基本として求めらるものであることが多い。　すなわち、ある指定された部分構造が検索キー構造式と被検索化合物とに共通に存在するか否かの情報をデータとするのである。　すなわち、検索キー構造式中に存在する部分構造（フラグメント）の種類とその数をチェックし、同時に被検索化合物中に存在する部分構造の種類と数をチェックする。　この部分構造の種類と数とをデータとして、先の計算式により類似度を算出するのである。

□　ＣＯＳＩＮＥ関数および谷本関数を用いた類似化合物検索例

ここでは例としてＣＯＳＩＮＥ関数と谷本関数とを用いて類似度検索を行った事例を紹介する。　尚、ＣＯＳＩＮＥ関数と谷本関数とはともに０～１の値を取り、その値が０の時は類似性が全く無い事を、１の時は互いに比較されたデータは全く同一のものである事を示している。　先程の２式は以下のように示される。

ＣＯＳＩＮＥ関数：

$$SI_{A,B} = \frac{\Sigma N[A,J]\,N[B,J]}{(\Sigma N[A,J]^2 \;\; \Sigma N[B,J]^2)^{1/2}}$$

谷本関数：

$$SI_{A,B} = \frac{\Sigma N[A,J]\,N[B,J]}{\Sigma N[A,J]^2+\Sigma N[B,J]^2-\Sigma N[A,J]\,N[B,J]}$$

ここで、$SI_{A,B}$は化合物ＡとＢとの類似度を、Ｎ〔Ａ，Ｊ〕及びＮ〔Ｂ，Ｊ〕はＪ番目の部分構造の検索キー構造式Ａ、及び被検索化合物Ｂにおける出現回数を、また総和は検索キー構造式Ａと被検索化合物Ｂとについて取る事を意味する。

類似検索は、予め設定されている部分構造の存在について、検索キー構造式と被検索化合物とについてチェックし、そのデータをもちいてＣＯＳＩＮＥ関数と谷本関数を導き出す。　この値を大きい順に並べ、値の大きい化合物から順に類似度の高い化合物として取り出す。　図　には検索キー構造式と最も類似度が高いとして取り出された上位４化合物が示されている。

図　と図　とをみた時、ＣＯＳＩＮＥ関数による類似検索の結果が谷本関数による結果よりも良くないことがわかる。　ＣＯＳＩＮＥ関数による結果では検索キー構造式と大きく異なる化合物が３個も存在している。　これは、類似度算出に用いたＣＯＳＩＮＥ関数では一方だけに部分構造が存在する時、その情報を類似度計算に充分に反映することが出来ないということに起因しているのであろう。

谷本関数による結果は類似度検索の問題を充分にクリアしているように見える。　ここでえらばれた上位４化合物の構造式はキー検索構造式と充分に類似したものである。　この結果は先のＣＯＳＩＮＥ関数を用いた類似検索の結果と比較することで、谷本関数の有効性がよくわかる。

用いた部分構造は、その構造単位が大きくなる程より良い結果をもたらす傾向がみられた。　これは、より大きな部分構造になる程結合パターンの詳細を反映することが出来るためと考えられる。　しかし、より小さな部分構造は分子のサイズや構成についての情報を特定するのに有効である。　この実験では、一般的により大きな部分構造とより小さな部分構造を個別に用いるのではなく、これらの部分構造を適度に組み合わせたデータを用いた時に最良の類似化合物検索結果が得られることが見出された。

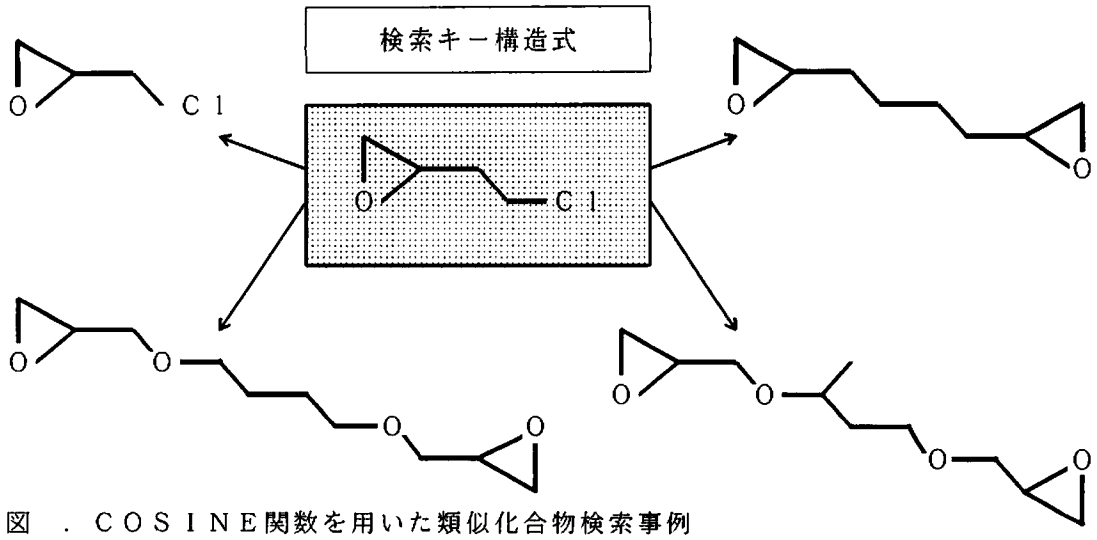


図　．ＣＯＳＩＮＥ関数を用いた類似化合物検索事例

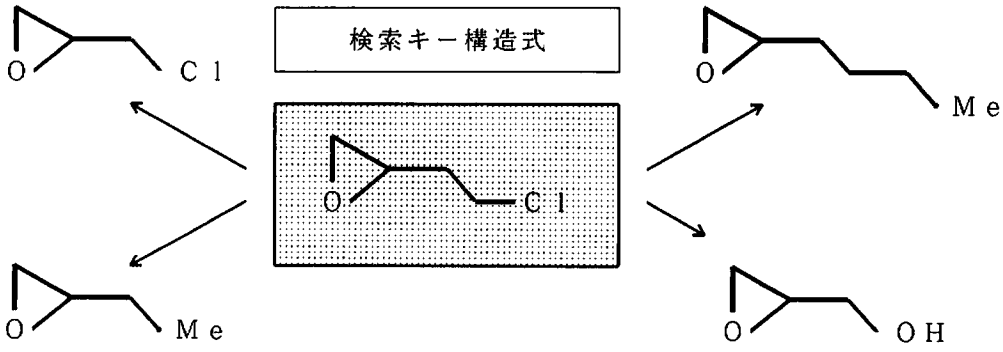


図　．　谷本関数による類似化合物検索事例

＊ Peter Willett,Vivienne Winterman and David Bawden," Implementation of Nearest -Neighbor Searching in Online Chemical Structure Search",
J.C.I.C.S. ,26 , 36-41(1986).

□　有機反応データベース（ＲＥＡＣＣＳ）における類似反応検索

米国ＭＤＬ（Ｍｏｌｅｃｕｌａｒ　Ｄｅｓｉｇｎ　Ｌｔｄ.）社が販売している反応データベースシステムＲＥＡＣＣＳ（Ｒｅａｃｔｉｏｎ　Ａｃｃｅｓｓ　Ｓｙｓｔｅｍ）を販売し、このシステムは世界中で利用されている。　このシステムでは類似反応検索を行なうことが可能である。　この検索技術については簇でに文献上に発表されているので、ここで簡単に解説する。

類似反応検索は単なる類似化合物検索と異なり、検索対象となる化合物は出発化合物群と成績体化合物群との複数を対象とすることが特徴である。　類似の基準により、さらに反応を行うための試薬等も考慮することも考えられる。

ここで説明するＲＥＡＣＣＳが採用している類似検索システムの基本は、検索対象を出発及び成績体化合物群に限定したアプローチである。　類似度算出の為の計算式としては谷本関数を用い、計算に用いるデータは部分構造の存在データを基本としている。　この部分構造データは、出発化合物群と成績体化合物群から得られるものと、ＲＥＡＣＣＳで定義している“反応中心（反応により構造式が変化する部分構造）”を対象として得られる部分構造データとの２種類で構成されている。

化合物構造式用の部分構造としては９３３個、反応中心用としては１１６３個を用意している。　反応中心用１１６３個のうち、２３０個は反応中心、及び反応中心から最大２結合分離れたところまでの環境を反映させる部分構造を利用している。　残る９３３個は出発化合物群と成績体との部分構造キーの論理ＯＲで得られるデータである。

検索キー構造式群と被検索化合物群との双方について部分構造の存在等がチェックされ、その結果が１（存在）および０（不在）で表現されるベクトルへと変換される。　このベクトルデータを用いて“谷本関数”を一部修正した式　を用いて類似度を計算する。

$$S = \frac{\Sigma W_k^2 D_{qk} D_{hk}}{\Sigma W_k^2 D_{qk} + \Sigma W_k^2 D_{hk} - \Sigma W_k^2 D_{qk} D_{hk}} \qquad (\ )$$

$$= \frac{C}{Q + H - C}$$

k　：部分構造のＩＤ
q　：検索キー構造式のＩＤ
h　：被検索化合物のＩＤ
D_{qk}：検索キー構造式ｑが部分構造ｋを含む時１、含まない時０
D_{hk}：被検索化合物ｈが部分構造ｋを含む時１、含まない時０
W_k　：部分構造ｋにアサインされたウエイト値
C　：検索キー構造式と被検索化合物とに共通に存在する部分構造ｋのウエイトの二乗和
Q　：検索キー構造式に存在する部分構造ｋのウエイトの二乗和
H　：被検索化合物に存在する部分構造ｋのウエイトの二乗和

この谷本関数の特徴は各部分構造ｋについて重みW_kを設けていることである。　従って、この重みの値の大小により類似度の値が大きく変化する。

ＲＥＡＣＣＳの類似度算出では、その算出基準となる部分構造データとして部分構造が存在する時１、存在しない時０のバイナリデータを用いている。　従って、化合物中に存在する部分構造の数をデータとして用いているわけではない。　この点で、大きい構造式を対象とした類似反応検索時には検索精度という点で不安がのこる。　しかし、ＲＥＡＣＣＳで用いている谷本関数による類似度の算出は、部分構造の存在情報というよりは、各部分構造にアサインされた重みW_kで行われている。　従って、この不安点は主として重みW_k値の決定方法でカバーされることになる。

重みの決定は種々の文献等を考察した結果、大きなデータベース内に存在する部分構造の“ユニーク”の程度を基準とすることが効率的であることが判明した。　この結果に従って部分構造の重みを決定していった。

算出の基準としたデータベースは反応データベースとして４６０００反応を、化合物データベースとして６４０００の化合物を用いた。　重みW_k値は以下の式に従って導かれた。　尚、W_kの値は出現頻度の低い（高度にユニーク）部分構造に類似度の評価が大きく影響を受けないように値は１～１０迄に制限をしている。

$$W_k = \sqrt{\frac{P_K}{P_{max}}} \qquad (\ )$$

W_k　：部分構造ｋのウエイト（１～１０）
P_K　：部分構造ｋの存在数
P_{max}：最も存在数の多い部分構造の存在数（約６０％）

・　ＲＥＡＣＣＳによる類似反応検索事例

ＲＥＡＣＣＳを用いて実際に類似反応検索を行った結果を図　に示す。　検索キーとしたのは有機リン化合物によるカルボアニオン生成によるカルボニルのオレフィン化反応であり、Ｗｉｔｔｉｇ反応に似た反応を検索キー反応式としている。

