□ 化合物構造情報の計算機上での表現

① 化合物構造情報表記の一般的問題点

② 線型表記法

(1) 構造式の再現不可能な構造表記法

・ ＣＡＳ番号
・ ＭＯＲＧＡＮ名
・ ＳＥＭＡ名

(2) 構造式の再現可能な構造表記法

・ 一般的命名法
・ ＷＬＮ，ＳＫＯＬＮＩＫ，ＧＲＥＭＡＳ，ＳＭＩＬＥＳ

③ 結合表による化合物の表現

④ 3次元構造式の表現

・結合表による表現
・Ｚ－マトリクスによる表現

⑤ 化合物の立体化学関連表示

⑥ 関連計算機技術

・グラフィックからの直接入力
・ＣＨＥＭＩＣＡＬタイプライター
・ＷＬＮ→ＣＴ変換
・化合物名→ＣＴ変換

1．　化合物構造情報の計算機上での表現（概論）

1．1　計算機の情報体系と化合物の情報体系の差と融合

□計算機における情報の特徴

計算機が理解できるデータは0／1からなるバイナリーデータである。　人間がこのバイナリーデータをみて必要な情報を取り出すことは困難である。　計算機では、このバイナリーコードを基本とし、より高度なコード体系として数値データや文字データを扱うことが可能である。　この数値／文字データは人間でも理解可能である。　計算機と人間の両方が理解できる数値／文字データをもちいることで、基本的な計算機と人間とのコミュニケーションがなりたつ。　通常はこの数値および文字データを計算機の入／出力として用いる。　この数値／文字データの他に重要な情報として画像や音声データがある。　この画像／音声データは0／1情報の集合体として扱うことで計算機上での処理が可能となる。　しかし、このアプローチはメモリーを大量に必要とし、処理等にも時間がかかることから最近になり実用が可能となってきたもので、初期の計算機では実現されていなかった。

□化合物構造式における情報の特徴

化合物構造式を情報という観点から考えてみると、①2次元／3次元の形を持つ絵としての情報、②原子同志の結合関係の情報の二種類存在することがわかる。

化合物の構造式を絵として表現することは大事であり、化学者間のコミュニケーション手段として重要である。　しかし、化合物構造式は絵であっても、化合物としての本質的な情報は原子の結合関係（トポロジカル／トポグラフィカル）であり、化学上の様々な処理はこの結合関係を変化させることで行われる。　従って、画像情報は化学の分野では特殊な目的（コミュニケーション、イメージ把握等）の時にのみ利用され、化合物情報の実質的な受け渡しは原子の結合関係を中心として行われる。

□化合物構造式情報の計算機への取り込み

先に、計算機と人間の接点は数値／文字データと画像情報の扱いにあることをのべた。このうち画像情報の扱いは初期の計算機にとり実行困難であり、且つ画像情報は化学の分野では必ずしも本質的な情報伝達手段ではないこともべた。　従って、初期の計算機による化合物構造式の取り扱いは例外を除き、必然的に数値／文字データによる扱いから出発した。

化合物構造式　⇨　数字および文字データ　⇦　計算機

1．2　化合物構造式の数値／文字データへの変換

化合物構造式を数値／文字データへと変換する手法として様々な手法が存在する。　これらの手法は化学の分野で古くから利用されてきたものから、計算機の進歩に伴って最近展開されてきたものまで多種多様存在する。　現在利用されている化合物表記法の代表的なものを表1にまとめる。

表1．化合物の数値／文字データへの変換手法

	線形表記	マトリクス形式
2次元構造データ	CAS番号* 化合物名（IUPAC） WLN．GREMAS SMILES MORGAN* SEMA*	結合表（2次元）
3次元構造データ		結合表（3次元） Zマトリクス

