解析に用いる数値データには大きく2種類のデータが存在する。

① 連続変数

② 不連続（カテゴリカル）変数、ダミー変数

とである。

3．5．数値データ間のスケールや精度について

□ 数値データのスケール／精度の統一

複数の数値データを同時に用いる時常に問題となるものとして、数値データのスケールと精度の問題がある。 解析は常に信頼度を高めた状態にしておく事が必要である。 この信頼度を低める要因としてさまざまなものがある。 この一つとして、数値データのスケールと精度の統一の問題がある。

解析に用いる数値データはそのままの値を用いて解析を実行する事が理想である。 これは解析結果を検討する時、生データを解析に用いればそのままの値で記述子間の貢献度等の比較／検討が可能であるからである。 スケールや精度が統一されないデータを用いる時は結果の解析が困難である。 特に重回帰手法ではこの係数の大きさが重要な意味を持つことが多く、従って重回帰による解析ではスケールの統一されたデータを用いる事が望ましい。

また、数値データ間の精度に差がある時、解析結果の精度は最も精度の粗いデータに支配される為、一部のデータにどんなに精度の高いデータを用いても無意味となる。

このような記述子間のスケールや精度の差による数値データ間の不当な扱いによる解析ミスを防ぐ目的で数値データに関し、場合によっては解析前になんらかの手続きを行う必要がある。

□ データ間のスケールの差

データ間のスケールの差は、様々な種類のデータを同時に利用する時、常に発生する重大な問題である。 例えば、化合物内の窒素原子の数と分子量とを同時に使う時には一般に窒素原子の数は1桁以内であり、しかも整数で表現される。 一方、分子量は通常3桁以上の値であり、しかも少数点以下の値を含む実数で表現されている。 このようにスケールの全く異なる2種類の数値データを混同して利用した時、解析結果の信頼性は下がることをさけることはできない。

このような問題点を解決する手法としてすべてのデータを解析に先立ち、平均が0、標準偏差1に統一して変換しておくことが行われる（オートスケーリング）。

□ データ間の精度の差

データ間の精度の差を統一し、最も精度の高いデータに基準を合わせるような数学的処理方法は存在しない。 精度に関してはあらかじめ解析を行う前に、用いるデータの精度を出来うる限り高めておく事しか対応のしようがない。

3．6． オートスケーリング

オートスケーリングはその名の通り、用いる数値データを平均＝0、標準偏差＝1の値に統一する事を意味する。 この結果すべてのデータのスケールが同じ基準に統一されるため、解析結果の信頼度が向上する。

$$Q = \sum_{i=1}^{n} (W_i - \overline{W}) \qquad (1)$$

$$W_k' = \frac{W_k - \overline{W}}{Q} \qquad (2)$$

Qは用いるデータの変量を示す。 W_k' はオートスケーリング後のWの値の内k番目の W_k の値、$\overline{W}$は用いる記述子の平均値である。

＊オートスケーリングを行う事により解析結果の信頼性は得られるが、この処理はすべてのケースについて万能ではない。 特にオートスケーリングにより失う物がある事を忘れてはならない。

□　判別式に関して行う記述子解析における問題

解析結果として得られた判別式の記述子について情報解析する時、オートスケーリングをかけていない時にはそのままの値を用いてその大小を比較する事が可能である。　しかし、オートスケーリングをかけた時はウエイトベクトルの要素値だけで単純に比較することはできない。

＊個々の数値データについて、ウエイトベクトルの絶対値の大小で比較する事はできないが、ウエイトベクトルのサイン（符号）は評価の対象となる。　つまり、サインが正の時と負の時とでその数値データの寄与の意味（方向）が反対になるということである。

例）　２クラス分類で、以下のような判別式が得られたとする。

判別式　　$D = 5X_1 + 2X_2 - 3X_3 - 2X_4 + 3X_{d+1}$　　――――（１）

式１中、式X_1, X_2ととX_3, X_4の符号が反対である。　正の値を係数として持つX_1, X_2がクラス１に貢献するとすれば、X_3, X_4はクラス２に貢献する事になる。　しかしこの場合、係数の大小で単純に貢献度の順位をきめる事は出来ない。　つまりオートスケーリングをかけたデータを用いた場合、クラス１に対する貢献度はX_1の方がX_2よりも大きいという結論を導き出すことは出来ない。