ＲＥＡＣＣＳでは反応検索条件として化合物の類似度に制限（０～１００％：０％は全く類似関係がないことを、１００％は類似関係が最も高いことを意味する）を設けて検索することが可能である。　この制限は化合物構造式群と反応中心とで個別に設定出来る。この例では化合物群に対して２０％、反応中心に対して８０％の制限を設けて類似反応検索をおこなっている。　従って、反応パターンは維持したまま適用化合物群に変化を期待した類似反応検索となる。

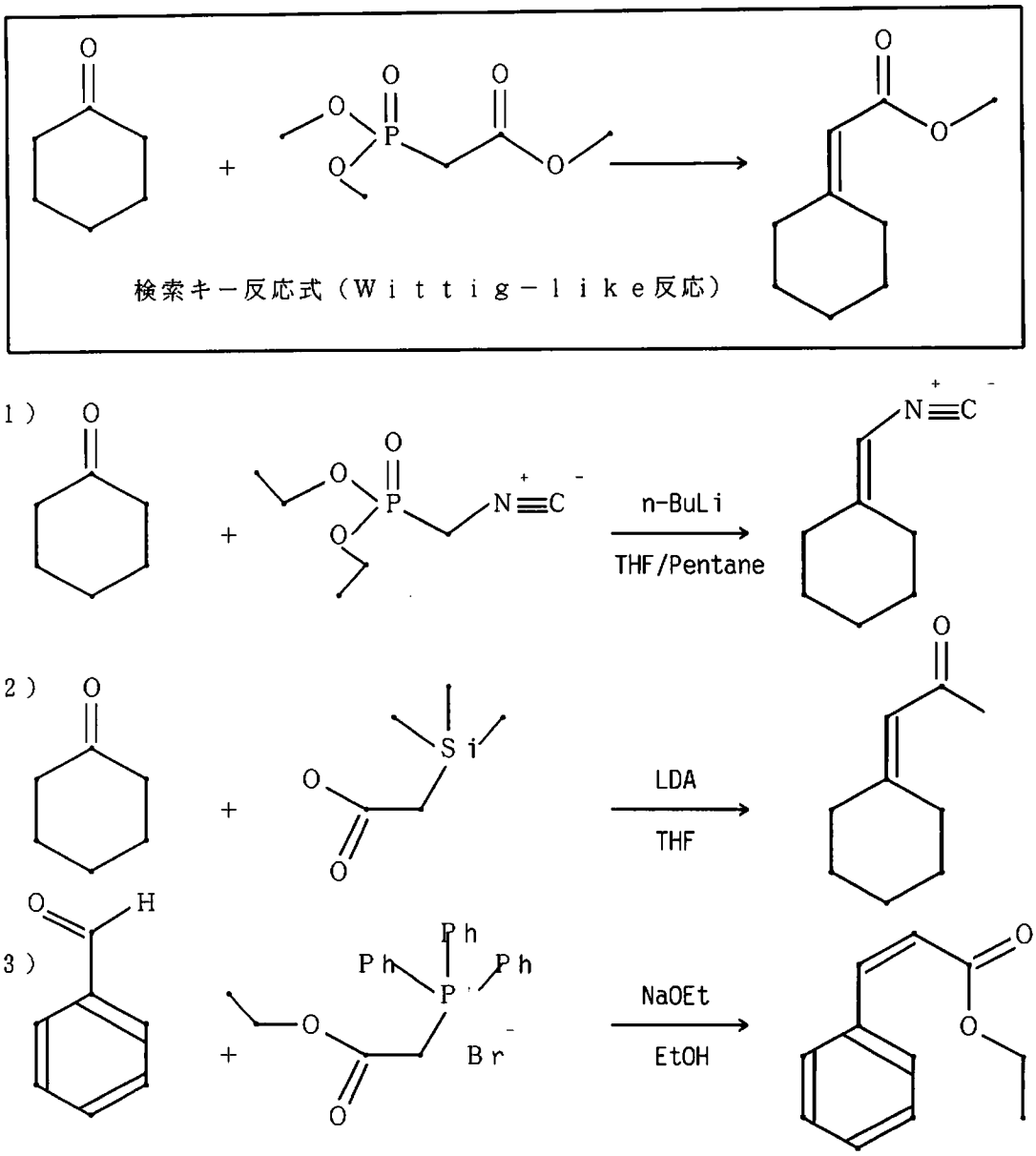


図　．反応類似性（８０％）、反応基質（２０％）の制限下でのヒットリスト。
１２ヒット中の３例

図　に示されるように、類似反応検索で取り出された反応は検索キーとした反応と充分に類似したものであることがわかる。　少なくともＷｉｔｔｉｇ反応と同じメカニズムで行く反応が選択されている。　また、３番目の反応は反応パターンとしては検索キー反応と同じであるが、化合物の構造式や反応部位は大きく異なっている。　これは、検索制限として設定した化合物群２０％、反応中心８０％の値が大きな効果を示した例といえる。

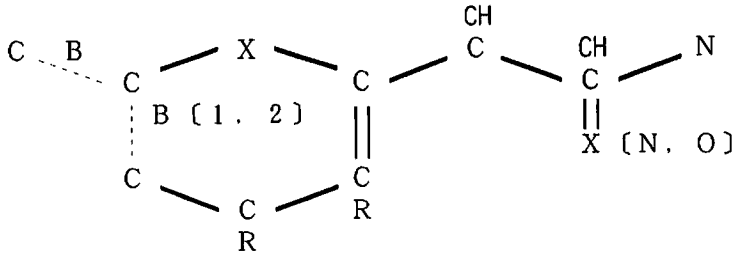
このように、類似度の制限をうまく利用することで検索目的に最適な反応群を効率的に取り出すことが可能となる。

□　総称（ＧＥＮＥＲＩＣ）検索

総称検索とは、化合物の検索キーのなかに原子および結合種に関するより包括的な検索条件（項目）を設けることで検索の制限をゆるめて検索を行うものである。　従ってこの検索は１対１の完全一致検索を行うものではなく、１対複数の検索を行うものである。検索においてこの総称検索は部分構造検索と並んで重要な検索形態の一つである。また、総称検索と部分構造検索は互いに独立して行われるのではなく、両方を組み合わせたかたちで実行されることが多い。

総称（ＧＥＮＥＲＩＣ）検索を行うためには、検索キー構造式中に総称的な情報が取り込まれていることが必要である。　この様式には様々なものがあるが、一般的には通常の原子表示には利用されえないアルファベット（例えばＸ、Ｒ等）や、特殊記号等を用いて行われる。　この総称検索に利用される典型的な検索キーの例を図　に示す。

総称検索には原子種や結合種の制限を解除（一部解除）したもの、逆にチェインやリング構造に関する制限を強化するもの等様々なものがある。　これらの検索条件は、それぞれの検索目的により使い分け利用することが必要である。



Ｂ；結合種は何でも良い
Ｂ〔１，２〕；限定付き結合種（単，２重結合）
Ｒ；リングを構成する原子
CH；チェインを構成する原子
Ｘ；水素以外の原子種
Ｘ〔Ｎ，Ｏ〕；制限付き原子種（窒素，酸素）

図　．ＧＥＮＥＲＩＣ検索用検索キー構造式例

図　に総称検索の事例が示されている。

総称検索の検索キー構造式を作成する時の指定方法は、結合種に関するものと原子種に関するものとに大きく２分類される。　原子種の場合、一般的には水素原子は対象外とすることが多い。　また、一部限定付きの検索を行う時は限定の内容についての指定を行うことが必要となる。　この指定は指定原子／結合に付随する括弧内に書き込むことで行われるのが多い。

検索の制限を緩める総称検索のアプローチとして原子種と結合種の２種類あることを示した。　総称検索にはこれら２種類の制限の他にも第３のアプローチとして、置換位置の指定制限解除がある。　この置換位置の制限解除はマーカッシュ（ＭＡＲＫＵＵＳＨ）検索という特別な名称で呼ばれている。　この検索手法については次節にて述べる。

□　マーカッシュ（ＭＡＲＫＵＵＳＨ）検索

特許等で多用される構造式表現手法を検索にも使えるようにするものである。　この表現形式は学術文献等においても頻繁に利用される形式である。　この表現形式も計算機を主体として考える時、多くの問題点を提供する検索手法となる。　このマーカッシュ検索は特殊な総称（ＧＥＮＥＲＩＣ）検索と考えられる。

１置換ベンゼン誘導体　　２置換ベンゼン誘導体

図　．マーカッシュ検索キー

図　にマーカッシュ方式による化合物表示の例がしめされている。　この図の左の構造式は１置換ベンゼン体を表現し、Ｒとして何らかの置換基が存在していることを示す。また、右側の構造式は２置換ベンゼン体を示し、Ｒ１、Ｒ２の２個の置換基がついていることを示している。　このような表現式を化合物の検索キー構造式として用いるのがマー

カッシュ検索である。
総称（ＧＥＮＥＲＩＣ）検索（　章　節）と異なり、置換位置や置換基の種類が特定されない形での検索を行うものであることがわかる。

□　互変異性体（ＴＡＵＴＯＭＥＲ）検索
化合物の互変異性体は化学者には自然に受入れられる考えである。　しかし、１対１の対応を基本とする計算機の世界では互変異性体の扱いは極めて厄介なものとなる。
互変異性体が存在する時、その存在を意識せずに検索キーを作成すると、例え同じ（化学的に）化合物であっても異なる化合物として認識されてヒットしないことが頻発する。これは計算機内部では化合物構造式を単なる図形として認識し、図形のマッチング問題として検索を行っている為である。　従って構造式（図形）の異なる化合物は異なる化合物として認識され、化学的には同一化合物であっても検索ではヒットされなくなる。
互変異性体検索ではこのような問題を解決することが要求される。

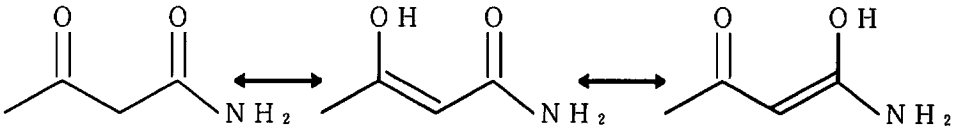


図　．ケトーエノール互変異性体

最も典型的で基本的な互変異性体はケトーエノール互変異性体であるが、システムが互変異性体をカバーする機能が無い時には検索キー構造式を工夫する必要が生じる。　例えば図　のようなジケトン体を検索するには検索もれをなくす為には３種類の構造式を検索キーとしなければならない。