＊　構造式の再現不可能

これらの表記法のうち、計算機と全く関係のない目的で開発されたものとして、CAS

番号と化合物命名法（ＩＵＰＡＣ名）がある。　ＣＡＳ番号は化合物に人為的に与えられる登録あるいはＩＤ番号であり、このＩＤ番号を発行する作業は米国のケミカルアブストラクツサービス（ＣＡＳ）が行っている。　ＣＡＳでは世界中の文献やパテントに発表された化合物をチェックし、新奇化合物についてこの番号を発行している。　従って、化合物構造式情報とＣＡＳ番号との間には何の相関もなく、単なる化合物のＩＤである。　この番号は現在一千万を超えている。　ＩＵＰＡＣ命名法は化学者が必要とする共通言語としての化合物命名法であり、世界の化学者の合意のもとに定められたものである。　残りの表記法は計算機での利用を前提として開発された。

□計算機の進歩と変換様式の変化

計算機による化合物構造式の変換様式は、計算機のハード上での進歩と密接に関係している。　計算機の当初はメモリーの容量（コア及び外部メモリー）が少なく、計算速度も遅いため、大量の数値／文字データを計算機上で扱うことは極めて困難であった。　この結果、当初の計算機による化合物構造式の扱いは、なるべく少ない数値／文字情報で扱うことを最優先としたアプローチに限定された。　この要求に答えるものとして、線型表記法が展開された。

しかし、化合物構造式そのものを文字データだけで表すこと自体が困難である事に加え、さらに可能なかぎり少ない数字／文字であらわすことは、抽象化が進み、ルールが複雑となり化学者にとって扱いにくくなることを意味している。　計算機からの一方的な要求を満たすために考案された線型表記法は、最後まで化学者とのギャップを埋めることは出来なかった。　現在では一部のアプローチで利用されているのみである。

計算機のハード技術の進歩（特にグラフィック関連技術）にともない、化合物構造式そのものをグラフィックディスプレイ上で表現することが現実のものとなった。　計算機上で化合物構造式を絵として扱うには従来の構造式変換技術とは異なる扱いが必要である。この目的で結合表に代表されるいくつかの変換技術が新たに開発された。　現在ではこの結合表形式が主流となっており、結合表自体も国際的に統一されつつある。

□計算機の進歩と利用形態の変化

変換様式のみならず、利用形態そのものも計算機のハード技術の進歩と密接に関係して変化してきた。　当初の利用は化合物データベースとしての利用が主であった。　分子軌道法での利用等も行われていたが、計算速度も遅く、メモリーの制限等から経験的及び半経験的分子軌道法の適用が限界で、実用に供するレベルとは程遠かった。　しかも、化合物は数字／文字で表現されるだけで、化合物構造式そのものを表示することが出来ないために化学者が計算機を利用することは困難であった。

グラフィックディスプレイの進歩とともに、化合物構造式をディスプレイ上に表示することが可能となった。　この結果、一般の化学者でも構造式をコミュニケーション手段とし、計算機と直接対話することが可能となってきた。　グラフィックのみならず、計算速度の高速化やメモリーの増大といったハードの進歩は化学分野における計算機利用を急速に拡大した。

構造式を直接扱うことでプログラム自体も構造式を主体とするマンマシンインタフェースを中心とした設計がなされるようになり、化学者と計算機とのギャップは従来と比べてかなり小さくなってきた。

１．３　化合物構造式変換法の分類

化合物構造式の数値／文字データへの変換は、その変換対象化合物が２次元構造式か３次元構造式かで大きく２分類可能である。　ここで２次元構造式は紙上に描かれる化合物構造式であり、３次元構造式というのは分子モデルのように３次元座標情報を持つものを意味する。

さらに、変換された数値データあるいは文字データを単に線条に並べた一次元の表現（線形表記）で行うのか、あるいは２次元のマトリクス形式で表現するかでさらに２種類に分類される。　マトリクス形式で表現する変換式では構造式を再現出来るが、線型表記法では構造式を再現できるものと出来ないものとの二種類存在する。　これらの分類は既に表１に示されている。　ここに示された手法以外にも化合物の変換形式は様々なものが存在し、ケースバイケースまたはローカルに利用されている。

□個々の化合物表記法の特徴

ＣＡＳ以外の表記法は、基本となる化合物構造式に対し個々の手法により定められた変換アルゴリズムを適用することで数字／文字データへと変換される。　これらの変換法のうち化合物名（ＩＵＰＡＣ命名法）については先にのべた。　ＷＬＮ、ＧＲＥＭＡＳ、ＳＭＩＬＥＳ等の線型表記および結合表形式の総ての表記法は、変換データから化合物構造

