1.パターン認識法の基本

1.1 パターン認識法による構造一活性相関解析の基本

① 構造一活性相関解析での基本原理

パターン認識法による構造一活性相関解析の基本原理は、図１で示されるような“情報等価関係”を基本とする。

つまり、“化合物構造式と薬理活性との間に潜む何らかの関係（ブラックボックス部分）は、パターン認識手法を用いて薬理活性を 100%正しく説明した時に用いられたパラメータ*1)（数値データ）中に潜んでいる”すなわち、ブラックボックスとパラメータは情報論的に等価である（図１）という前提に立っている。従って、構造一活性相関を目的とする時には、用いるパラメータは化合物構造式を基本として創出したものである事が絶対条件となる。

＊１）このパラメータは、以下に述べるケモメトリクス分野では‘記述子(Descriptor)’と呼ばれる事が多い。

なお、薬理活性が化合物の種々物性データ・スペクトルデータ等に置き換わるならば、それぞれ構造一物性相関、構造一スペクトル相関と言われる研究であり、これらの解析はケモメトリクス*1)といわれる分野に属している。このケモメトリクスに関しては専門書があるのでそちらを参照していただきたい。

＊１）‘Chemometorics’；統計やパターン認識手法を用いて化学に関連する種々の問題を解析するアプローチの総称である。米国ワシントン州立大学の B.R.Kowalski により提唱された。

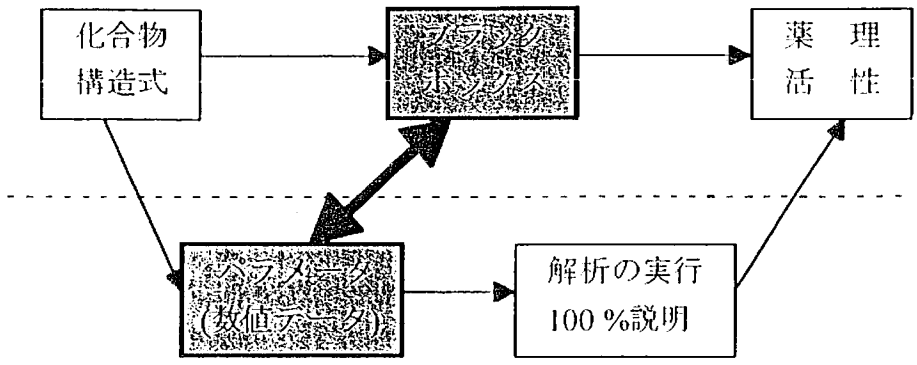


図１.パターン認識による構造一活性相関の基本

図１で示されるように、ブラックボックスに盛り込まれている情報の内容は解析に用いられたパラメータ中に潜んでいる。従って、構造一活性相関に関する解析とはこの薬理活性の説明に用いられたパラメータ中に潜む化合物構造

式に関する情報を読みとる事である。従って、この解析で利用されるパラメータはブラックボックスの内容を説明出来る内容を持ち、同時に化合物構造式との関係が明確であることが求められる。当然のことであるが、分類能力が高いことは大前提である。

②パターン認識法による構造一活性相関の特徴と利点

パターン認識法による構造一活性相関の最大の特徴は、解析の柔軟性の高さと、発見型のアプローチを取るという二点である。これら二点の特徴について簡単にのべる。

・柔軟性の高さ（適用目的、化合物構造式、活性データ、毒性、他）

・適用目的／パターンの広さ

パターン認識法による構造一活性相関は前項で述べたように“情報等価原理”を基本としているために他の手法と比べて制限事項が少なく、様々な点での柔軟性が非常に高いことである。つまり、解析を行う時に必要となる薬理活性データの種類、パラメータでの自由度が高い為に、様々な目的への適応が可能となる。

例えば、著者が行ってきたLead候補化合物検索 (Lead Retrieval) やLead候補化合物再構築 (Lead Re-construction)のアプローチは最近注目されつつあるコンビナトリアルケミストリ／HTSでのプレスクリーニングやLead候補化合物群の創出という観点で大いに利用出来る。このうちLead候補化合物再構築はDe Novoデザインの基本機能を備えている。また、同様にコンビナトリアルケミストリ／HTSで重要な分子多様性の評価の問題は化学パターン認識の世界では既に確立されたアプローチである。これらの詳細については第　章にて述べる。

・化合物構造式の多様性

パターン認識法による構造一活性相関を他の手法と比べた時の最大の特徴は、構造の大きく異なる多様な化合物群を同時に扱う事が可能という点である。このように構造式が大きく異なる化合物群の活性を予測したり解析することはHansch-Fujita法では不可能である。また、ドラグレセプター理論に基づくドッキング主体のアプローチはHansch-Fujita法と比べて化合物構造式の自由度は高いが、パターン認識法と比べれば低くなる。また、最近展開されている3-D QSARではHansch-Fujita法と同様に定量的な活性予測を行えるが、化合物構造式が大幅に異なっている場合は手続き上での限界が生じてくる。つまり、3-D QSARの実施には三次元化合物の重ね合わせが必要であるが、構造式が大きく異なる場合はこの重ね合わせが事実上不可能となる。