・共鳴構造体の問題
互変異性体の検索には特殊な技術が必要と理解できても、共鳴構造体の検索が互変異性体検索と全く同じ問題を抱えていることを認識できる化学者は少ないであろう。　化学に親しめば親しむほどこの傾向はつよくなる。
計算機に近い立場から化学を眺めた時、共鳴構造体の問題も、化合物の検索という点では厄介な問題である。　例えば最も身近な例ではベンゼン環の問題があり、２重結合の書き方により２種類の図を描くことが出来る。　このベンゼン環の問題を拡張すると芳香族縮合多環化合物に対しては多数の共鳴構造体が存在し、検索キー構造式もこれらの共鳴構造体の数に対応するだけ必要となる。　またこれ以外にも、ニトロ基は窒素の原子価を５とした表現と＋／－のチャージを持たせた表現の２種類が存在する。　これらの構造式もシステムに何らかのサポート機能が無い限り、同一物であると計算機で認識することは困難である。
現在実用化されている大部分のシステムでは、ベンゼン誘導体や簡単な官能基レベルでの共鳴構造体に対する保証機能を備えている。　しかし、必ずしも総てのシステムがこの保証機能を備えているとは限らないので化合物データベースを用いる時はこの点での留意が必要である。　この問題を避けるためには、データベース内にストアされている化合物構造式（特に共鳴構造体）の表現方法は統一されていることが必要で、そのデータベースを用いて化合物検索を行う人はデータベース毎に定められた共鳴構造体のパターンを認識していることが必要である。

図　．共鳴構造式による複数表現

□　共通部分構造(SUPER STRUCTURE) 検索
複数化合物群に共通して存在する部分構造を検索して特定する検索手法である。　このアプローチはいままで述べてきたような、検索キーの持つ条件を満たす化合物／化合物群をデータベースから取り出すものではなく、選ばれた化合物群（データベース等）に共通に存在する部分構造をあらたに探し出すというものである。　従ってこの検索の場合、データベースは単に解析対象となる化合物をストアしているだけのものであり、最終的な

検索目的は共通に存在する部分構造である。　このため、検索当初では検索の回答となる部分構造は見えていない。
　このように共通部分構造検索はかなり特殊な検索様式であるが、目的を限定して利用することで強力な検索様式となる。　例えば、構造－活性・物性相関等の研究分野では、活性を出現させる為の重要部分構造を活性化合物群の中から見出す等の目的で利用されている。　残念ながら、実用的レベルでこの共通部分構造検索を実現したシステムは存在せず、研究レベルに止まっているのが現状である。

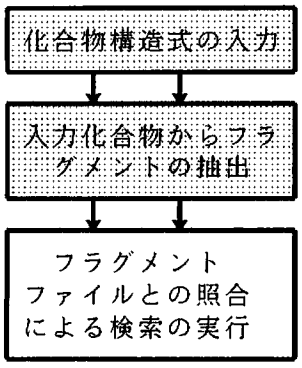
④　今後の化合物検索様式（ＡＩ（人工知能）検索）
　今後の検索様式として様々なものが考えられる。
　検索対象となる化合物構造式は３次元であることは常識となるであろう。　さらには、化合物検索とその他の解析業務との連繋が簡単にとれることも重要なポイントとなってくるであろう。　従って、解析で得られた候補化合物に対する関連データの収集等はその場でデータベース検索を行うことで簡単に行うことが可能となる。

・ＡＩ（人工知能）を利用した化合物検索
　最近、ＡＩ（人工知能）関連の技術が急速に進展してきた。　従来手法では解決が困難な分野への人工知能の積極的な利用がはじまり、その成果が様々な分野で報告されるようになった。　計算機化学と人工知能との関係は古く、人工知能の解説書等で最初の実用人工知能システムとして必ず引用されるシステムは以外にも化学分野における応用システムであり、このシステムは種々スペクトルデータを用いて化合物の構造式を推定する“ＤＥＮＤＲＡＬシステム”である。　これは、このスペクトル解析問題は方程式等で規定されることが少なく、ルール主体で物事を決定する事が多いという特徴が人工知能研究者の恰好の標的となったようである。　人工知能分野ではバイブル的存在であるこのＤＥＮＤＲＡＬも、化学者自体にはあまり受け入れられることなく、その人工知能技術の展開という役割を遂げた後、この開発プロジェクトも解散した。
　現在、化学分野における人工知能の利用という点では華々しい成果をあげているものはすくないが、この技術は高いポテンシャルを持つことは確実である。　今後、計算機化学の基礎が確定し、従来の技術では補いえない分野／技術上の問題が出てきた時この人工知能技術は重要なものとなろう。

・化合物検索における人工知能技術の適用
　化合物検索にもこの人工知能技術を利用することは大いに考えられることである。　このＡＩ検索により、人間が行う高度な検索を計算機が自動的に実行出来るようになる。
　ＡＩ（人工知能）技術は検索の様々なステージでの利用が可能である。　実際に化合物検索を検索のエキスパートが行った場合と素人が行った場合とで検索のヒット率に大きな差がでることも明白であり、実際にＣＡＳオンラインでの検索ではその利用料に大きな差が出てくることになる。
　このような検索エキスパートのノウハウをＡＩ（人工知能）にのせた形での化合物検索が行えるようにするならば、初心者でも効率よく目的とする化合物や関連情報を取り出すことが可能である。
　このような検索ノウハウをＡＩにのせるアプローチの他に、検索手法自体にＡＩの技法をのせるアプローチが存在する。　このようなＡＩ検索としては類似化合物検索が考えられる。　即ち、類似という高度な判断事項をＡＩにまかせてしまおうということである。類似化合物検索は、検索する研究者自身の研究歴、対象とする被検索化合物の領域、検索内容等様々な条件により検索結果が異なってくる。　この変化性が類似化合物検索の実現を困難にする要因である。　現在実用化されている類似化合物検索ではこのような変化性をある程度犠牲にする事で実現されている。　ＡＩ検索ではこのような変化性の犠牲を最少限に押さえた形での検索を期待することが出来る。　しかし、この実現の為にはさらに多くの研究の積み重ねと技術の蓄積が必要であろう。

⑤　化合物検索技術と計算機に由来した検索技術との融合
　化合物の検索は計算機化学における基本である事は先にも述べた。　この検索技術は化学に関する様々なシステムを開発する時に必要となってくる。
　計算機による化合物検索が困難であった理由は、主として化合物がトポロジカル及びトポグラフィカルな情報を基本とするのに対し、計算機による検索技術はこの種の情報の扱いに不得手であった事に起因する。
　この章で述べた技術は化合物検索に必要となる化合物のコード化及び化学特有の検索オプションに対応するものである。　しかし、通常のシステムではここで述べた技術の他に計算機に由来する様々な技術との組み合わせを行って初めて効率的な検索が可能となる。

化合物はトポロジカルであっても、一旦それが数値データへと置き換えられたならばその後は計算機本来の検索技術を十二分に利用することが可能である。
計算機に由来する検索技術は第　章でも述べたように様々なものがある。　これら性格の異なる2種類の技術を融合する事が、より高度な検索システムへと導く基本である。



・入力構造式から結合表への変換
・部分構造検索によるフラグメントの抽出
・最も検索効率の高いフラグメントの選択
・インバーテッドファイル作成
・ハッシング　・ソーティング
・ＡＮＤ／ＯＲ検索

図　．フラグメント検索を行う時の簡単な作業フロー、および関連技術

化合物検索で最も頻繁に利用される部分構造検索を例にとって説明する。　図中点線部分は化学に起因する特有な技術が中心となる部分であり、その他の部分は計算機に依存して発展した技術である。　この図からもわかるように、効率の良い化合物検索システムを構築するには両方の技術の融合が必須である事がわかる。

④　ＣＡＳ／ＤＡＲＣ　等における化合物検索技術

ＣＡＳやＤＡＲＣでは大量（数百万）の化合物検索を行う必要から独自検索システムを開発しながらオンライン検索サービスを行っている。　しかし、これら大規模検索システムといえども、その中で利用している技術は先に述べた技術の延長上にあるものである。
ＣＡＳで行っている技術のうち主なものについて簡単にのべる。

(1)　計算機上での工夫

ＣＡＳは一台の大型計算機の配下に数多くのミニコンピュータが配置され、実際の検索はこのミニコンピュータ上で検索が実行されるようになっている。　大型計算機は検索に必要となる検索化合物の入力、入力された化合物の検索キーへの変換、検索結果の出力、ＪＯＢ管理、オンライン管理等の作業を行い、実際の検索作業は行っていない。　検索化合物が大型計算機に入力されると検索に必要となる検索キーを発生し、その検索キーを配下のミニコンピュータに同時に入力し、検索は多数のミニコンピュータ上で平行して実行される。　従って、ミニコンピュータの台数だけ検索スピードが向上し、且つその台数だけ化合物ファイルの大きさを小さく出来る為さらに検索スピードが向上する。　化合物数が増えてもミニコンピュータの台数をふやす事で検索スピードを保つ事が出来る。　また、化合物ファイルのアップデートはファイル容量が小さい為に容易である等様々な利点がある。　このシステム構成は基本的には現在における最新の技術である並列型コンピュータの思想をかなり前から実現していたことを意味する。
システムは化合物数の増大があってもミニコンピュータの数を増やす事で簡単に現状のレスポンスを保証する事は可能である。　しかし、この構成は利用者数の増大に対しては効果は期待出来ない。　このユーザ数増大への対応は個々の計算機の能力向上、検索スピードの向上等で対応する事が必要である。　以下に簡単にその特徴をまとめ、そのシステム構成図を示す。

表　．　ＣＡＳのシステム構成を取ることによるデータベース上での利点

ＣＡＳシステム構成による利点	Ｎ台のミニコンピュータ
1．検索スピードの向上が簡単に実現可能	検索速度はＮ倍
2．検索スピードがファイルの大きさに左右されない	ファイルを一定の大きさにする
3．化合物ファイルのアップデートが簡単である	アップデートはＮ分の１

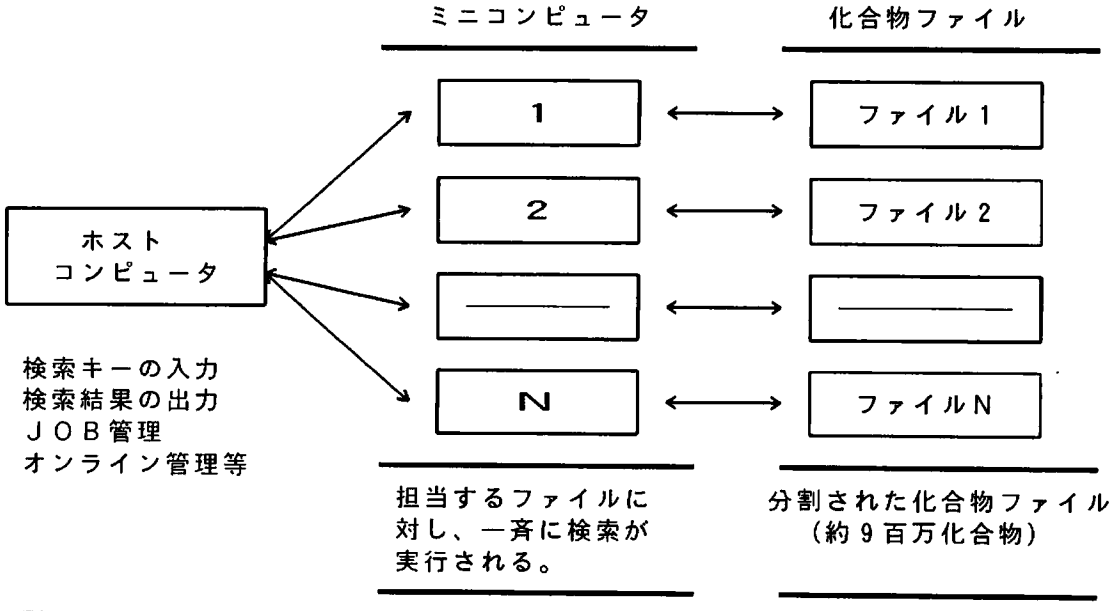


図　．　ＣＡＳにおけるシステム構成図

＊ Robert E. Stobaugh," Chemical Substructure Searching",
J.C.I.C.S,25,271-275 (1985)