式を再現することが可能である。　これら以外のＭＯＲＧＡＮやＳＥＭＡといった表記法は化合物の一元／一項対応による化合物検索を主目的とし、変換コードからの化合物構造式の再現は不可能である。

□化合物の表記における一元一項対応の問題について

化合物構造式を何らかの形で表現する時に問題となることとして、化合物構造式と化合物名との一元一項対応という問題が存在する。　ここでは簡単にこの一元一項対応についてのべる。

この一元一項対応が特に問題となるのは化合物の命名法を定める時である。　化合物名というものは厳密に１化合物に１つの名前というように１対１で対応していることが理想的な形態である。　しかし、３次元的に複雑な形態を持つものを１次元の文字集合で完全に表すのは困難である。　さらに、化合物命名法が定まる前から使われていた慣用名等が存在し、このような事実が化合物と化合物名の一元一項対応を困難にしている。

一元一項対応の元と言う言葉はこのばあい化合物に対応している。　また、項という言葉は化合物名に対応している。　従って、一元一項対応とは化合物とその化合物名とが１対１で厳密に対応していることを意味する。

多元一項対応とは一つの化合物名が複数の化合物に対応する時を、一元多項対応とは一つの化合物が複数の名前をもっていることになる。　このように、一つの名前に複数の化合物が対応したり、一つの名前に複数の化合物が対応することは、化合物検索を行う時に化合物名をもちいて検索することは能率が悪いことを意味している。

一元一項対応 ・ＣＡＳ番号	化合物（１）	化合物名（１）
多元一項対応 ・	化合物（Ｎ）	化合物名（１）
一元多項対応 ・	化合物（１）	化合物名（Ｎ）
多元多項対応 ・	化合物（Ｎ）	化合物名（Ｎ）

１．４　線型表記法（Linear Line Notation）による化合物構造式の表現

線型表記法は化合物構造式を一次元の数字／文字列として表すものであり、化合物命名法の基本である。　本来、二次元及び三次元である化合物を一次元で表現するため、情報を完全に保ちながら一次元に落とすには様々な変換ルールが必要となる。　使用にあたっては、その変換ルールを充分に理解することが必要であり、化学者にとっては大きな負担となる。　特に化合物の複雑な３次元情報を完全な形で１次元情報へと変換することには種々の点で限界が存在し、その限界をカバーする為にルールがさらに複雑になるといった悪循環に陥ることになる。

化合物名（ＩＵＰＡＣ命名法）は化学者の常識、コミュニケーション手段として必須であり、この変換ルールを知っていることは化学者である必須条件である。　従って、このＩＵＰＡＣ名を用いて計算機を利用することができれば化学者の負担は少ないのだが、不幸にしてこのＩＵＰＡＣ化合物名をそのまま計算機上で扱うことは困難である。　この主な原因として以下のようなことがある。

①キーボードには無い特殊な記号を用いている
②化合物名の順序づけに自由度があり過ぎる
③構造が複雑になると表現が長くなる

①の問題は化合物構造式を計算機に入力出来ないという決定的な問題となる。　②の問題は化合物検索時に、該当する化合物の検索が出来ない、検索効率が悪いといった問題が発生する。　これはＩＵＰＡＣ命名法が化合物の命名を目的とするものであり、効率的な検索を目的とするものではないことによる。　③ではメモリー等の制限に関する問題が発生する。

これらの欠点をカバーし、特に計算機での取扱を目的として開発されたのがＷＬＮやＧＲＥＭＡＳ等で代表される線型表記法であり、後にのべる結合表とよばれる表記法である。

これらの表記法を用いることで、先にのべた様々な欠点が解消される。

以下の表には化合物名とＷＬＮ、ＧＲＥＭＡＳ等の線型表記法との特徴的な差がまとめてある。

特徴	化合物名（ＩＵＰＡＣ）	線型表記法（WLN,GREMASその他）
Unique／Unambigous	同一化合物名の表現に種々の名前が存在し、化合物と名前が１対１に対応しない。	化合物と変換コードは良く対応する。厳密な一元／一項対応を目的とした表記法もある。
使用文字	特殊記号を用いる。『添字、大文字／小文字の混在』	キーボード上にある文字のみを使用
表現文字数	一般的に、化合物の表現に必要な文字数は多くなる。（慣用語は除く）	化合物表現の為の文字は少なくなるように設計されている。