・活性データの種類

薬理活性データとして、パターン認識法ではほとんど総ての種類のデータを扱える。但し、解析手法の種類が数値データの種類により異なってくるので注意が必要である。例えば、効く／効かないの二クラスデータは殆どの判別分析手法が適用出来るが多クラスデータの場合は適用出来る手法は限定される。さらに、実際の手続き上で重要な特徴抽出手法も多クラスデータに適用されるものは少ない。このように、データの種類により適用可能な解析手法が制限される点への配慮が必要である。

・毒性予測への適用

毒性予測をパターン認識法以外の構造一活性相関手法で行うことは困難である。毒性は通常の薬理活性と異なり、メカニズムが明確でなく、毒性の定量化も困難であり、従って仮説検証型のアプローチを取ることは極めて困難である。さらには、評価対象とする化合物構造式の変化が大きすぎ、この点からも通常の構造一活性相関手法で扱うことは殆ど不可能である。

毒性予測はパターン認識法でしかアプローチ出来ないと言い切れる。これは、先にも述べた毒性独特の様々な問題をクリア出来るのは事実上自由度の高いパターン認識法だけであるためである。

・発見型のアプローチ

パターン認識法による構造一活性相関の基本原理は“情報等価関係”であることは既にのべた。これは他の構造一活性相関手法が独自の理論に基づいて展開されているのとは大きく異なる。この差は、パターン認識法が“発見型”の手続きを取り、他の手法が“仮説検証型”の手続きを取ることに反映される。

ここで発見型の手続きとは、化合物構造式と薬理活性との因果関係に関する前提条件を設けることなくデータ解析し、その解析結果に従ってサンプル中に存在する因果関係を導き出すことを意味する。一方、仮説検証型の手続きとはあらかじめ何らかの因果関係を想定しておき、その因果関係に基づいて導かれた様々なツール（例えば、パラメータ等）を用いて解析するアプローチである。解析が成功すれば前提とした因果関係が存在することになり、解析が失敗すればその系には前提とした因果関係が存在せず、他の因果関係の存在可能性が証明されることになる。従って、仮説検証型のアプローチを取る時は利用されるパラメータの種類は仮説の内容により大きく限定される。このように解析の手続きが異なれば、手続きを実施するのに準備するデータも異なってくる。

パターン認識法による解析では、前提条件を設けずに解析を行い、目的とする因果関係はデータ解析後に取り出される。従って、解析の初期段階では総ての可能性を加味した解析が必要である。このために、解析に先だって用意されるパラメータは可能な限り総ての可能性を網羅したものを揃えることが求められる。結果として、パターン認識法による解析で利用される記述子の数は他

の解析手法と比較した場合、桁違いに大きな数の種類の記述子群を用いることになる。一方、パターン認識法以外のアプローチではその仮説となる因果関係に関与した記述子しか利用しないため、解析時に利用される記述子数は極端に少なくなる。この場合、前提とする因果関係と関係の無い記述子群はたとえ薬理活性と何らかの関係があったとしても解析には用いない。

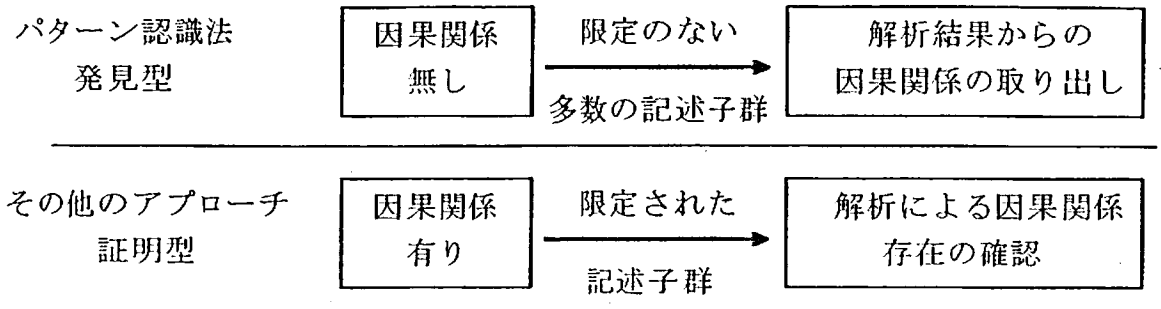


図　．パターン認識法とその他の構造ー活性相関手法とのアプローチの違い

1.2　パターン認識法による構造ー活性相関で用いられるパラメータの特徴

①パターン認識法と Hansch-Fujita 法で用いられるパラメータの違い

パターン認識法でも Hansch-Fujita 法でもその解析過程において種々のパラメータを用いる。これらのパラメータは単なる数値データではあってもその内容は大きく異なっている。以下にこれらの差異について簡単にまとめる。

・　パラメータの基本と特性の違い

パターン認識法で用いられるパラメータと Hansch-Fujita 法で用いられるパラメータとでは、データの形式や情報の内容等に大きな差異がある。これは、パターン認識法による解析で用いられるパラメータは単に目的活性の変化を効率良く説明するという単目的であるためにパラメータの内容に対する自由度が非常に高いのに対し、Hansch-Fujita 法で用いられるパラメータは厳密な薬物熱動力学理論に基づいているために運用上いくつかの制限事項を持つ事である。