3．7．パターン認識に利用される数値データ（化学関連分野）の詳細

パターン認識に利用される数値データとして様々なものが利用されている。　これらの数値データのうち、特に化学の分野において利用されるデータは何らかの意味で化合物の基本的な特性に相関を持っている事が必要である。

これらの数値データの大部分は化合物構造式を基本として展開されるものが多い。　この数値データを大きく分類すると、トポロジカル、トポグラフィカル、物理化学的データ、その他と分類できる。　以下にこれらの数値データの特徴とよく用いられる典型的な数値データについて説明する。

表３－　パターン認識／多変量解析等に用いられる数値データ

■　トポロジカルデータ
分子構造インデックス：原子数（原子種）、結合数（結合種）、リング数、その他
様々なインデックス値：ＨＯＳＯＹＡインデックス、分子結合インデックスＭＣ値
パス値インデックス、

■　トポグラフィカルデータ
化合物の３次元的形状に関するパラメータ
化合物全体構造　：ボックスパラメータ、対称パラメータ、
立体格子パラメータ、その他
化合物部分構造　：ステリモルパラメータ、

■　物理化学データ
分子に関する様々な物性データ　：　分子屈折率、分子量、ＬＯＧＰ、融点、沸点
分子容積、分子表面積、その他
分子軌道法より得られる様々なパラメータ：電子密度、ＨＯＭＯ、ＬＵＭＯ、他
分子力学計算から得られるパラメータ：　種々歪みエネルギー
種々スペクトルより得られるデータ：　種々スペクトルデータ

■　その他のデータ
部分構造パラメータ：　部分構造の有無、部分構造数、
部分構造単位の様々なパラメータ値計算、
演算パラメータ１　：　記述子間の演算により得られるパラメータ（＋－ｘ÷ Log）
演算パラメータ２　：　他の解析手法より算出されたパラメータ
ダミーパラメータ　：　有るパターン存在の有無（１／０）に関するパラメータ

3．8．　トポロジカルパラメータ

□　トポロジカルパラメータの基本的特徴

トポロジカルパラメータは対象とする構造体（ここでは化合物構造式をさすものとする）内部の点と点との結合関係に関する情報を数値化する。　化学の分野ではこの点を原子に、点と点とを結ぶ線を結合と見立てて化合物構造式を数値データへと変換する。　ここではトポロジカルデータの定義を拡張し、点（原子）や線（結合）そのものに関するデータもトポロジカルパラメータと称することにする。

定義の拡張により以下に示すような情報もトポロジカル情報として扱われる。

①　化合物中に含まれる原子の数。　これは原子の種類毎の値や、複数種類の原子種をまとめた値等がある。

例：Ｃの数、Ｎの数、Ｓの数、等
：ハロゲン原子の数（Ｆ＋ＣＬ＋ＢＲ＋Ｉ）等

②　化合物中に含まれる結合の数。　これも①同様に結合の種類によりことなり、また複数種類をまとめる事もある。

例：単結合の数、２重結合の数、３重結合の数、芳香族結合の数等
・不飽和結合の数（２／３重結合＋芳香族結合）等

③　その他幾つかのパラメータがこの分類に入れられる事もある。
・部分構造パラメータ、リングパラメータ、官能基パラメータ等

□　トポロジカルデータの特徴

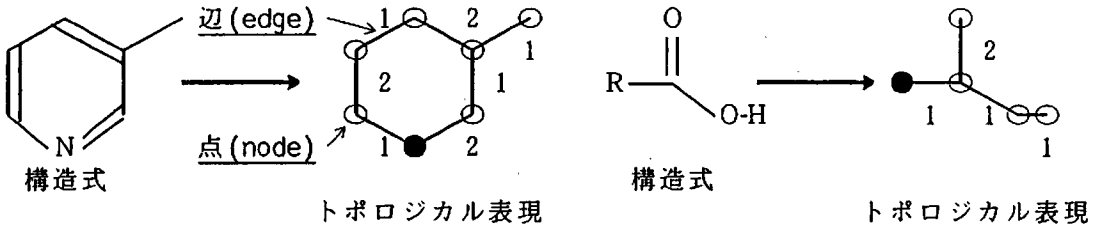
トポロジカルデータは化合物を構成する原子と結合とを、それぞれノードとエッジとに

定義する。　トポロジー的な問題として化合物構造式を捕らえ、化合物の原子（ノード）間の相互的な関係（つながり状態）を数値データに変換したものである。

このトポロジカルデータの特徴を簡単にまとめると以下のようになる。

長所：　化合物の複雑な結合情報を数値データに変換できるので、通常の数値データでは説明出来ないような複雑な情報を扱う事が可能となる。　この結果、分類能が飛躍的に向上することが期待される。