例）　ＢＥＮＺＥＮＥ　⟹　Ｒ
ＥＴＨＡＮＥ　⟹　２
1-METHYL-2-METHOXY-4-AMINOBENZENE　⟹

上記線型表記法の開発により、化合物構造式の計算機への入力が可能となった。　つまり、２次元および３次元の複雑なトポロジカル情報が１次元情報へと変換され、この結果構造式が単なる文字列へと変換されたためである。　この結果、化合物検索は単なる文字検索の問題へとすりかえられ、計算機による従来からの検索技術を利用する事で化合物検索が可能となった。

しかし、前記ＷＬＮやＧＲＥＭＡＳ等を用いるためには特殊な変換ルールをマスターすることが必要である。　しかし、主なユーザである化学者はこれらの変換ルールを知らない事が多い。　従って、化合物検索を行おうとするならば化学者は普段利用している化合物命名法（ＩＵＰＡＣ命名法）の他にＷＬＮ、ＧＲＥＭＡＳ等の線型表記法を改めて学ぶ事が必要となる。　この事実は多くの化学者にとって大きな負担となる。

化学者の思考体系は化合物構造式を起源とし、総てのコミュニケーションも化合物構造式が基本である。　従って、ＷＬＮ等の数字／文字データを基本として思考することはありえず、必ず構造式に変換する事が必要である。　従って、計算機を離れた途端ＷＬＮはその存在意義を失い、化合物構造式への変換が必要となる。　このように、日常的に構造式→ＷＬＮ→構造式といった変換業務が発生し、特に複雑な化合物の時は、この業務が大きなネックとなってくる。

１．５　個々の表記法による化合物構造式の変換

□ＷＬＮ(WISWESSER LINE-FORMULA CHEMICAL NOTATION)の概要

ＷＬＮ表記法は計算機による使用を目的とし、W.J.WISWESSER により提唱された化合物の線型表記法である。　この表記法は計算機での使用を前提とするため、使用する文字は数字、２６個の英大文字、３個の区切り記号＆、－、／およびブランクだけである。

これらの数字／文字／記号は一字単位で特有の情報を表現することを原則とし、例外（出現頻度の低い原子で通常二字で表現される（例、Ｆｅ→ＦＥ、Ｎａ→ＮＡ等）ものやＵＵで示される３重結合等）を除き二文字以上で一つの単位となることはない。　表２にＷＬＮで利用される代表的な文字／記号とその情報内容をまとめて示す。

表２．ＷＬＮで利用される英文字／特殊記号とその情報内容

Ｑ　－ＯＨ	Ｅ　Ｂｒ	Ｎ　３級窒素
Ｖ　ケトン	Ｆ　Ｆ	Ｋ　４級窒素
Ｗ　－ＮＯ２、－ＳＯ２－	Ｇ　Ｃｌ	Ｘ　４級炭素
Ｍ　ｉｍｉｎｏ、ｉｍｉｄｏ	Ｉ　Ｉ	Ｙ　３級炭素（メチレン可）
Ｚ　ａｍｉｎｏ、ａｍｉｄｏ	Ｊ　ハロゲンの総称的表現、環情報の区切り	
Ｈ　水素（通常は省略される）	Ｕ　２重結合	ＵＵ　３重結合
Ｒ　ベンゼン環	Ｌ　炭化水素環	Ｔ　ヘテロ環及び飽和環
＆　切断を示す	－　繰り返し	／　区切り

化合物構造式はこれらの文字／記号を組み合わせることで線型な文字列へと変換される。

文字列は右、左のどちらからでも書くことができるが、読む時は左から右へと読むことが決められている。　このために文字列創出時には左右どちら向きに並べるかが問題となる。　この問題をクリアするべく、ＷＬＮで利用される数字／文字／記号はそれぞれ順位付けがなされている。　順位は低い方から順番に特殊記号、数字及び文字のようにつけられている。　この順位は最終的な文字列の向きを決定する基準となる。　すなわち、化合物を表現する文字列が決まったならば、その最初と最後の部分を比較し、順位の高いもの（Ｚの方）が文字列の最初にくるようにしたものが最終文字列となる。