しかし、相関解析という観点に立つならばパターン認識法で利用されるパラメータの自由度が高いだけその分情報の解析も的が絞りにくくなることは否めない。一方、Hansch-Fujita 法は自由度が小さい分、薬理活性やメカニズムとの相関がより直接的であり、明白である。

・　パラメータを基本とした要因解析力の違い

パターン認識法で利用されるパラメータは分類のみの単目的であるため、実際の解析で利用するパラメータはどのような種類のパラメータを用いてもかまわない。実際、パターン認識法による構造ー活性相関で用いられている大部分のパラメータは種々の変換アルゴリズムに基づいて化合物構造式から数値

データへと変換されたものである。従って、これらのパラメータは化合物構造式と直接的な相関を持つが、薬理活性やメカニズムとの相関は持たない。しかし、化合物構造式と薬理活性間に相関が存在することは明白であり、第一節で述べたように、間接的ではあっても解析で用いられたパラメータは構造一活性相関に関する情報を有しているはずである。このように、パターン認識法による構造一活性相関では薬理活性やメカニズムと直結した議論が出来ないために、要因解析という観点からは多少切れ味が悪くなる。この故に、パターン認識法による構造一活性相関での情報解析力は Hansch-Fujita 法よりも弱くなることは否めない。

Hansch-Fujita 法で用いられるパラメータは必ず何らかの形で生体内の薬理活性メカニズムに関与、あるいは代表している事が前提である。このために、ひとたび回帰式が得られれば、その得られた回帰式から構造一活性相関に関する情報を取り出す事では非常に強力なパワーを発揮出来る。つまり、単なる構造式との相関のみならず、薬理メカニズムをも考慮した議論が可能である。これが Hansch-Fujita 法の最大の特徴であり、強みである。

パターン認識法： パラメータ　・　化合物構造式　・　薬理活性
Hansch-Fujita 法： パラメータ　・　薬理活性メカニズム／構造式　・　薬理活性

・パラメータに対する制限事項の違い

前節で説明したようにパターン認識法による構造一活性相関は“情報等価性”を基本としており、この故に用いるパラメータの種類や特性等に関する大きな制限事項は特に存在しない。　さらに加えて、パターン認識法による構造一活性相関では化合物構造式、解析手法の種類や薬理活性データの形式等に関する特別な制限事項を持たない。これが化合物構造式、使用するパラメータの種類、さらには用いる薬理活性データの種類や解析手法等に多くの制限事項を持つ Hansch-Fujita 法との大きな違いである。さらには、同一レセプターに作用する化合物であるとの前提条件を必要とするドラグレセプター理論に基づくアプローチ（第　章）との最大の違いである。このように、パターン認識法による構造一活性相関は様々な点での自由度の高さが最大の利点である。但し、解析目的にあわせたデータセットを用意する必要があり、また、解析内容や結果はデータセットが有するデータ内容に左右される事を意識しなければならない。

ここではパターン認識法と Hansch-Fujita 法、ドラグレセプター理論に基づいたアプローチ間の種々特性の大まかな違いについて簡単にまとめる。

パターン認識法　：	
・パラメータの種類	化合物構造式に関するものデータあれば利用可能
・薬理活性データ	連続データ、バイナリデータ、カテゴリーデータ
・化合物構造式	構造式上で特に制限は無い
・その他	定量的／定性的構造ー活性相関

Hansch-Fujita 法　：	
・パラメータの種類	薬物熱動力学に基づいた物理化学的パラメータ
・薬理活性データ	連続データ
・化合物構造式	基本骨格が固定された同族体化合物群
・その他	定量的構造ー活性相関

ドラグレセプター理論に基づいたアプローチ：	
・パラメータの種類	レセプター・リガンド相互作用に関するパラメータ
・薬理活性データ	用いるデータに制限はないが、解析は定性レベル
・化合物構造式	同一レセプターサイトに関与する類似化合物群
・その他	定性的構造ー活性相関

②起源や情報内容の異なる各種パラメータの混在使用の可否

パターン認識による解析に Hansch-Fujita 法で利用されるパラメータや、その他のアプローチで用いられるパラメータを用いる／混在させることは解析上特に問題ない（但し、統計上の問題に起因するいくつかの制限事項をクリアしていることが必要である）。この時、利用されたパラメータは通常のパターン認識法による構造ー活性相関で用いられるパラメータ（構造式から創出されたパラメータ）で得られない情報を提供する。たとえば、Hansch-Fujita パラメータは薬物熱動力学上の薬理メカニズムに関する情報を提供し、ドラグレセプター理論を基本としたアプローチによるパラメータはレセプターサイトでの薬物相互作用に関する情報を提供する。