短所：　数値データの変換ルールと化合物構造式との関係が不明な時が多い。　アルゴリズムが数値データへの変換の為のルールとなっている事が多く、最終目的である目的変数に対する情報の説明や解釈が困難な事が多い。　即ち、分類の為のデータに陥り易く、分類だけが目的の時は強力なパラメータとなりうるが、そのパラメータの持つ意味（情報）を解釈する事が重要となる解析には不向きである。



このトポロジカルデータは現在様々なものが提唱されている。　特に有名なものとして化合物の物性予測に用いられる事の多いＨＯＳＯＹＡ　ＩＮＤＥＸ　と、構造活性相関分野で利用実績の多い分子結合インデックス（Ｍ．Ｃ．）（Molecular Connectivity Index）等が有名である。

① ＨＯＳＯＹＡ　ＩＮＤＥＸ

② 分子結合インデックス（ＭＣ：Molecular Connectivity Index）

このパラメータはＭＣＩと呼ばれており、ＫＩＥＲとＨＡＬＬにより提唱されたパラメータである。　このパラメータは化合物の様々な物性との相関が良く、数多くの種類の物性との相関研究の実績がある。　構造－活性相関の分野でも多用されている重要なパラメータである。

□ ＭＣＩ値の算出法

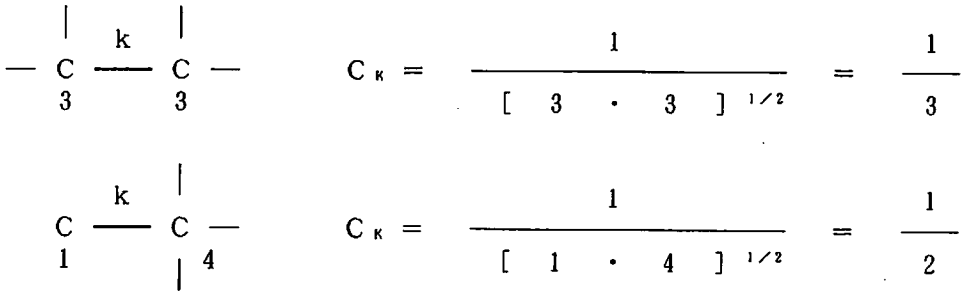
まず化合物を構成している個々の結合についてＣ$_k$値を求める。　続いて、このＣ$_k$値を化合物中の総ての結合について総和した値が分子に対するＭＣＩ値となる。

$$MCI = \sum_{k=1}^{m} C_k = \sum_{k=1}^{m} \frac{1}{[\ L_i \cdot L_j\ ]_k^{1/2}}$$

k：　ある一つの結合のＩＤ番号
i：　結合ｋを形成する原子２個のうちの一つの原子に関するＩＤ
j：　結合ｋを形成する原子２個のうちｉ以外の原子に関するＩＤ

上式中、L_iは原子ｉの結合の多重度であり、L_jは原子ｊの多重度を示している。この多重度とは現在注目している原子から飛び出している結合の数を意味し、この時水素原子とつながっている結合の数は無視して計算する。

例）　Ｃ$_k$値の求め方



□　ＭＣＩへの結合次数及び結合タイプの導入

C_K 値を求め、この値を基準としてＭＣＩを求める時、化合物構造式の複雑さを情報として取り入れるべく結合次数（ＢＯＮＤ　ＯＲＤＥＲ）という概念と結合タイプ（ＢＯＮＤ　ＴＹＰＥ）という２つの概念を導入する。

・結合次数（ＢＯＮＤ　ＯＲＤＥＲ）は C_K を求める時の対象となる結合と、その結合を形成する原子の数を拡大してゆくものである。
・結合タイプ（ＢＯＮＤ　ＴＹＰＥ）とは、結合が複数集まって一つの C_K を形成する時の集合形態に関する情報である。

□　結合次数（ＢＯＮＤ　ＯＲＤＥＲ）について

結合次数は基本となる C_K 値を求める時に対象とする結合や原子数を規定するものである。　次数が小さければＭＣＩの値は大きく、次数が増大するにつれてＭＣＩの値は小さくなる。