(2) 検索システム上での工夫

ＣＡＳ　ＯＮＬＩＮＥ　で利用可能な検索様式にはさまざまなものがある。　この中でも頻繁に利用される部分構造検索についてのべる。

ＣＡＳにおける部分構造検索の基本はフラグメントを多数用いたフラグメント検索である。　このフラグメント検索は、フラグメント１個について一個のインバーテッドファイルを作成し、このファイルを用いて高速な検索を実現するものである。　但し、検索精度の保証と数百万化合物という大量の高速検索を実現するために、通常の検索システムと比べると１、２桁も大きい数（数千）の部分構造フラグメント（以下フラグメントと略す）を用意している。　１９８２年の報告ではこの部分構造フラグメントは総数２０４７（但しバージョン３）個用意されている。

表　にはＣＡＳの化合物検索で利用されている部分構造フラグメントの種類と数が示されている（但しバージョン３）。

＊W.Graf,H.K.Kaindl,H.Kniess and R.Warszawski," The Third BASIC Fragment Search Dictionary", J.C.I.C.S., 22 ,177-181 (1982)

表　．　ＣＡＳで用いられるフラグメントリスト（バージョン３）

TYPE OF FRAGMENT	TOTAL OF ALLOCATED FRAGMENT NUMBER
1．LINEAR SEQUENCE (LS)	
ATOM SEQUENCE (AS)	512
CONNECTIVITY SEQUENCE (CS)	215
BOND SEQUENCE (BS)	207
2．AUGMENTED ATOM (AA)	
AUGMENTED ATOM, GENERAL (AA)	16
AUGMENTED ATOM, SPECIFIC (AA)	757
HYDROGEN AUGMENTED ATOM (HA)	113
TWIN AUGMENTED ATOM (TW)	17
3．RING	
RING COUNT (RC)	10
TYPE OF RING (TR)	51
4．OTHER FRAGMENT TYPES	
ATOM COUNT (AC)	19
ELEMENT COMPOSITION (EC)	111
GRAPH MODIFIER (GM)	19
TOTAL	2047

ＣＡＳで用いられている２０４７個のフラグメントは大きく４種類に分類される。　これらはそれぞれ、①ＬＩＮＥＡＲ　ＳＥＱＵＥＮＣＥ（ＬＳ）：原子の線型結合状態に関する情報、②ＡＵＧＭＥＮＴＥＤ　ＡＴＯＭ：ある特定の原子に注目した時、その原子のまわりの環境に関する情報、③ＲＩＮＧ：化合物内の環に関する情報、④ＯＴＨＥＲ　ＦＲＡＧＭＥＮＴ　ＴＹＰＥＳ：その他の関連情報となっている。

以下の節ではこれらフラグメントのうち代表的なものを示す。　表　にはこれらフラグメントの表記に用いられる記号が示されている。

表　．フラグメントの表現で用いられる記号リスト

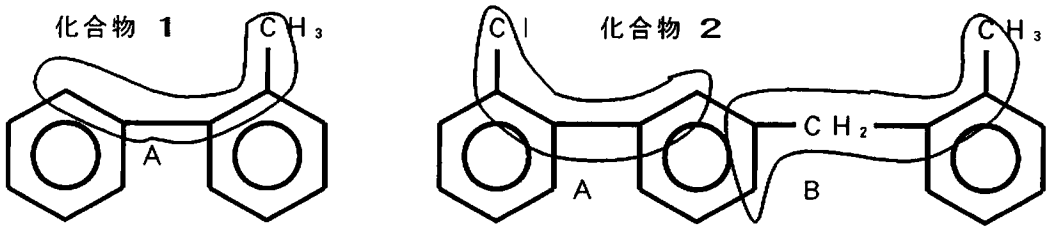
DEFINITION OF THE SYMBOLS	
A	= ANY ATOM EXCEPT HYDROGEN
D	= NON FUSION NODE(cf.TYPE OF RING SECTION)
G	= COMMON SYMBOL FOR F,Cl,Br,I
M	= COMMON SYMBOL FOR METALS
T	= FUSION NODE(cf.TYPE OF RING SECTION)
Y	= COMMON SYMBOL FOR O,S
Z	= COMMON SYMBOL FOR NONMETALS B,Si,P,As,Se,Te
$	= SYMBOL INDICATING THAT SEVERAL FRAGMENTS HAVE THE SAME FRAGMENT NUMBER
*	= RING BOND
−	= CHAIN BOND
1，2，3，4	= BOND VALUE

1．フラグメントの具体的事例

先に述べたフラグメントについて、以下には実際の化合物を用い、フラグメントがどのように利用されるかについて簡単に述べる。

1．1．ＬＩＮＥＡＲ　ＳＥＱＵＥＮＣＥ（ＬＳ）事例

□　ＡＴＯＭ　ＳＥＱＵＥＮＣＥ（ＡＳ）による検索



上記化合物１と２について部分構造フラグメントとして線で囲まれた部分を取り出すとする。　この時これらのフラグメントに相当するものがフラグメント辞書内に存在するか否かをチェックする。

複数原子の直線的な結合関係に関する情報はＬＩＮＥＡＲ　ＳＥＱＵＥＮＣＥ（ＬＳ）に登録されており、この中で結合情報を無視した原子だけの結合関係はＡＴＯＭ　ＳＥＱＵＥＮＣＥ（ＡＳ）にある。　このＡＳフラグメント５１２個の内、６番目のフラグメント（Ｓ６）を取り出すと、このＳ６は更に４個のフラグメントから構成されている事がわかる。　これら４個のフラグメント中、化合物１のＡに該当するフラグメントはＳ６の①である事がわかる。　以下同様に、化合物２のＡはＳ６の②に、ＢはＳ７の②に該当する事がわかる。（Ｇ，Ｎ，Ｙ，＊等の定義に関しては表　を参照）

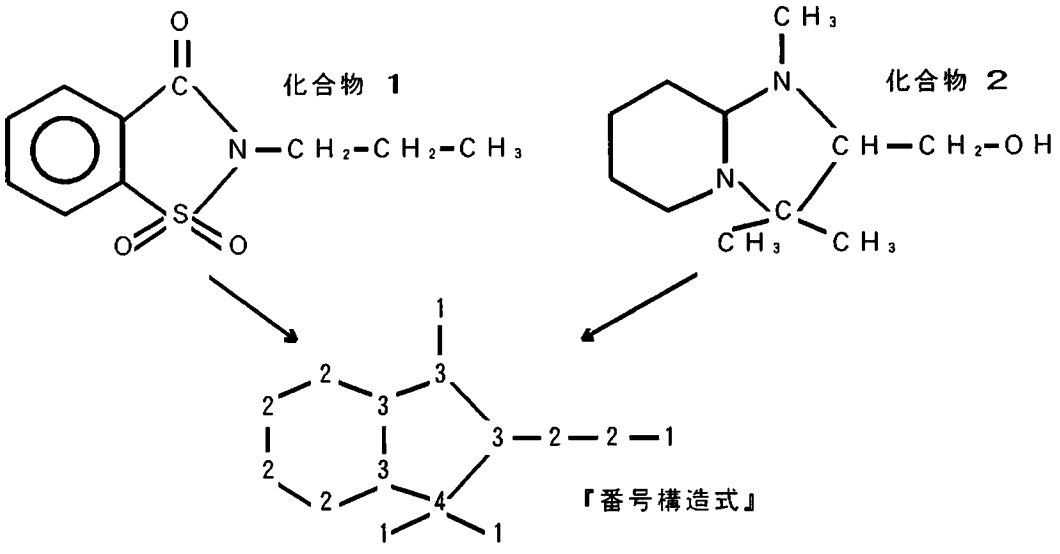
ＡＳ＝Ｓ６＝
① Ｃ＊Ｃ－Ｃ＊Ｃ－Ｃ　⟶ 化合物１のＡ
② Ｃ＊Ｃ－Ｃ＊Ｃ－Ｇ　⟶ 化合物２のＡ
③ Ｃ＊Ｃ－Ｃ＊Ｃ－Ｎ
④ Ｃ＊Ｃ－Ｃ＊Ｃ－Ｙ

ＡＳ＝Ｓ７＝
① Ｃ＊Ｃ ── Ｃ＊Ｃ－Ｃ
② Ｃ＊Ｃ－Ｃ－Ｃ＊Ｃ－Ｃ　⟶ 化合物２のＢ
③ Ｃ＊Ｃ－Ｎ－Ｃ＊Ｃ－Ｃ
④ Ｃ＊Ｃ－Ｙ－Ｃ＊Ｃ－Ｃ

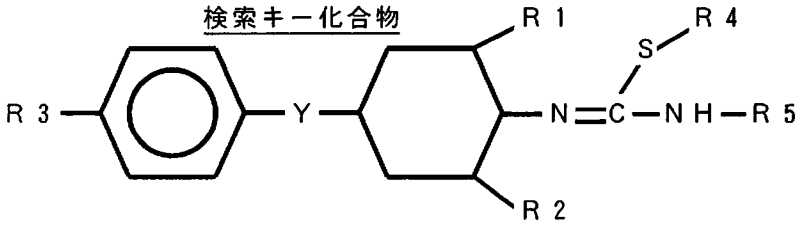
□ CONNECTIVITY SEQUENCE（CS）による検索

前記ASは原子の種類を考慮するが、原子の結合（特に分岐）に関する情報は単に直線的なものだけである。 これに対しCSは原子の種類を無視し、一方で原子が他の原子と結合している状態（分岐等）に関する情報をまとめたものである。

手順1：構造式中の原子の種類を無視し、原子が水素以外の原子と結合している数をチェックし、その番号で原子と置き換えた構造式（仮称『番号構造式』）を作成する。



手順2：検索キー化合物を手順1に従い総ての原子を結合番号に置き換える。



CONNECTIVITY SEQUENCE（CS）は番号で表現された原子の中、連続する3～6個の数値を数多く取り出したものであり、検索キー化合物について取り出されるCSの数に制限は無い。