・利点および留意事項

このように情報ソースの異なるパラメータ群を用いる／混在させることはパターン認識法による構造ー活性相関の弱点である要因解析力を強化すると同時に、単一種類のパラメータだけでは不足する情報を相互補助させる事が可能となる。ただし、情報の種類や内容が同じでデータ（値）のみが異なるパラメータを混在した場合は要因解析の妨げになるばかりか、解析の精度や信頼度を低下させることになるので注意が必要である。また、これらのパラメータが持つ固有情報は、解析に用いたパターン認識手法の種類やデータ処理法によってはその情報内容を失う時があるので注意が必要である。さらに、個々のパラ

メータが前提とする条件が解析で満たされていることも必要である。このような点に留意しつつ解析するならば、特性や情報の異なるパラメータを混在させて解析することは要因解析という観点上非常に望ましいアプローチとなる。

既に、3-D QSAR の CoMFA 法等では化合物周辺状況の情報を表す CoMFA パラメータと Hansch-Fujita 法で利用される LogP パラメータとを混在させた解析が多数報告されている。従来は Hansch-Fujita 法、パターン認識法、そして 3-D QSAR といった解析手法が先にあり互いに独立していた。従って、解析に用いるパラメータ群も必然的に決まっていた。しかし、今後は解析目的に応じ、最適と思われる多様なパラメータを混在させつつ解析することが主流となるであろう。パターン認識法は原理上このような使用にも耐えうるアプローチである。

1.3 パターン認識法で用いられるパラメータ

1.3.1 パラメータの種類

パターン認識法で利用されるパラメータは化合物構造式を基本として創出されることが基本である。もちろん、前項で述べたように Hansch-Fujita 法やその他の解析手法で用いられるパラメータを用いることも可能である。

現在、パターン認識法で頻繁に利用されるパラメータ群は、トポロジカルパラメータ、トポグラフィカルパラメータ、物理化学的パラメータ、部分構造パラメータ、その他の五種類に大きく分類される。これらのパラメータはそれぞれ化合物構造式に関する異なる情報を取り出してくる。

①トポロジカルパラメータ

化合物を構成する原子や結合に関する二次元結合情報の数値データ化。

トポロジカルパラメータの歴史は古い。当初は化合物の融点や沸点を構造式から求める時に用いるパラメータとして利用されていたが、1970 年代後半より構造一活性相関にも L.B.Kier と L.H.Hall らにより本格的に利用されはじめた。

Kier & Hall らにより導入されたトポロジカルパラメータはモレキュラーコネクティビティインデックス（分子結合インデックス：Molecular Connectivity Index) [*1)] と呼ばれ、当初は炭化水素化合物群への適用から始まり、機能強化に伴い一般有機低分子へと拡大適用されてきた。

＊１）L.B.Kier, L.H.Hall, W.J.Murray, and M.Randic, J.Pharm.Sci., 64,1971 (1975).

このモレキュラーコネクティビティインデックス（以降、ＭＣＩと略す）は計算するときの条件の差異により多種多様のインデックスを創出する。これらの創出されたインデックスは記号で表されるが、$^{1}\chi$、$^{2}\chi$、$^{3}\chi_{p}$、$^{3}\chi_{pc}$、$^{1}\chi^{v}$等の多種類存在する。これらの詳細については文献を参照されたい。

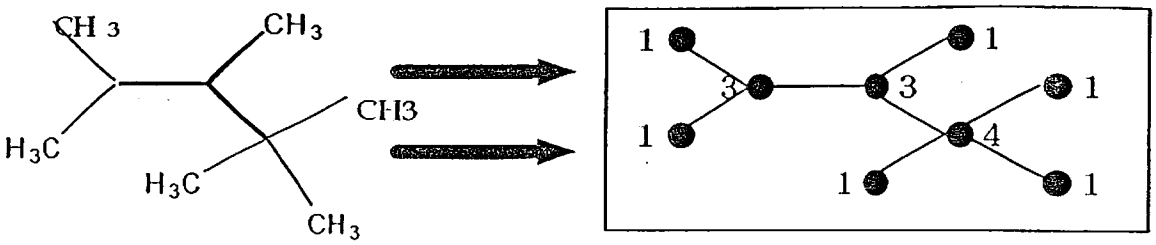


図１．構造式と基本骨格図。炭素原子が●で示され、水素原子は省略されている。基本骨格図中、原子に付随された数値は炭素原子が結合する水素原子以外の原子の数である。

ここでは2,2,3,4 - テトラメチルペンタンを例に取り、ＭＣＩのうちで最も単純な$^{1}\chi$を求める基本的な計算手続きを順番に説明する（図１）。

・化合物の結合情報に従って基本骨格図を作成する。

この時、原子種と結合の種類は無視される。トポロジー的には、結合部分は“パス(Path)”または“辺(Edge)”、原子部分は“点(Node)”と呼ばれる。

・基本骨格図上の点に値を付ける

個々の点（以下、ノードと呼ぶ）の結合状態により値を付ける。値は該当原子が結合している水素以外の原子の数を取る。この値はδ_iで示される。なお、原子種や結合種を考慮した計算も可能であるが、ここでは言及しない。以下に、図１の基本骨格図における個々のパス単位でδ値を示す。

（δ_i、δ_j）：(1,3), (1,3), (3,3), (1,3), (3,4), (1,3), (1,3), (1,3)