□　結合タイプ（ＢＯＮＤ　ＴＹＰＥ）について

ＴＹＰＥは C_K としてまとまった単位（特に次数が大きくなった時）の形を規制するものである。

・ＰＡＴＨは最も単純な形をしており、結合が直線上に繋がっているものを意味する。この時、次数が１のものは直線であり、ＰＡＴＨとみなす。
・ＣＬＵＳＴＥＲは分岐した形状を持つ C_K となる。　従って、次数が３以上で現れる
・ＰＡＴＨ－ＣＬＵＳＴＥＲは C_K 内部にＰＡＴＨ部分とＣＬＵＳＴＥＲ部分を持つ。

ここまでに述べたＭＣＩは情報として化合物の構造式中の結合関係に関する情報だけに関して数値データ化合物しているものである。　実際の化合物はその構造式中に様々な原子を含み、また結合も様々なものがある。　この情報をＭＣＩの中に組み入れない限りＭＣＩの情報の価値は高くない。

実際のＭＣＩではこのような原子種及び結合種を考慮した形で値が求められている。ＭＣＩの値の求め方の違いにより同じ化合物から様々なＭＣＩが算出される。　以下にその計算により得られるパラメータの表示方を示す。　詳細な計算方法については成書を参考されたい。

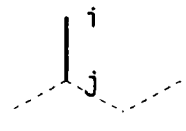
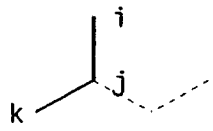
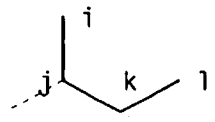
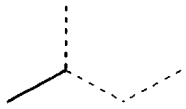
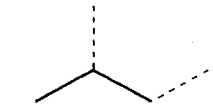
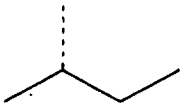
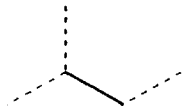
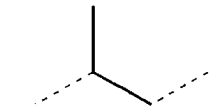
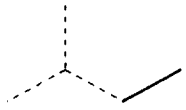
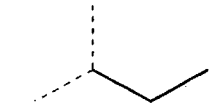
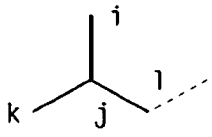
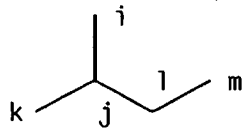
例）　$^{1}\chi_{P}$ ：　結合次数１、ＰＡＴＨタイプの C_K 値を基本として求めたＭＣＩ値

$^{4}\chi_{PC}$ ：　結合次数４、ＰＡＴＨ－ＣＬＵＳＴＥＲタイプの C_K 値を基本として求めたＭＣＩ値

$^{1}\chi_{P}$ ：結合次数１、ＰＡＴＨタイプの C_K 値を基本として求めたＭＣＩ値にリング補正を加えた値

$^{1}\chi_{P\nu}$ ：結合次数１、ＰＡＴＨタイプの C_K 値の計算にヘテロ原子を考慮して求めたＭＣＩ値

例）ＭＣＩにおける次数と結合タイプの概念及びC_k計算式

TYPE	ORDER 1	2	3	4
PATH				
				
				
				
CLUSTER				
PATH－CLUSTER				

□ 次数（ＯＲＤＥＲ）が異なる時のC_kの計算式（次数１～４について）

$$\text{次数 1} = \sum_{s=1}^{N_m} (\delta_i \delta_j)_s^{1/2}$$

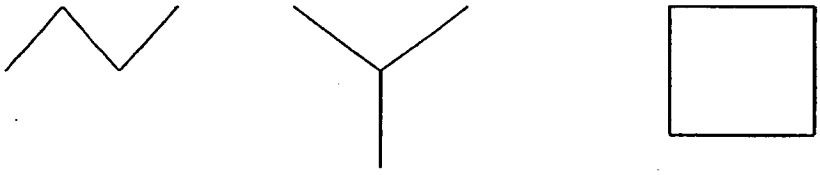
$$\text{次数 2} = \sum_{s=1}^{N_m} (\delta_i \delta_j \delta_k)_s^{1/2}$$

$$\text{次数 3} = \sum_{s=1}^{N_m} (\delta_i \delta_j \delta_k \delta_l)_s^{1/2}$$

$$\text{次数 4} = \sum_{s=1}^{N_m} (\delta_i \delta_j \delta_k \delta_l \delta_m)_s^{1/2}$$