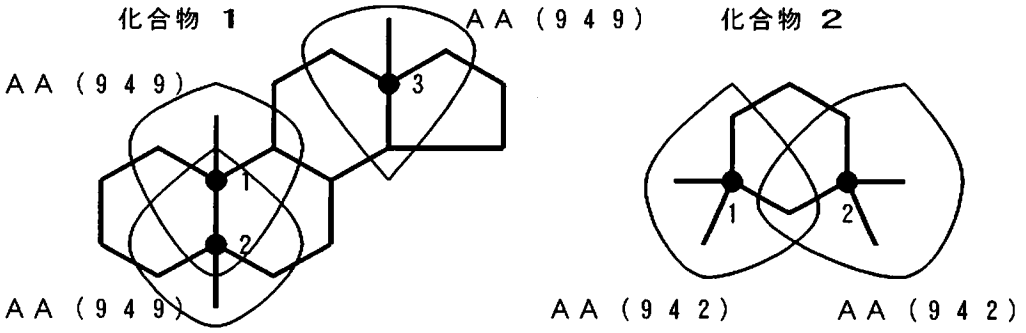
検索キー化合物に該当するCS ＝

6 3 0 ＝ 2＊3－2－3＊2
6 4 3 ＝ 2＊3＊2＊3＊3＊3
6 1 2 ＝ 2＊2＊3＊2＊2＊3
6 3 8 ＝ 2－3－2－3＊3
6 3 0 ＝ 2－3＊2＊3＊3－2

連続する3～6個の数値

１．２．　ＡＵＧＭＥＮＴＥＤ　ＡＴＯＭ　事例

□　ＡＵＧＭＥＮＴＥＤ　ＡＴＯＭ，ＳＰＥＣＩＦＩＣ（ＡＡ）による検索



化合物中の各原子に注目し、その原子が結合している総ての原子に関する情報をまとめて数え上げるものである。　形式は最初に原子の環境パターンで一つの化合物中に含まれる同じ環境の原子の数が来る。　数値データの次に、現在対象とする原子の情報が来る。その原子に続いて、該当原子に直接結合している原子に関する情報が総て連続して表記される。

ＡＡ＝（パターンの数）、（該当原子）、（該当原子に直結する原子の結合情報）

化合物１に対するＡＡ　＝　９４９　＝　３　Ａ　＊Ａ＊Ａ＊Ａ－Ａ
化合物２に対するＡＡ　＝　９４２　＝　２　Ａ　＊Ａ－Ａ－Ａ

１．３．　フラグメントによる一般的検索事例

フラグメントを用いた検索としては部分構造検索が一般的である。　ここではＣＡＳのデータベースの部分構造検索事例について説明する。　また、このフラグメント検索を利用したより高度な検索である“類似部分構造検索”について簡単に述べる。

１．３．１．部分構造検索事例

化学研究においては完全一致検索と並んで部分構造検索を利用する事が多い。　これは、ある共通な部分構造を持つ化合物群や、指定された化合物と似た構造を持つ類似化合物群等、一連の共通な構造的特徴を持つ化合物群を幅広く取り出したいという要求が多いことが一因である。　さらに検索様式という観点から考えた時、計算機にしか出来ない検索様式として部分構造検索がある。　つまり、化合物構造式が決まったものを探し出すことは人間にとってもさほど困難な作業ではないが、共通な部分構造を持つ化合物群を多数の化合物中から探しだすということは人間にとって殆ど不可能な作業である。

ＣＡＳを用いた部分構造検索事例としてβ－Sympatholytic activityを有する可能性のある化合物の検索を例にとって説明する。　一般的に、β－Sympatholytic activityを有する為の必要条件として図　に示された部分構造を有している事が必要という事実がしられている。　従って、この部分構造を有する化合物を化合物データベースの中から取り出す事で、新たなβ－Sympatholytic activity化合物の発見につながる事が期待出来る。このような検索を行う時には部分構造検索が必須である。

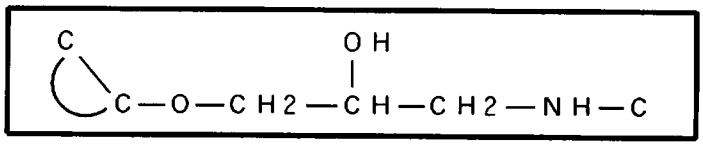
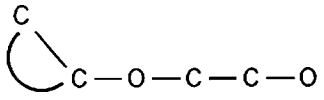
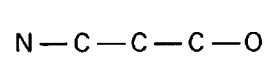
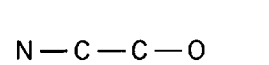
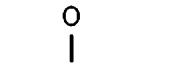
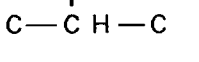
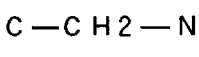
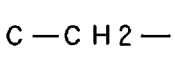


図　．β－Sympatholytic activity発現の為の基本部分構造

図　の部分構造に対し、ＣＡＳで用意されている２０４７個のフラグメントの中、部分構造検索に必要と思われるフラグメントが８個抽出された。（表　．）

部分構造　に関しては、この８個のフラグメント総てを用いたフラグメント検索で部分構造検索が実行される。　この時ヒットする化合物はその構造式内に図　の部分構造を含む化合物であり、β－Sympatholytic activityを有する可能性の高い化合物である。ＣＡＳ　ＯＮＬＩＮＥ　ｓｅｒｖｉｃｅ（１９０万化合物、１９８１年１月）を用いた実験では３４７４個の化合物が取り出された。

表　．　部分構造検索に用いられた８個のフラグメント

ACS	Fragment 1 No. 187	
	Fragment 2 No. 378	
	Fragment 3 No. 405	
CS	Fragment 4 No. 613	
HA	Fragment 5 No. 1149	
	Fragment 6 No. 1235	
	Fragment 7 No. 1297	
	Fragment 8 No. 1597	

１．３．２．類似部分構造検索への展開

部分構造検索で用いた８個のフラグメントのうち、少数のフラグメントを除いた残りのフラグメントを用いて検索すれば図　の部分構造と類似の部分構造をもつ化合物の検索をおこなったのと同じ効果が得られる。

先の部分構造検索と同様にＣＡＳ　ＯＮＬＩＮＥ　ｓｅｒｖｉｃｅを用いて前記８個のフラグメントを用いた検索を８種類行っている。　つまり、８個のフラグメントから１個を除いた残りの７個のフラグメント総てを用いて、総計８回の検索を行っている。

表　には８個のフラグメント総てを用いた検索と７個のフラグメントを用いた８回の検索により得られたヒット化合物数及び検索ヒットされた３４７４化合物の中、個々のフラグメント単位で識別される化合物数（抽出化合物数＝８個のフラグメントによる検索ヒット化合物数－７個のフラグメントによる検索ヒット数）が示されている。

表　．　８及び７個のフラグメントを用いた検索結果

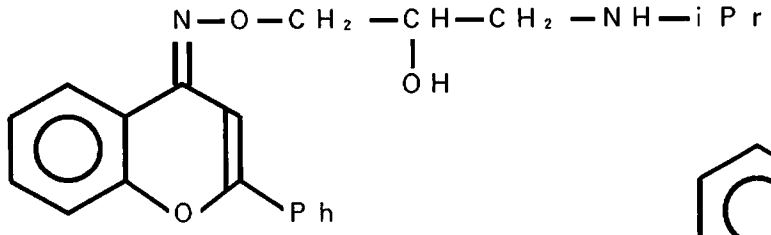
フラグメント数	フラグメントの種類	検索ヒット化合物数	抽出化合物数
８（総て）	１２３４５６７８	３４７４	
７	□２３４５６７８	３７６０	２８６
７	１□３４５６７８	３４９６	２２
７	１２□４５６７８	３５４１	６７
７	１２３□５６７８	３４８１	７
７	１２３４□６７８	３４８６	１２
７	１２３４５□７８	３４７４	０
７	１２３４５６□８	３４８１	７
７	１２３４５６７□	３５６９	９５

フラグメント１を除いた残りのフラグメントで検索を行った時は３７６０個の化合物が取り出されている。　一方、フラグメント１を加えた全フラグメントによる検索では３４７４個の化合物が取り出されている。　この結果、フラグメント１により２８６個の化合物が取り除かれている事がわかる。　一方、フラグメント６では一個も取り除く事が出来

ないでいる。　この事実はフラグメント６の持つ情報は他のフラグメントで充分補われている事を示している。

図　の２化合物は７個のフラグメントによる検索ではヒット化合物として取り出されるが、８個のフラグメントによる検索では取り除かれた化合物中、代表的なものが示されている。

Ｒｅｇ．Ｎｏ．
66928-88-9



Ｒｅｇ．Ｎｏ．
71106-77-9

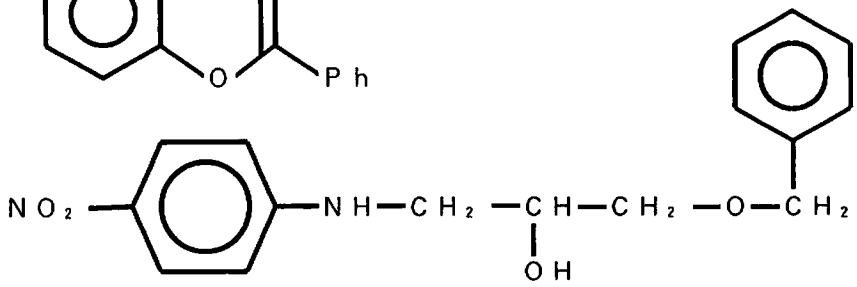


図　．　７個の７フラグメントではチェックされない構造式

(3)　ＣＡＳにおける化合物検索システムについて（まとめ）

以上ＣＡＳで利用されている化合物検索システムについて、ハード上での観点とソフト上での観点の両方から簡単に考察してきた。　この他にも様々なタイプのフラグメントが利用されて検索効率向上の為に利用されている。　ここでは総てのフラグメントについて解説する事はしない。　関心がある人は元著を参照されたい。

ＣＡＳは単に化合物構造式を計算機で扱うという問題の他、極端な迄の多数（約９００万）の化合物を扱わなければならない為に通常展開されている化合物検索システムよりも数多くの工夫が盛り込まれている事がわかったかと思う。　しかし、ここで利用されている技術はスケールの大きさと、それに伴う細々とした工夫に関する部分を除くならば個々の基本技術は他のシステムで利用されている技術と大幅に異なる事は無いということも明白である。

1．　部分構造検索

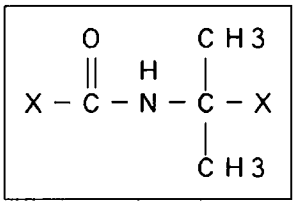
部分構造検索は化合物の部分構造を検索キーとし、構造式中に検索キーに用いた部分構造を含む化合物群を取り出してくる検索手法である。