・各パス単位にその両端のノードの値を掛けることでパスの値を求める。

各パスの値（$\delta_i \cdot \delta_j$）：３、３、９、３、１２、３、３、３

・各パス値のルートを求める。

$\sqrt{}$（$\delta_i \cdot \delta_j$）：1.732, 1.732, 3.000, 1.732, 3.464, 1.732, 1.732, 1.732

・分子を構成する前記パス値の逆数をとる。

$1/\sqrt{}$（$\delta_i \cdot \delta_j$）：0.577, 0.577, 0.333, 0.577, 0.289, 0.577,0.577,0.577

・前記パス値総和のルートを取り、その逆数を最終のＭＣＩ値とする。この値は$^{1}\chi$として示される。

$$^{1}\chi = \Sigma\,(1/\sqrt{}\,(\delta_i \cdot \delta_j)) = 4.084$$

この結果、2,2,3,4 - テトラメチルペンタンのＭＣＩの$^{1}\chi$値は4.084となる。

②トポグラフィカルパラメータ

化合物の三次元構造に関する情報で、化合物の三次元座標データから求められる。このようなパラメータとしてはHansch-Fujita法で一時期頻繁に用いられたSTERIMOLパラメータ、分子を長方形のボックスに入れ、そのボックスのX,Y,Z三長軸に関する情報を数値データ化した分子主成分パラメータ、分子の写像情報をパラメータ化したもの等多数存在する。ここでは分子主成分パラメ

ータを例に取り、説明する。

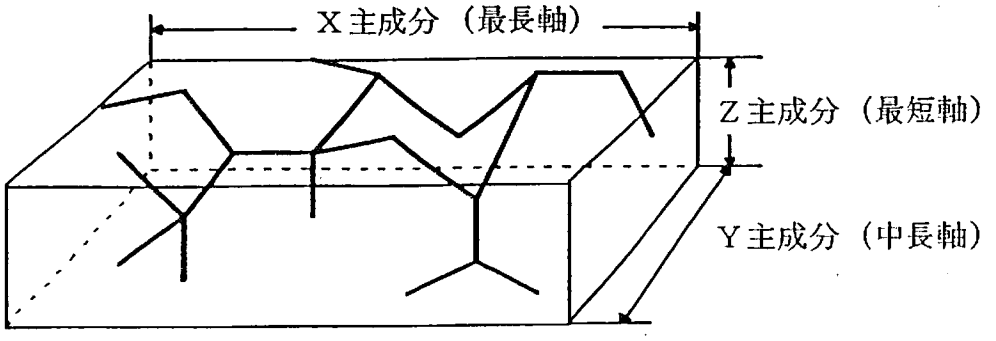


図　．三次元ボックスによるトポグラフィカルパラメータ例

図に示されるように化合物を三次元ボックス中に納める。この時のボックスの三長軸の値（X,Y,Z）とその比である（X／Y, X／Z, Y／Z）の値をパラメータとする。これらのパラメータにより化合物の三次元形状に関する情報が得られる。例えば、平面性の高い化合物はZ軸が小さく、かつ三種の比の値が大きいことで示される。同様に、球に近い形状、棒上の形状等の情報はこのパラメータを検討することで容易に得ることが出来る。

③物理化学的パラメータ（化合物の物性や物理化学的特性に基づく情報）

物理化学的パラメータは化合物の物性データをそのままパラメータとして利用するものである。従って、数値データとして表現出来る総ての物性はパラメータとして利用することができる。

一般的に、解析対象とする化合物の総てに実測値が得られる事は少ない。現在、物性の種類は限定されるが化合物構造式から直接物性を求める手法がいくつか発表されており、これらの手法を利用することで解析対象化合物群の物性をコンピュータ上で計算できる。このようにコンピュータで簡単に求められ、かつ構造--活性相関に利用されて比較的精度高く求められる物理化学的パラメータとしては、logP、MR（分子屈折率）、分子容積、分子表面積、分子軌道法から求められる種々の電子的パラメータ、分子力学法による歪みエネルギー等がある。

④部分構造パラメータ

ここで①から③にかけて説明してきたパラメータは何らかの形で化合物構造式と相関を持つ。しかし、このパラメータが持っている情報を化合物構造式へとフィードバックすることのし易さはパラメータの種類により大幅に異なる。構造一活性相関情報を構造式に置き換える作業は、化合物合成を担当する合成化学者にとり非常に重要なものとなる。従って、用いるパラメータが化合物構造式と単に相関を持つだけでなく、その情報は化合物構造式に簡単にフィ

ードバックされることが必要である。この構造式へのフィードバックが他のパラメータと比べて最も容易なものが部分構造パラメータである。この部分構造記述子を用いることが出来るのが、パターン認識法による解析の強みでもある。