Ｑ Ｕ Ｉ Ｚ

設問：　以下の化合物を分子結合インデックス値の大きい順に並べなさい。
並べた結果を見て、並び方の特徴を化合物の構造式とどのような関係にあるのか吟味しなさい。　但し、平方根は求めなくても良い。



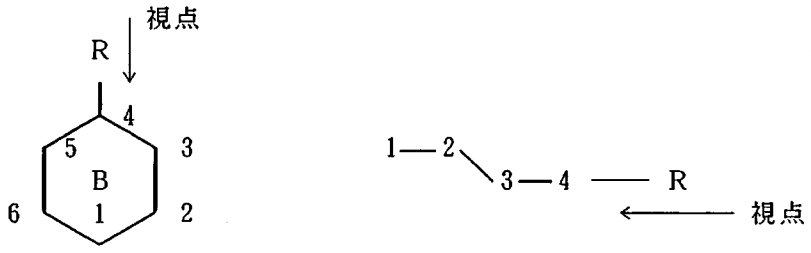
3．1．　トポグラフィカル記述子

トポグラフィカルデータは化合物の３次元情報を数値データに置き換えたものであり、トポロジカルデータが化合物の２次元的情報を扱うデータである為、相互に化合物の情報を補完しあう関係にある。

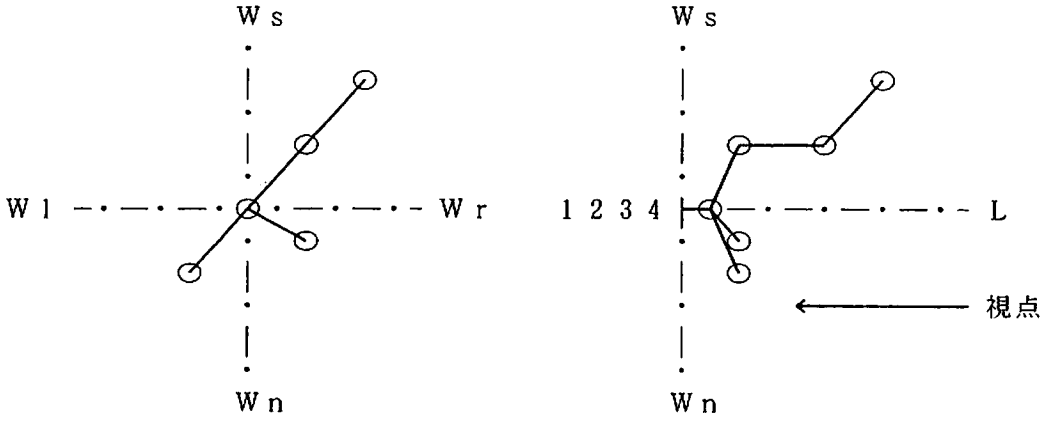
化学の分野においては立体情報を反映した数値データの必要性が高いが、現在この要求を完全にみたした変換方法は存在していないといえる。　現在比較的頻繁に用いられるトポグラフィカルパラメータとしてはTaftのＥｓ定数、Verloop のSTERIMOLパラメータ（後述）、分子の形状（後述）等のパラメータがある。

①ＳＴＥＲＩＭＯＬ　ＰＡＲＡＭＥＴＥＲ

このパラメータは化合物の置換基Ｒの３次元的立体情報を記述するのに用いられる。特に、重回帰手法によるＨＡＮＳＣＨ／ＦＵＪＩＴＡ法等に用いられて数多くの実績を有する構造活性相関には重要なパラメータである。



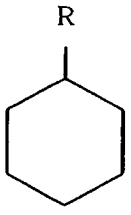
パラメータは化合物の基本構造部分（図中１～６で示されるＢ部分）と置換基Ｒ部分とに分けた時、基本構造部分と置換基Ｒ部分とが直結している結合をＲの方からＢに向かって見た時の置換基Ｒの占める空間上の領域をそれぞれの軸方向について分割した時の値を要素データとするものである。