この部分構造検索は完全一致検索と異なり、つねに一個以上の化合物を検索ヒット化合物群として化合物データベースから取り出してくる。　従って部分構造検索の目的は、目標とする化合物を高速にデータベースから取り出すという完全一致検索と異なり、ある共通の要因（部分構造）を構造式内に持っている化合物群を取り出すことである。　この共通要因は単純な分類目的の構造要因である時もあれば、構造－活性相関を目的とする薬理作用団(PHARMACOPHORE)であったりする。　従って、部分構造検索はライブラリ的利用というよりも研究開発を主目的とした利用に供されることが多い。

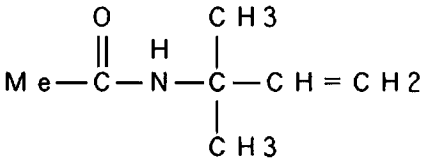
図　に部分構造検索の例を示す。　検索用キーとして化合物の部分構造が利用される。検索キー構造式中、Ｘでしめされた原子は水素以外の原子を表す。　この検索キーを用いて部分構造検索を行う時、化合物構造式中１及び２には該当する部分構造は含まれていない。　１はケトンがチオケトンとなっており原子種の違いでヒットしない。　２は被検索化合物構造式自体が検索キー構造式よりも小さいため検索キー構造式を包含することが出来ない。

化合物構造式３及び４は検索キー構造式を包含している（太線部分）。　４の構造式では検索キー構造式を環構造の一部としてとらえている。　このように、部分構造検索は被検索化合物が複雑になるほどその検索効果が高くなる。

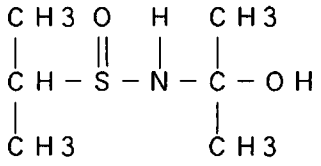
検索キー構造式



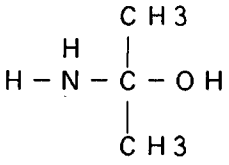
3．



1．



2．



4．

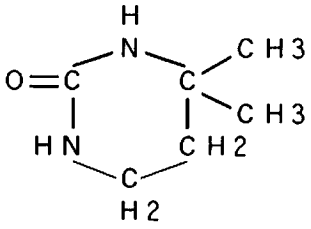


図　．部分構造検索事例

要チェック⇨　K(x)= [{Rk(x) ∈R (X) } ∈D (R (X))] ――― ( )

この部分構造検索をおこなう手法として現在様々なものが提唱されている。　ここでは部分構造検索を行う時に一般的に行われている２種類のアプローチについてのべる。　一つはフラグメント検索で、通常はこのフラグメント検索だけで部分構造検索を完全に完了することはできない。　この検索は検索速度が比較的おそい部分構造検索の検索速度向上のための１次検索として利用されるのが一般的である。

もう一つは部分構造検索そのものを行うもので、ここでは“セットリダクション法”と呼ばれている手法についてのべる。　この手法のアルゴリズムが発表されたのは比較的古いが、部分構造検索の基本となるものであり、現在でも利用されているので特に取り上げて説明する。

(1)　フラグメント検索（あらかじめ個々の化合物に関するフラグメントの情報をまとめておく方法）

□　フラグメント検索の特徴

フラグメント検索は、化合物の小さな構造単位であるフラグメントの存在情報を用いて化合物検索を行うものである。　この検索の特徴は後で述べる部分構造検索よりも検索速度が早いことである。　しかし、検索単位が構造的に小さなフラグメントで行ない、総ての原子に関する完全一致を行うわけではない。　このため、フラグメント検索の検索精度は高くない。　以上の特徴からフラグメント検索は単独で利用されることは少なく、後に述べる部分構造検索に先立つ１次検索手法として利用されることが多い。

□　フラグメント検索の内容および手続き

フラグメント検索は実際の検索に先立ち、データベース中に保存されている個々の化合物についてフラグメントの存在をチェックし、これらのフラグメント単位で検索用の転地（インバーテッド）ファイルを用意する。

実際に化合物を検索する時の手順は以下のようになる。

①検索キー構造式中に存在する総てのフラグメントを取り出す。

②取り出されたフラグメント単位でデータベース内の転地ファイルを参照し、ヒット化合物群を取り出してくる。

③②で取り出されたヒット化合物リストに関し、ＡＮＤ検索を行う。

④検索に用いた総てのフラグメントが存在する化合物で最終ヒットリストを作成する。

□　フラグメント検索の具体的内容

フラグメント転地ファイルの構造を表　に示す。　この表には各部分構造やフラグメントを構造式中に含む化合物のリストが登録されている。　これにより、データベース内の化合物がどの様な部分構造／フラグメントから構成されているかがわかる。

表　．フラグメント検索用転地（インバーテッド）ファイル例。

NO.	フラグメント	ヒット化合物ＩＤリスト
1．	－ＣＨ３	3 6 7 10 15 16 18 19 …… 999
2．	－ＮＨ２	12 23 45 78 92 123 132 …… 970
3．	＝ＣＨ－	4 5 8 12 17 19 22 25 …… 996
	……	……
N．	－Ｃ６Ｈ５	2 8 11 14 18 23 25 31 …… 996

□　フラグメントについて

検索の効率という観点からは検索に用いるフラグメントをどのようなものとするかが大切な問題となる。　検索対象化合物群中に存在しない、あるいは出現頻度の低いフラグメントを用いても検索効率は上がらない。　また、逆に出現頻度が高いフラグメントを用いても同様に検索効率の向上は期待出来ない。　適度の出現率を持ち、しかもその出現パターンがフラグメント間で重なることの少ないものが理想である。　この出現率や出現の重なり程度、さらには用いる部分構造／フラグメント数についてどの程度あれば良いという目標値は残念ながら発表されていない。　経験的にケースバイケースで行われているのが現状である。

理想的にはデータベース単位で最適な検索用フラグメントを用意するのが良い。　部分構造やフラグメントの出現頻度等を参照しつつ、最小限度の部分構造／フラグメント数を用いて、最大の検索効率が期待できるフラグメント群を取り出す機能が必要である。　実用システムにはこの様なフラグメント選択機能を基本機能として備えているものもある。

一般的に利用されているデータベースでは数千～数万の化合物を検索対象とする場合、数百～千程度のフラグメントを用意し、数万以上の化合物を扱う大規模なデータベースでは数千以上、場合によっては万程度のフラグメントを利用して高速検索を実現している。また、用意するフラグメントの内容も個々のデータベースでそのシステムに最適なものであるように工夫されている。

この検索用転地ファイルはデータベースに化合物を新たに登録する時に更新する事が必要である。　小さなデータベースでは化合物登録時にまとめてフラグメント転地ファイルの更新も同時におこなうことが多いが、大規模なデータベースでは計算機の運営／管理上の問題から、多くはあるまとまった数の化合物を一度に登録するアプローチが取られる。

(2) 結合表を用いた部分構造検索（セットリダクション法）

部分構造検索を行う為の手法としてグラフ理論に基づいたセットリダクション（SET REDUCTION)法が１９６５年、E.H.SUSSENGUTH,Jr.[1]により最初に発表された。 その後、T-K.MINGおよびS.J.TAUBER[2]らにより改良が加えられ、より実務的に計算機上で行うレベルに至るものとして、１９７２年J.FIGUERAS[3]らが、ブール代数を用いたより具体的なアルゴリズムが発表されている。 ここではこのFIGUERASの論文に従って説明する。

1） E.H.Sussenguth,"A Graph-Matching Algorithm for Matching Chemical Structures", J.Chem.Doc.,6,36(1965).
2） T-K.Ming and S.J.Tauber,"Chemical Structure and Substructure Search by Set Reduction", J.Chem.Doc.,11,47(1971).
3） J.FIGUERAS,"Substructure Search by Set Reduction",J.Chem.Doc.,12,237(1972).

□セットリダクション法概要

セットリダクション法による検索は３段階の手続きから構成されている。

① プレスクリーニング
② セットリダクション
③ アトムバイアトム検索

部分構造検索はセットリダクション部分にて行われ、①と③の部分はこのセットリダクションを行う時の検索速度の向上（①）、および検索精度の向上／最終ヒット化合物のチェック（③）に用いられるものである。 ここでは、②のセットリダクションの説明に先立ち、①と③の機能についてのべる。

□ プレスクリーニング

①のプレスクリーニングは後に続くセットリダクションを効率的に行うためのもので、機能的には非常に簡単なものである。 すなわち、検索キーを構成する原子数と検索対象化合物の原子数とを比較し、検索対象化合物の原子数が検索キーの原子数よりも小さい時は検索せずにパスするというものである。 このチェックは原子の種類毎に行われる。従って出現頻度の小さな原子が検索キー中に多くある検索程、このプレスクリーニングの効果は大きくなる。 例えば、検索部分構造式中に３級アミンが２個あるならば、化合物中に３級アミンが２個に満たない化合物は続いて行われるセットリダクションをパスされて次の化合物の検索が実行される。

□ アトムバイアトム検索

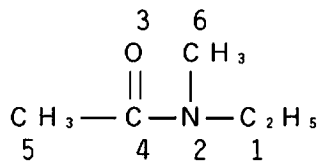
この検索の詳細は後にのべる。 ここではこのアトムバイアトム検索がセットリダクションという部分構造検索において果たす役割について述べる。

□　セットリダクション法

1．結合表の作成（セットリダクションを行うための準備）

セットリダクション法は最初に検索用の結合表を作成することから始まる。　この結合表の目的は、通常利用されているような構造式の数値データ化及び結合表からの構造式の再生といったことではない。　この結合表の目的とする所は、後に続く部分構造検索を行うのに必要な原子単位の情報をまとめて表示することにある。　従って、化合物の表現に利用される結合表とはかなり構造が異なっている。

図　にこの結合表をしめした。



原子ＩＤ	原子コード	コード数	結合相手原子・結合種
1	0206202	1	21
2	0107202	1	41 11
3	0108102	1	42
4	0106304	1	21 32 51
5	0106101	2	41
6	0106101	2	21

図　．セットリダクションを行うための結合表形式

最初の原子コードはＭＭＡＡＸＢＢの４種類の情報から構成されている。　ＭＭは原子の繰り返しをあらわし、０～９９の値を取る。　この繰り返し数は、例えば側鎖のメチレンの繰り返し（－ＣＨ２－）ｎのｎの値がはいる。

ＡＡの部分には該当する原子の周期率表に従った原子番号がはいる。　従って、炭素は６、窒素は７、酸素は８といった値をとる。　Ｘの部分には該当する原子が結合している原子（結合）の数が入る。　最後のＢＢには該当する原子に結合している総ての結合の多重度の合計がはいっている。　但し、水素との結合は無視する。

結合の多重度は以下のような値が用いられる。

- ■　０　ベンゼノイド結合
- ■　１　単結合
- ■　２　２重結合
- ■　４　互変異性結合
- ■　８　チャージおよび非局在化結合
- ■　１６　３重結合

これらの情報から、原子コードは個々の原子単位でその表現単位、原子種、その原子が結合している最隣接原子の数とその結合関係等の情報を簡潔にまとめたものといえる。２番目のコード数は個々の原子コードの存在数を示す。　従って、図　中末端メチルの５と６の原子は全く同じ原子コードを持つので、コード数は２となる。　３番目の項目は個々の原子の結合する相手の原子ＩＤと、その原子に結合している結合の種類をまとめて一つとし、これを結合原子の数だけリストするものである。　図中、１番目の原子団に結合する原子は２番目の酸素原子であり、これは単結合で結合している。従って、この部分の記号は２１が入ることになる。