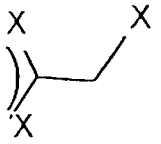
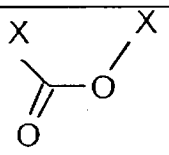
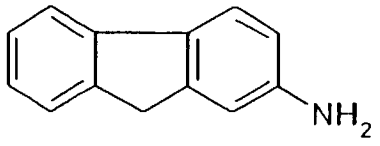
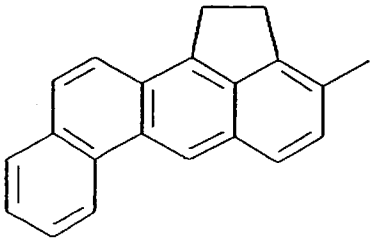
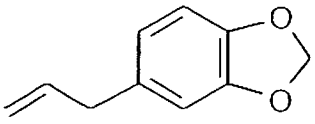
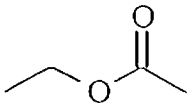
	部分構造式			
				
被検索化合物	部分構造数	MCI	部分構造数	MCI
	2	3.166	0	0.000
	2-Aminofluorene			
	2	3.301	0	0.000
	3-Methylcholanthrene			
	1	2.690	0	0.000
	Safrole			
	0	0.000	1	1.904
	Ethyl acetate			

図　. 部分構造式と個々の被検索化合物の部分構造数とＭＣＩの値

部分構造パラメータは指定された部分構造が被検索化合物の中に存在するか否かをチェックし、部分構造が存在する時はその部分構造数、あるいは検索の時に指定したした部分構造にＸ部分（水素以外の原子を意味する）の情報も含めて何らかのアルゴリズムにより数値データへと変換する。

図　にこの部分構造パラメータの例が示されている。部分構造数は部分構造が被検索化合物中に存在する数を、ＭＣＩはＸ（水素以外の原子）原子を含めた部分構造のＭＣＩ（モレキュラーコネクティビテイインデックス）値を求め、その平均値を出したものである。最初の部分構造では部分構造数だけでは 2-Aminofluorene と 3-Methylcholanthrene とを識別出来ないが、ＭＣＩの情報を用いることで区別が可能であることがわかる。このように、部分構造パラメータの基本形は識別能力が非常に小さいが、より高度な変換をすることでこの欠陥

をある程度補えることがわかる。しかし、二番目の部分構造で分かるように、部分構造が存在しない時の値は総て0となるために識別能力が他のパラメータと比較した場合弱いことがわかる。

この部分構造パラメータは単に情報の化合物構造式への変換が容易になるだけのみならず、例えばDe Novoデザインを行う時にLead候補化合物群を創出するのに必要な部品となる部分構造や官能基群の選択に利用することが出来る。ここで選択される部品はパターン認識による構造一活性相関により目的薬理活性の向上に重要な働きをすると判断されたものである。これら選択された部分構造や官能基群の組み合わせにより新規のLead候補化合物群を容易に創出する事が可能である。パターン認識による構造一活性相関の一つの展開形として筆者は、De Novoデザインの概念が確立する以前に"Lead候補化合物最構築(Lead Re-construction)"という概念を提唱し、実際に実験を行い、発表*1)している。このように部分構造パラメータを用いることでパターン認識法によるDe Novoデザインの実施が可能である。また、この"Lead候補化合物最構築(Lead Re-construction)"の技術は、コンビナトリアルケミストリ／HTSで利用されるスクリーニング対象化合物群の構造式創出にも利用可能である。このアプローチに関する詳細はDe Novoデザイン、およびコンビナトリアルケミストリ／HTSの節にて説明する。

⑤その他のパラメータ

以上のパラメータ群の他にも多数のパラメータ群が存在する。これらのパラメータのうち、パラメータ同士の演算により創出される演算パラメータについて簡単に説明する。この演算パラメータはパラメータの絶対数が少ない時にパラメータ数を増やす目的で利用されることが多い。しかし、この場合には以下の点で注意が必要である。単純なパラメータ同士の演算はパラメータ数を増加することは出来ても、パラメータが持つ情報の意味を演算により失ってしまう場合がある事である。構造一活性相関で最も重要な要因解析ができる可能性を低くする危険がある。従って、パラメータ同士の演算にて新規パラメータを創出する時は演算後のパラメータが持つ情報の意味が明確であることが望ましい。

図　はベンゼン環のオルト、メタ、パラ効果に関する情報を個別に取り出した部分構造パラメータ同志を加え合わせることで、ベンゼン環における総ての置換基効果の影響を一つのパラメータに集約したものである。パラメータ同志の演算式も問題で、この場合には個々のパラメータの総和を求めるべきで、積和を求めた場合は構造一活性相関解析上で無意味なパラメータとなる。