ＳＴＥＲＩＭＯＬ＝（Ｗｌ，Ｗｒ，Ｗｓ，Ｗｎ，Ｌ）
＝（1.5，2.5，2.0，1.5，4.0）

ＱＵＩＺ：

以下の化合物についてＳＴＥＲＩＭＯＬパラメータを求めよ。



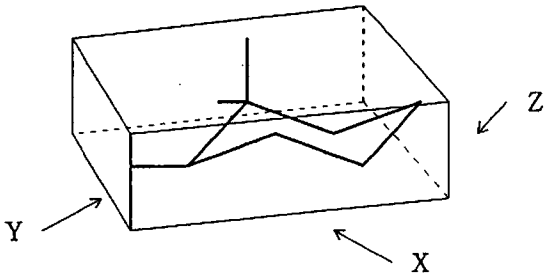
①　Ｒ＝　ＣＯＯＨ

②　Ｒ＝　ｔ－Ｂｕ

③　Ｒ＝　ｃｙｃｌｏｈｅｘａｎ

② 分子全体の形状に関する幾何学的情報（ボックスパラメータ）

化合物の３次元構造式をそのまま長方形のボックスにいれる。 このボックスの各軸の長さとその比とをパラメータとする。



パラメータ１＝	X
パラメータ２＝	Y
パラメータ３＝	Z
パラメータ４＝	X／Y
パラメータ５＝	X／Z
パラメータ６＝	Y／Z

このパラメータにより、分子全体の立体的な形状についての情報がえられる。 例えば、分子が平面に近い、細長い、立方体に近い等の情報である。

３．２． 物理化学パラメータ

物理化学パラメータは化合物の物理化学的特性に関するパラメータを総称して述べたものである。 この様なパラメータは数多く存在するが、その性格上説明変数のみならず目的変数として用いられる事も多い。 この物理化学パラメータの特徴を長所と短所としてまとめると、

長所： 実際の物性を数値データとして用いている為、その情報と原因との因果関係を特定する事が容易な時が多い。
短所： 実際の物性データを用いる為、データの種類によってはそのデータの収集が困難である事が多い。

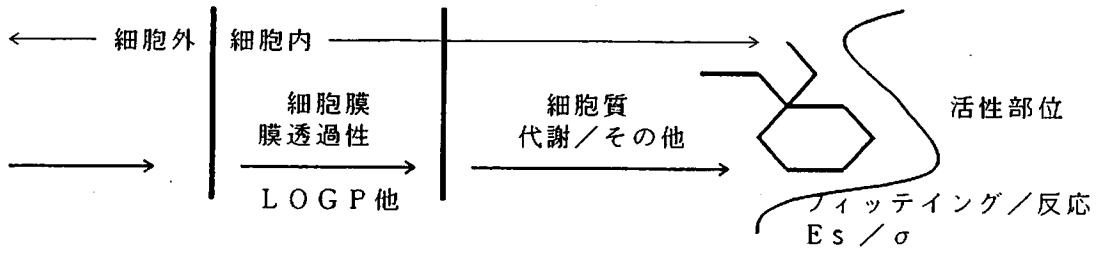
□ 物理化学パラメータの種類

この物理化学パラメータには多くのものがあるが、物性では融点、蒸気圧、沸点、その他のパラメータがある。 さらに分子軌道法計算により様々な種類のパラメータを求める事が可能である。

例）物性等 ： 分子量、融点、沸点、分子屈折率、ＬＯＧＰ、Hammett σ他
分子軌道法： 電子密度、ＨＯＭＯ、ＬＵＭＯ、分極率、双極子モーメント他

□ 構造活性相関等で用いられている物理化学パラメータ

構造活性相関分野では薬物の生体内の挙動に直結するパラメータが好んで用いられているが、特にＬＯＧＰパラメータは薬物の細胞外から細胞内への移動をモニターするパラメータとして重視されている。 この他、構造活性相関に使われるパラメータとしては薬物の動態を基本とし、薬物の細胞内における挙動の各ステージ毎に特定のパラメータが用いられている。



構造活性相関、特にＨＡＮＳＣＨ／ＦＵＪＩＴＡ法に用いられるパラメータは上図で

示したような薬物の生体内における移動過程を設定し、その個別の過程を最も良く説明するパラメータを数多く用意している。

中でもＬＯＧＰは薬物の細胞膜透過をシミュレートするのに最適なパラメータとして最も使用頻度の高いパラメータである。　このパラメータは平衡状態にある水層と油層との間の分配定数をパラメータとするものである。　この時、油層は細胞内の状態に最も近い状態を再現するものとしてｎ－Ｏｃｔａｎｏｌが用いられる。