〔文字ランキング〕

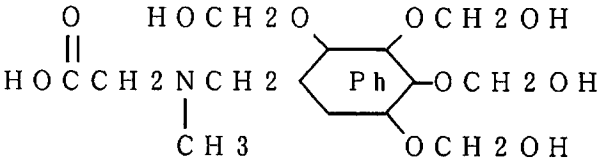
＆－／０１２３４５６７８９１０１１１２（・・・）ＡＢＣＤ（・・・）ＷＸＹＺ

高 ⟶ 低

このようにＷＬＮを構成する文字等はシンプルで明快なものであるが、これらを運用する変換ルールは複雑であり、ルール数も多くなっている。　例えば、線形化合物の表記に１９ルール、単環化合物の表記に１０ルール、縮合環に８ルール、その他より複雑な化合物の表現に６１ルール必要である。　個々のルールがさらに複数のサブルールに別れているものも多く、実際のルール数は数百のレベルに達し、このルールをマスターすることは多くの化学者にとり負担となるものである。

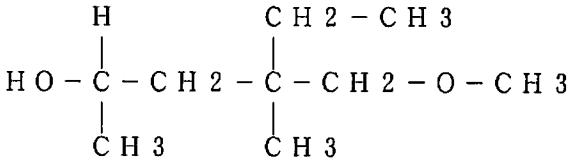
計算機の進歩に従い、利用者に負担を強いることの多いＷＬＮは殆ど利用されなくなった。　ここではＷＬＮのルールを解説することはしない。　そのかわりに幾つかの化合物構造式とそのＷＬＮ表記を示す。　これらの例から、ＷＬＮ表記の優秀さと、変換ルールの複雑さとを感じていただければよい。

構造式１



ＷＬＮ表記　　ＱＶ１Ｎ１＆１Ｒ　Ｂ－　Ｃ－　Ｄ－　Ｅ－／Ｏ１Ｑ　４

構造式２



ＷＬＮ表記　　ＱＹ＆１Ｘ＆２＆１Ｏ１

□ＳＭＩＬＥＳ（Simplified Molecular Input Line System）の概要

ＳＭＩＬＥＳは化合物の線型表記法として開発されたものであるが、他の多くの線型表記法と異なり、①表記ルールの単純さ、②同じ一化合物を表現するのに必要な字数の少なさの２点を最大の特徴とする。

構造式のＳＭＩＬＥＳへの変換は化学者が構造式を紙上にペンで書くのと同じような感覚で行われる。　すなわち、構造式をペンで書く順に従って原子と結合を順に文字コードで置換し、分岐等のためにペンを紙上から一旦離すところはマーカーをつけて括弧でくくる。　この繰り返しにより、単純な化合物から複雑な化合物まで簡単に文字情報へと変換される。　表記文字数の少なさは、互いの結合関係（トポロジカル）が暗黙のうちに表記中に取り込まれていることによる。　即ち、結合関係の情報を特に明記する必要がないことが表記文字数削減の大きな要因となっている。

□　ＳＭＩＬＥ表記のためのルール

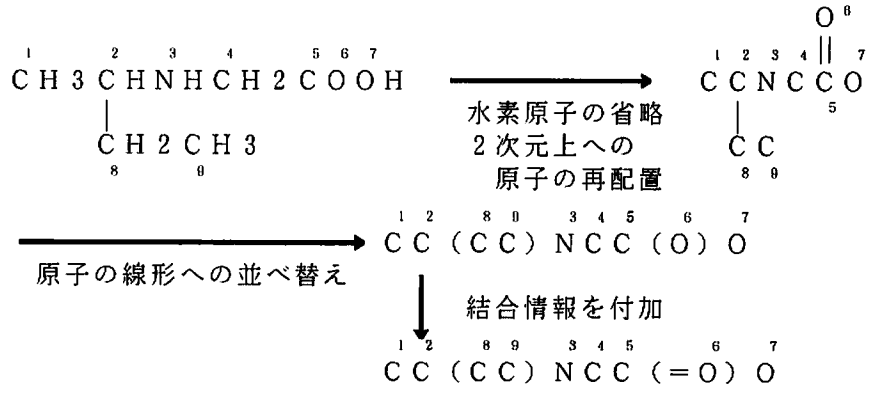
ＳＭＩＬＥＳの表記では原子と結合（単結合は省略）を一つの文字にあてはめ、それらを化合物の適当な場所から順に絵を書くように文字と置換することで線型表記へと変換される。　この変換時に必要なルールの主要なものとして以下に示すものがある。