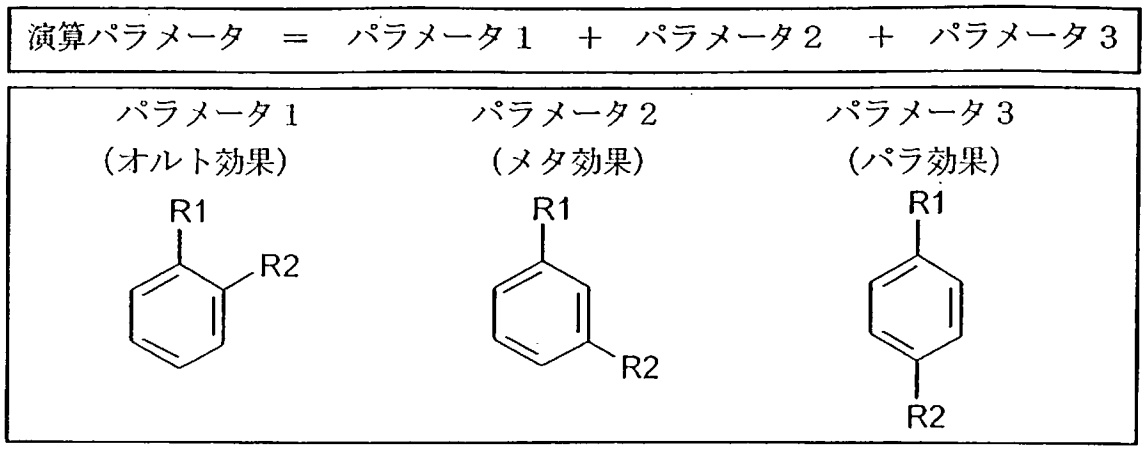


図　．パラメータ同志の演算による新規パラメータの創出

1.3.2 代表的な四種類のパラメータ群の一般的特徴

前項で示された四種類のパラメータは解析薬理活性構造一活性相関という観点から一般的な傾向を有している。ここで、これら四種類のパラメータについて分類能力と化合物構造システムきとの相関という観点から簡単に比較する。

分類能力は解析が成功するか否かを左右する大事な項目である。パラメータの分類能力が低い時は分類が成功しないばかりか、分類出来たとしてもパラメータ数が多くなり、解析の切れが鈍くなり、要因解析も困難となる。また、化合物構造式との関係は構造一活性相関の観点から重要なもので、この関係が明確でない時は解析情報を化合物構造式へとフィードバック出来ないためにドラグデザインや新規化合物への展開が不可能となる。従って、構造一活性相関で理想的なパラメータとは分類能力が高く、同時に化合物構造式へのフィードバックが容易なものが理想となる。

	トポロジカル	トポグラフィカル	物理化学的	部分構造
分類能力	◎	○	○	X
要因解析力	X	○	◎	○
構造との相関	X	○	○	◎

トポロジカルパラメータは化合物構造上の細かな差異を数値データ化することが可能であり、このために分類能力が高い。しかし、化合物構造式を数値データ化する過程は原子や結合単位で計算が行われ、結合の方向性等も情報としては取り出されない。従って、このパラメータから化合物構造式にフィードバックすることは極めて困難である。また、要因解析も構造しきと同様にこのパラメータから求めることは困難である。

トポグラフィカルパラメータは化合物の三次元情報を取り出す。この種類の

パラメータはトポロジカルパラメータと比較した場合化合物構造式との相関がイメージとして捕らえやすいことが特徴である。しかし、構造上の細かな差異を数値データの大きな差異へと変換することが出来ないために分類能力はトポロジカルパラメータより劣る。

物理化学的パラメータは物性との関係が明確であることから要因解析には最も優れたパラメータとなる。しかし、化合物構造との関係は大部分が明確にはなっていないため構造と活性との相関をダイレクトに認識する事は困難である。

部分構造パラメータは構造との相関が最も良く理解できるパラメータである。化合物の部分構造情報がそのままダイレクトに数値データへと変換されるため、解析後に行う要因解析時にイメージをつかみやすい。しかし、分類能力は部分構造の存在が前提となるために、該当する部分構造が存在しない化合物群は総て同じ０の値を取る事になる。つまり、同一の化合物と見なされることになる。この点で部分構造パラメータの分類能力は弱い。

1.4　解析手法と用いるデータの種類

1.4.1 解析手法概要

パターン認識手法（日本では多変量解析[*1]の一部として議論される事が多い）としては現在多種多様な手法が存在する。この分野の言葉として数年前から一般的に聞かれる言葉としては“ニューラルネットワーク”がある。また、基本概念を拡張するならばファジィといった言葉も一度は聞かれたことがあると思うが、このファジイ理論に基づいたパターン認識手法も多数開発され、家電製品の制御機構等に利用されている。

本著の主眼はこれらの解析手法を自分の研究に正しく使えるようにすることであり、個々の解析手法を議論するものではない。従って、パターン認識や多変量解析に関する詳しい内容等については他の専門書を参照されたい。著者としては“ケモメトリクス”関連著書の購読を勧める。これはケモメトリクスが化学の一分野である事と、パターン認識や多変量解析の専門者が書いたものではなく、化学者がパターン認識や多変量解析を道具として利用する立場から書かれているからである。本著では、パターン認識や多変量解析の概要説明に止め、説明も数式を極力避けて、可能な限り概念図を用いて説明する。

＊１：パターン認識と多変量解析

パターン認識と多変量解析の共通点は多次元のデータを扱うことである。違いは計算過程で統計的な手続きを踏むか踏まないかにある。パターン認識は統計量に関する情報（平均、分散等）を用いる事なく実行されるが、多変量解析はこれらの統計量を用いて実行される。