原子コードとコード数はセットリダクションに入る前のプレスクリーニングに利用される。　つまり、検索用部分構造と被検索化合物の各原子について原子コードが調べられ、その原子コードの種類毎にコード数がチェックされる。　被検索化合物は原子コードの種類の数と個々の原子コードのコード数が検索キーの部分構造の数と同じかそれ以上存在することが求められる。　この条件を満たさない化合物は、セットリダクションにゆく前にふるい落とされる。

2．セットリダクションの実行

前節で述べた結合表を用いて検索キー構造式と被検索化合物との関係についてまとめた

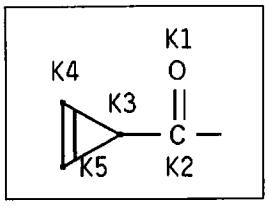
種々セット（１／０の数値データからなるベクトル）を作成し、これらのセットを利用しながら部分構造検索を行うのがセットリダクション法である。　ここでは実際の構造式を例に取りながらセットリダクションの手続きを順を追って説明する。

(1)　検索キー構造式及び被検索化合物構造式の結合表作成

セットリダクションを行うに先立ち、検索に用いる検索キーと検索対象のおのおのについて結合表を作成することが必要である。　例に用いる検索キー及び被検索化合物の構造式とその結合表とを図　に示す。

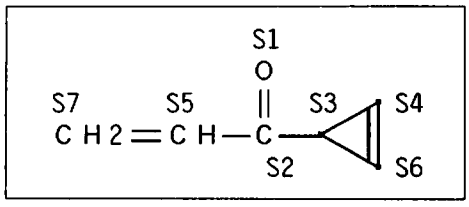
表１．検索キー構造式および被検索化合物構造式の結合表

検索キー構造式



原子ＩＤ	原子コード ＭＭＡＡＸＢＢ	コード数	結合相手原子・結合種
Ｋ１	０１０８１０２	１	２　２
Ｋ２	０１０６３０４	１	１　２　３　１
Ｋ３	０１０６３０３	１	２　１　４　１　５　１
Ｋ４	０１０６２０３	２	３　１　５　２
Ｋ５	０１０６２０３	２	３　１　４　２

被検索化合物構造式



原子ＩＤ	原子コード ＭＭＡＡＸＢＢ	コード数	結合相手原子・結合種
Ｓ１	０１０８１０２	１	２　２
Ｓ２	０１０６３０４	１	１　２　３　１　５　１
Ｓ３	０１０６３０３	１	２　１　４　１　６　１
Ｓ４	０１０６２０３	３	３　１　６　２
Ｓ５	０１０６２０３	３	２　１　７　２
Ｓ６	０１０６２０３	３	３　１　４　２
Ｓ７	０１０６１０２	１	５　２

(2)　検索用１次セット作成のための手続き

検索キー及び被検索化合物に関する前記結合表を用いて、セットリダクションの実行に必要な種々セットを段階的に作成する。　以下、この手順に従って説明する。

□　手続き１：　原子基準セットの作成

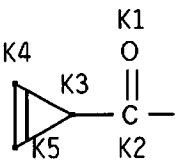
最初に検索キー構造式中の各原子について、被検索化合物中のどの原子が対応（原子の種類や結合特性が検索キー原子と被検索化合物原子とで全く同じ）しているかをチェックすることが必要である。　この原子対応に関する情報をまとめたものが原子基準セット[1]である。

＊１：この原子基準セットは原文ではCHARACTERISTIC VECTOR と呼ばれている。　ここでは説明上後にのべる結合基準セットとの対応から原子基準セットとした。

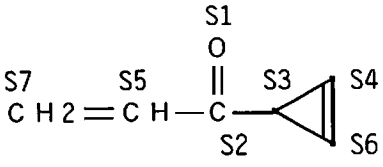
表　に検索キー原子に対する被検索化合物中の対応原子のリストが示されている。　表中の原子コードの内容は既にのべた。　原子基準セットのデータは、被検索化合物の各原子(この場合７原子)の順に従って並べられている。　このセットでは、検索キー構造式中の原子と同じ特性を持つ原子は１で、そうでない時は０の値で示されている。

表　.　検索キーと被検索化合物構造式とセットリダクション用セットコード例

検索用キー構造



被検索化合物



検索キー原子	検索キー原子コード	被検索化合物構造式原子基準セット
	M M A A X B B	S-1 2 3 4 5 6 7
K 1	0 1 0 8 1 0 2	1 0 0 0 0 0 0
K 2	0 1 0 6 3 0 4	0 1 0 0 0 0 0
K 3	0 1 0 6 3 0 3	0 0 1 0 0 0 0
K 4	0 1 0 6 2 0 3	0 0 0 1 1 1 0
K 5	0 1 0 6 2 0 3	0 0 0 1 1 1 0

例えば、図　の検索キー及び被検索化合物の結合表から検索キー構造式の１番目の原子（Ｋ１）と同じ特性（ケトンの酸素原子：原子コード＝０１０８１０２）を持つ原子は、被検索化合物中の１番目の原子[1)]（Ｓ１）であり、残る６原子は対応していないことがわかる。　そこで、対応する原子には１を、対応しない原子には０を割り当てることで（１００００００）の原子基準セットが作成される。　同様に、４番目のキー原子（Ｋ４：０１０６２０３）に対応する被検索化合物中の原子はこの場合Ｓ４、Ｓ５、Ｓ６の３原子が対応していることがわかる。

＊１：厳密にいうならば被検索化合物のＳ１原子はαβ不飽和ケトンの酸素原子であり、これはケトンの一種である。　化学的には検索キー構造式の意図するものと被検索化合物の内容とでことなる事が考えられるが、この場合の検索キー構造式の内容ではケトンの種類を特定することは不可能である。　また、原子コードだけの情報ではケトンの種類を特定することは出来ない。

□　手続き２：　結合基準セットの作成

手続き１で得られた原子基準セットは原子コードを基準とした対応チェックリストであり、あくまでも１原子についての対応である。　実際の部分構造検索では構造式を扱うのであるから、原子間の結合関係（隣接関係）のチェックが必要となる。

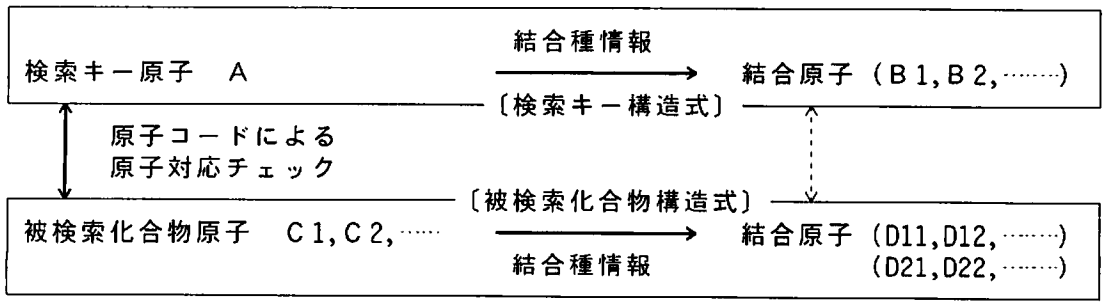


図　. 結合情報によるセット（結合基準セット）創出のための手続き流れ図

手続き２の目的は、検索キー構造式及び被検索化合物の両方の原子について隣接関係をチェックし、新たなセットを作成することである。　この作業を行う過程が図　に示されている。　この図に従って説明すると、新たなセット（結合基準セット）作成のための具体的な手続きは以下のようになる。

検索キー原子Ａに注目しているならば、そのＡ原子と結合している原子（隣接原子）のリスト（Ｂ１，Ｂ２，…）をその結合の種類の制限つきで求める。　同時に、検索キー原子Ａに対応する被検索化合物原子群Ｃ１，Ｃ２ ……の個々の原子についても、その隣接関係を検索キー原子Ａと同じ結合の種類の制限下にチェック（Ｄ11、Ｄ12、…）する。
この作業を検索キーの全原子について行う。

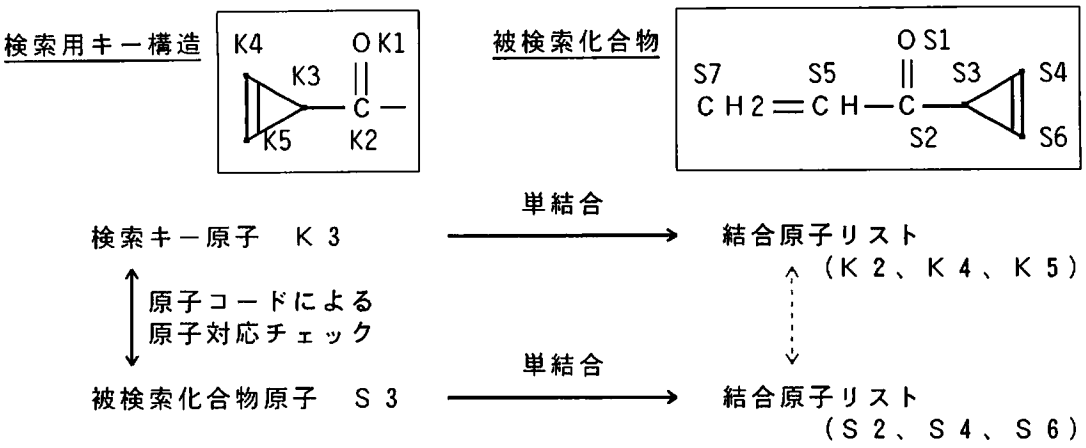


図　．Ｋ３原子に関する結合原子リストの創出例

この過程を実際の構造式を用いた図　に従って説明する。　今、検索キー構造式中の３番目の原子（Ｋ３）について考える。　この原子に対応する被検索化合物中の原子は３番の原子（Ｓ３）のみであることは先に求めた原子基準セットから簡単にわかる。　続いて、この注目している検索キー原子Ｋ３が結合している隣接原子に関する情報を集める。検索キーのＫ３原子の隣接原子はその原子コード情報（１０６３０３）より、３個の原子が単結合で結合していることがわかる。　この、Ｋ３原子に単結合でつながる隣接原子という情報を前提に検索キーの結合表を調べると、この条件を満たす原子はそれぞれＫ２、Ｋ４、Ｋ５の原子であることがわかる。　また、検索キー原子Ｋ３に対応する被検索化合物中のＳ３原子についても同様に、単結合で結合しているという制限下で隣接原子情報を取り出すと、Ｓ２、Ｓ４、Ｓ６がこの条件を満たす隣接原子であることがわかる。