・水素原子は特別の場合を除き省略する。

・原子は大文字で表現する。
・芳香族原子は小文字を用いる。
・原子の直後にある数字は同じ数字の所で結合していることを意味している。
　一つの原子が数ケ所と結合している時、原子に続く数値は複数となる。
・結合は単結合は省略する。　２重結合は＝で示し、３重結合は＃で示す。
・同じ原子に何個も結合している時（分岐がある時）は括弧でくくる。
　括弧は何重になってもかまわない。

図　に線形化合物、図　に環状化合物構造式のＳＭＩＬＥＳ表記への変換仮定の手順を示す。

１．線形化合物



２．環状化合物

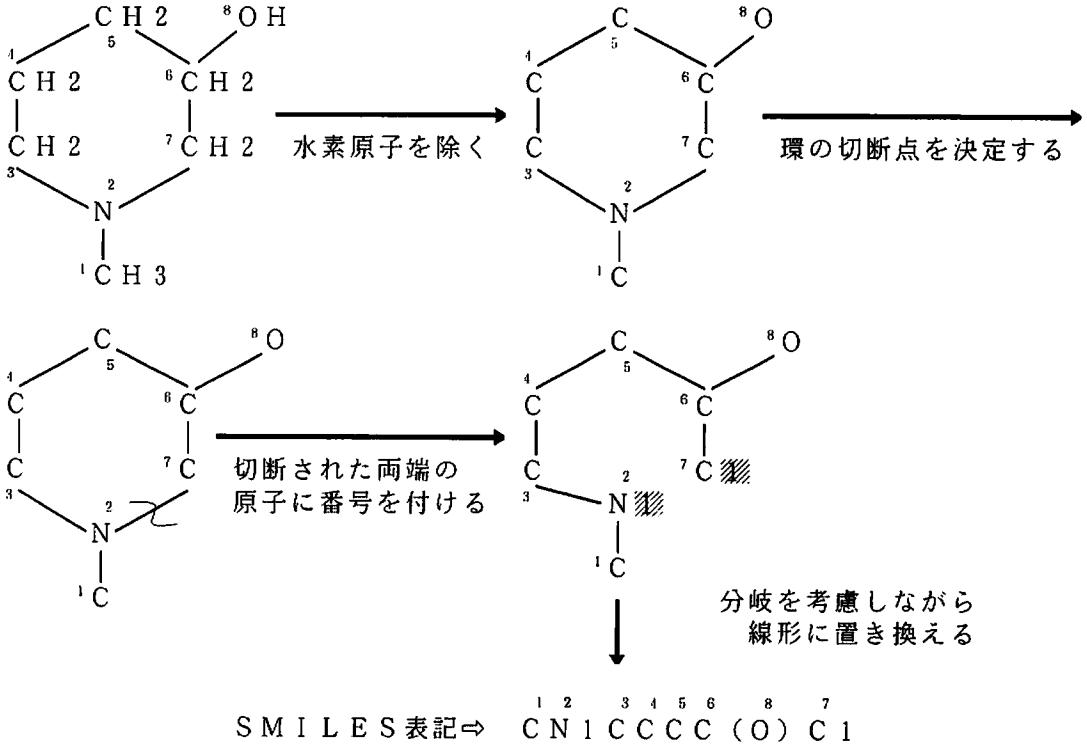


図　．線形及び環状化合物のＳＭＩＬＥＳへの変換手順
　＊各原子の添字は原子の識別を目的としており、ＳＭＩＬＥＳとは関係ない

□　ＳＭＩＬＥＳ表記された名前のユニーク化
　化合物構造式をＳＭＩＬＥＳ表記にすることで線型表記が可能となるが、最初に作成されるＳＭＩＬＥＳ名は決してユニークなものではない。　従って、ＳＭＩＬＥＳ名を使っ