□　ＬＯＧＰパラメータの定義式

・ＨＡＮＳＣＨ－ＲＥＯによるフラグメント付加方式によるＬＯＧＰ値推算。

$$LOGP = LOG \frac{[\ C\ ]\ lipid}{[\ C\ ]\ aqueous}$$

[C] lipid　：　平衡状態における油層中の濃度
[C] aqueous ：　平衡状態における水層中の濃度

□　ＬＯＧＰ推算プログラム

(1)ＣＬＯＧＰ

以下に述べるHANSCH-REOによるフラグメント付加方式によるLOGP推算方式に従ったプログラムであり、ADAPT のなかの記述子創出プログラムの一つとして開発された。
このプログラムは後にREO の元に移されてMEDCHEM ソフトウェアの中心プログラムの一つとして展開されている。　現在はＣＬＯＧＰ II となっている。

(2)Ａb-initio　ＬＯＧＰ

ＬＯＧＰ推算のパラメータとしてab-initi分子軌道計算より得られる静電ポテンシャルと水－薬物間電荷移動量、及び分子表面積を用いてＬＯＧＰを推算する。

この他、様々な形式による化合物構造式表示機能等を備えている。

(3)ＭＥＤＣＨＥＭ　ＳＯＦＴＷＡＲＥ

ＰＯＭＯＮＡ大学のＲＥＯが中心となって開発しているプログラムで、現在はDAYLIGHT COMPANY　が開発提供している。　日本は京大化研の西岡助教授が窓口となっている。

特徴は化合物構造式のシステムへの入力から、入力された構造式を用いてのLOGPパラメータ及びMRパラメータ等の推算機能と、構造活性相関(特にＨＡＮＳＣＨ／ＦＵＪＩＴＡ法用) において頻繁に用いられる物理化学パラメータのデータベース機能がある。

データベース機能は、ＨＡＮＳＣＨ／ＦＵＪＩＴＡ解析用パラメータのデータベースとしては現在世界1 の規模と内容を誇っている。

主機能：
① 化合物構造式のSMILES線型表記法による入力及び構造表示
② 物理化学パラメータ推算機能(MR:分子屈折率、LOGPパラメータ)
③ 物理化学パラメータ(HANSCH-FUJITA解析用) データベース機能

□　ＬＯＧＰパラメータ推算

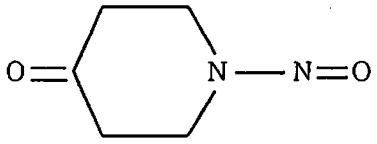
ＬＯＧＰパラメータは構造活性相関分野では極めて重要なパラメータである。　しかしながらこのパラメータを実験で求めるのは大変な作業であり、化合物が存在しなければこのパラメータは求められない。　従ってこのＬＯＧＰ値を化合物の構造式から自動的に求めるという手続きが重要となる。　このＬＯＧＰ推算方式も様々なアプローチがされているが、最も幅広い化合物に適用出来、且つ得られた値の精度も高いアプローチとして、ＨＡＮＳＣＨ－ＲＥＯが提唱しているフラグメント付加方式による推算、及び守口らが行っている幾つかの定められた数値データを用いて線型重回帰手法によりＬＯＧＰ値を推定するアプローチとがある。

① フラグメント付加方式によるＬＯＧＰ推算式

$$LOGP = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_j F_j$$

a_i ： i番目のフラグメントの出現回数
f_i ： i番目のフラグメントに対するフラグメント定数値
b_j ： j番目の修正因子の出現回数
F_j ： j番目の修正因子の修正定数値

ＬＯＧＰ値計算例）



4-KETO-N-NITROSO-PIPERIDINE

フラグメント定数

フラグメント	出現回数	フラグメント定数	総　和
— CH_2 —	4	0．66	2．64
ケトン	1	−1．90	−1．90
N−N＝O	1	−2．45	−2．45

修　正　定　数

リングボンド	（n−1）（−0．09） ＝ 5（−0．09）	−0．45
極性基修正	2X［−（0．20）（$f_1 + f_2$）］ 2X［−（0．20）（-2.45 - 1.90）］ 2X［0．87］	1．74

$LOGP_{CALC}$ ＝ −0．42
$LOGP_{OBS}$ ＝ −0．47