以上の結果から新たにセットを作成する。　このセットは先に調べた検索キー構造式中の注目する原子（この場合Ｋ３）に隣接する原子群（Ｋ２、Ｋ４、Ｋ５）と、Ｋ３原子に対応する被検索化合物中の原子Ｓ３の隣接原子群（Ｓ２、Ｓ４、Ｓ６）を取り出したものである。　以上より、検索キー原子のＫ２、Ｋ４、Ｋ５の原子に対応する結合基準セットは被検索化合物の原子ＩＤ順に０１０１０１０となる。　このセットはＫ４及びＫ５の原子に関しても全く同じである。
Ｋ３の隣接原子各々に対応する結合基準セットを表　に示す。

表　．検索キー原子Ｋ３の隣接原子Ｋ２、Ｋ４、Ｋ５に対応する
被検索化合物原子のリスト

Ｋ３隣接 検索キー原子	結合基準セット S-1 2 3 4 5 6 7
Ｋ３－Ｋ２	０１０１０１０
Ｋ３－Ｋ４	０１０１０１０
Ｋ３－Ｋ５	０１０１０１０

□　手続き３：　結合基準セット及び特性ベクトルセットを用いた検索１次セットの作成
手続き１および２で求めた被検索化合物に関する原子基準セットと結合基準セットのブール代数積をとることで検索１次セットを作成する。

表　．Ｋ２、３、４原子に対する検索１次セットの作成

K３隣接 検索キー原子	原子基準セット S-1 2 3 4 5 6 7	結合基準セット S-1 2 3 4 5 6 7	ブール 代数積	検索１次セット S-1 2 3 4 5 6 7
K３－K２	０１０００００	０１０１０１０	⇨ ⇨ ⇨	０１０００００
K３－K４	０００１１１０	０１０１０１０	⇨ ⇨ ⇨	０００１０１０
K３－K５	０００１１１０	０１０１０１０	⇨ ⇨ ⇨	０００１０１０

さきにキー検索原子（この場合Ｋ２、Ｋ４、Ｋ５）のそれぞれと対応（同一の環境にある）する原子をチェックした原子基準セットを作成した。　このセットでは、検索キー原子と被検索化合物原子とで原子的な特性／環境が全く同じ時には１が、異なる時に０が立っている。　従って、検索キーのＫ２原子と特性／環境が同じ原子は被検索化合物中ではＳ２の原子が相当し、Ｋ４検索キー原子に対してはＳ４、Ｓ５、Ｓ６の原子群が対応している。

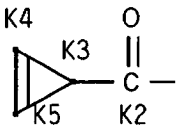
また、

表　には、このCONNECTIVITY-BASED SETとATOM-CODE-BASED SET のブール代数積を取った結果のＳＥＴが示されている。　このＳＥＴで１が立つのは、検索キー原子と被検索化合物原子とでその結合状況と原子状況とが全く同一であることを意味している。　従って、構造式検索ではこの１の立つ部分は検索構造式と被検索構造式とが一致している可能性を持つという必要条件を満たす部分であることになる。

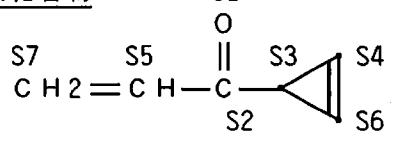
この検索１次セットの情報内容は、検索キー原子に対応する（原子及び結合環境が全く同じ）原子群を被検索化合物中から取り出すものである。

以上の手続きを検索キー原子の総てについて行った結果を図　に示す。

検索用キー構造



被検索化合物



QUERY ATOM	CHARACTERISTIC VECTOR
K１	１００００００
K２	０１０００００
K３	００１００００
K４	０００１０１０
K５	０００１０１０

図　．最終第１次特性ＳＥＴ（ベクトル）の構造

(2)　検索用第２次特性セットの構築

部分構造検索は図　で示された最終第１次ＳＥＴにおいて、検索キー原子と被検索化合物原子とが１対１で対応するならば、この検索キー構造式は被検索化合物に含まれていることになり、この時点で検索が完了して検索対象化合物は次の化合物へと移る。

しかし、図　における検索キー原子４及び５のように、対応する被検索化合物原子が２個以上ある時(MULTIPLE ISOMORPHORISM)は、最終的な１対１の対応を決定するための手続きが必要となる。　この多重＿＿＿(MULTIPLE ISOMORPHORISM)は分子の対象性に起因している。

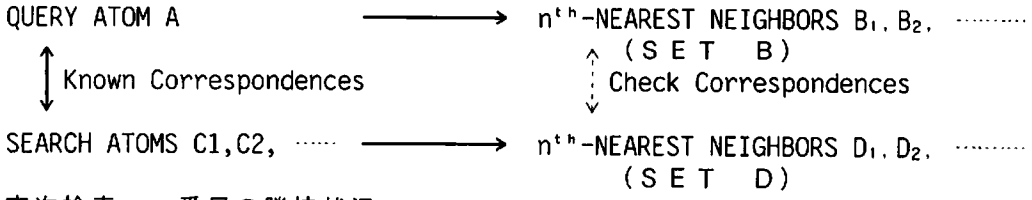


図　．高次検索：ｎ番目の隣接状況

この多重＿＿性をチェックし、真の包含関係を認識する目的で多重＿＿性原子の隣接原子の環境を調べる手続きを行う。　具体的には以下のような手続きを取る。

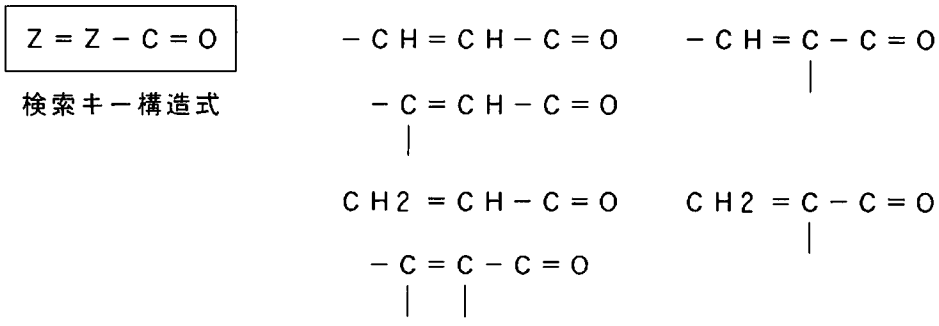
①検索キー構造式中の多重＿＿性原子（A）について、その最隣接原子情報（B1～B2 ……）を求める。〔セット　Bの作成〕
②先の段階でチェック済の検索キー原子（A）に対応する被検索構造式中の原子（C1～C2 ……）について、①と同様の手続きにより最隣接原子情報（D1～D2 ……）を求める。〔セット　Dの作成〕
③①と②で求めたセットBおよびセットDのブール代数積を求めて新たな特性セットEとする。
④③で得られたセットEが、その前のセットBおよびセットDと同じ時、この検索　はこの時点で終了する。　検索構造式は被検索化合物に一致したと見なされる。
　セットEがセットBおよびセットDと異なる時、さらに高次の隣接原子に移り、　①からのステップを繰り返す。

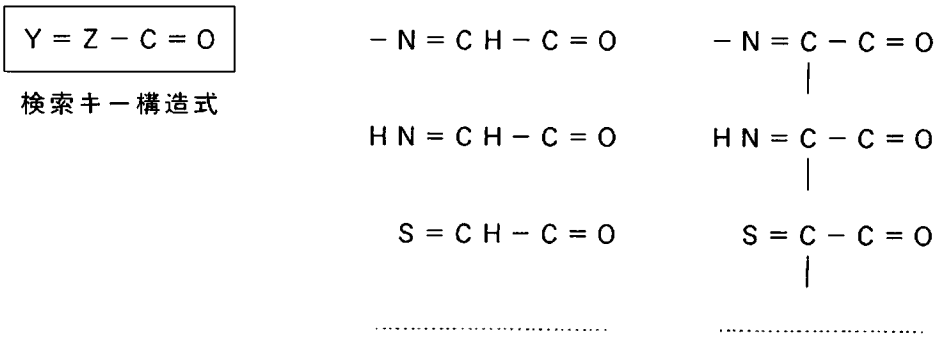
□　ミスマッチの検出
1．特性（CHARACTERISTIC VECTOR)ベクトルが0となる時。
2．検索キー原子のある特性を持つ原子（X）が、被検索化合物原子中で同じ特性を持つ原子（Y）よりも多数存在する時。　　X＞Y

□　特殊TERMINATORS

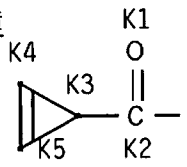
X：ヘテロ原子、共役
Y：ヘテロ原子、非共役
Z：炭素原子

ここで示されたX、Y、Zの表示はいわゆる特殊検索にも利用できるものである。　例えばこれらの機能を利用すればGENERIC検索は簡単に実行可能である。

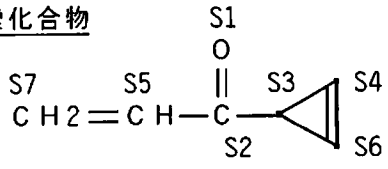




検索用キー構造



被検索化合物



検索用キーの多重＿＿の状態

検索キー原子	最近隣原子情報（検索キー構造式） K-1 2 3 4 5
K 4	0 0 1 * 1
K 5	0 0 1 1 *

手続き①：

(1)検索用キー原子４、５に対応する（全く同じ環境の原子）被検索化合物原子リスト

検索キー原子 隣接原子	原子基準セット S-1 2 3 4 5 6 7
K 4	0 0 0 1 0 1 0
K 5	0 0 0 1 0 1 0

(2)検索用キー原子４の１次隣接原子３、５に対応する被検索化合物原子の原子状況

検索キー原子 隣接原子	原子基準セット S-1 2 3 4 5 6 7
K 4 − 3	0 0 1 0 0 0 0
K 4 − 5	0 0 0 1 0 1 0

(3)検索用キー原子５の１次隣接原子３、４に対応する被検索化合物原子の原子状況

検索キー原子 隣接原子	原子基準セット S-1 2 3 4 5 6 7
K 5 − 3	0 0 1 0 0 0 0
K 5 − 4	0 0 0 1 0 1 0

手続き②：

(1)検索用キー原子４、５に対応する被検索化合物原子リスト

検索キー原子	原子基準セット S-1 2 3 4 5 6 7
K 4	0 0 0 1 0 1 0
K 5	0 0 0 1 0 1 0

(2)被検索化合物原子４、６に最隣接する原子リスト

被検索化合物原子	１次隣接原子情報 S-1 2 3 4 5 6 7
S 4	0 0 1 * 0 1 0
S 6	0 0 1 1 0 * 0

手続き③：

(1)セットＢとセットＤとのブール代数積を求める

検索キー	原子基準セット S-1 2 3 4 5 6 7
K 4	0 0 0 1 0 1 0
K 5	0 0 0 1 0 1 0

4 － 3	0 0 1 0 0 0 0
4 － 5	0 0 0 1 0 1 0
K 4	0 0 1 * 1
S 4	0 0 1 * 0 1 0
S 6	0 0 1 1 0 * 0
K 4 － 3	0 0 1 0 0 0 0
K 4 － 5	0 0 0 1 0 1 0
S 4	0 0 1 * 0 1 0
S 6	0 0 1 1 0 * 0