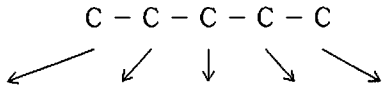
て化合物検索等を行うならば、このＳＭＩＬＥＳ名を検索に都合の良いユニーク性をもたせることが必要となる。
ここではＳＭＩＬＥＳをユニークＳＭＩＬＥＳとするための手続きを簡単にのべる。

〔手順１〕
最初に個々の原子について以下の条件を見出す。
１）　結合数
２）　非水素原子の結合数
３）　原子番号
４）　チャージの符号（０，－，＋）
５）　チャージの絶対数（０－１０）
６）　結合水素数（０－９）

例）末端メチル原子：１１６００３
　　メチレン　　：２２６００２

〔手順２〕

各原子にアサインされた前記数値データを比べて、各原子の順位付けを行う。



116003-226002-226002-226002-116003

順位	１	２	２	２	１

〔手順３〕

先に付けられた順位を基準とし、その隣接順位の数を加えて新たな番号を付ける。この番号を比較し、新たに順位をつけなおす。

順位	１	２	２	２	１
	+2	1+2	2+2	2+1	+2
番号のつけなおし	２	３	４	３	２
新順位	１	２	３	２	１

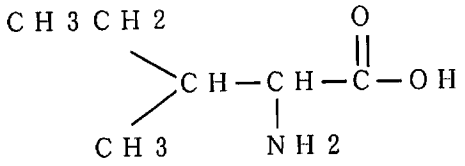
〔手順４〕

手順３を繰り返し、順位の変化がなくなる、または順位の数が減少した時はその一つ前の順位で停止する。

新順位	１	２	３	２	１
	+2	1+3	2+2	1+3	+2
	２	４	４	４	２
新順位	１	２	２	２	１

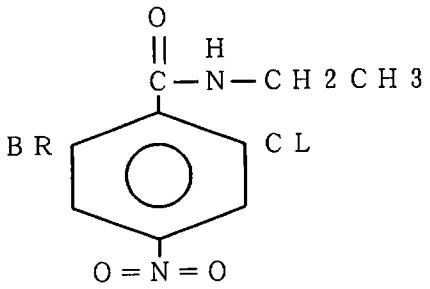
ここで付けれた順位は１と２だけで２種類しか存在しない。　従って、３種類の順位が存在していた一つ手前の順位付けを最終順位として採用する。

最終順位	１	２	３	２	１
	Ｃ	－	Ｃ	－	Ｃ	－	Ｃ	－	Ｃ



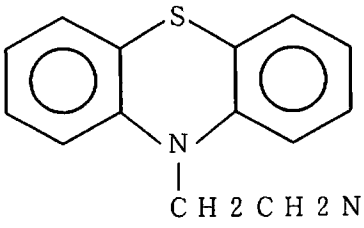
QVYZY2&1

CCC(C)C(N)C(=0)0



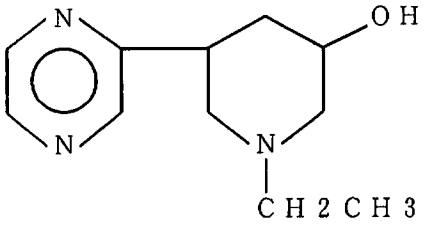
WNR CG EE DVM2

CCNC(=O)c1c(CL)cc(N(=O)=O)cc1(BR)



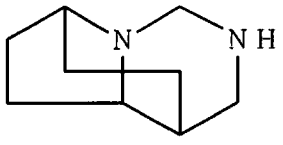
T C666 BN ISJ B2N1&1

C(C)NCCn1c2ccccc2sc3ccccc13



T6N DNJ B- CT6NTJ A2 EQ

CCN1CC(c2cnccn2)CC(O)C1



T566/BH 2AB K AN JMTJ

C1CC2CCC3C1CNCN23

図 . WLN及びSMILESによる化合物構造式の線型表記