1.4.2 解析手法各論

多数存在するパターン認識法は解析の内容／形態から一般的に、①判別分析、②フィッティング、③クラスタリング、④マッピングの四種類に分類して議論されることが多い。ここではこれらの手法の他に五番目の分類としてグラフ解析法を加え、これらの解析手法に関する基本概念の簡単な説明とその代表的な手法を例に取り説明する。個々の解析手法の詳細については専門書を参照されたい。

①判別分析

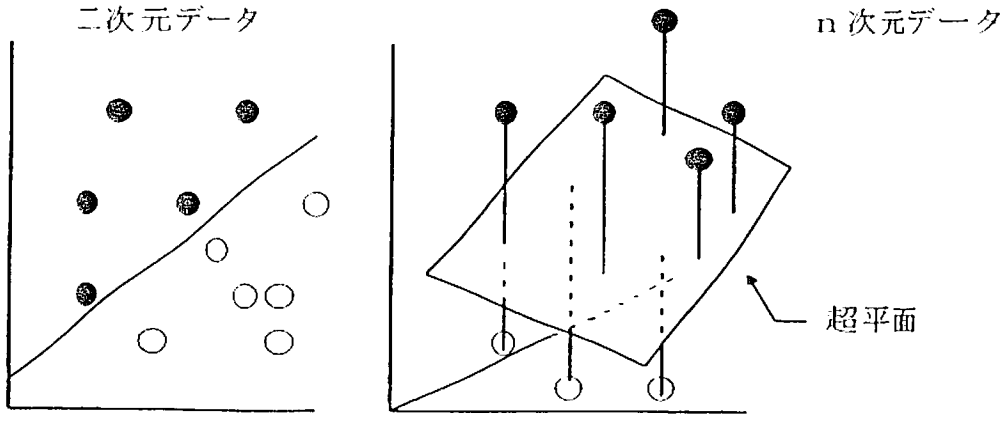
判別分析は解析対象となる母集団をクラス別（二クラス分類も含める）に分類するアプローチであり、パターン認識による構造ー活性相関解析では最も頻繁に用いられる。活性が効く／効かないの二値データであれば二クラス分類が、二クラス以上の多クラスを扱うのであれば多クラス分類手法が適用される。この二クラス分類と多クラス分類を一つの解析手法で解析出来る手法は少ないが、ニューラルネットワークはこの要求を実現できる。

判別分析法はサンプル群の分類を線形（超平面）で行うか、非線形（超曲面）で行うかで大きく二分類される。また、分類する為の道具（判別関数と呼ぶ）をどのようにして求めるかで、教師付き学習と教師無し学習法との二種類に分類される。この時用いられる判別関数は以下のような式となる。 d_i はi番目のパラメータを示し、 a_i はi番目のパラメータの係数である。

$$Y = a_1 d_1 + a_2 d_2 + \cdots + a_i d_i + \cdots a_n d_n + C$$

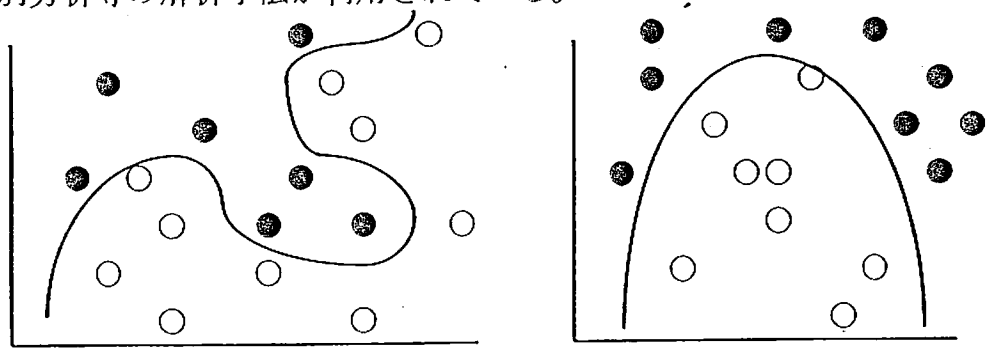
例えば二クラス分類であるならば、Yの値が正か0の時はクラス１に、負の時はクラス２へと分類することでクラス分けが実現される。この判別関数の善し悪しを計る評価基準は分類率や予測率を用いる。なお、分類率は内挿の、予測率は外挿の評価基準となる。

図　に線形の二クラス分類の概念図を示す。左図は二次元データを扱った時に直線で分類される様子を示し、右図は多次元（n次元）データを扱った時に超平面でクラス分類がなされる様子を示す。



図．線および面（三次元）／超平面（三次元＜）による二クラス分類

図　には二種類の非線形型の分類の概念図が示されている。左図で示されるように、曲線が複雑な形を取る非関数型の分類はニューラルネットワークが実施可能である。また、右図にあるような単純な関数型の曲線を取る分類はBayes判別分析等の解析手法が利用されている。



ニューラルネットワークによる
非関数型非線形分類

Ｂａｙｅｓ非線形判別分析による
関数型非線形分類

図　．二種類の非線形分類例

得られる情報）

本解析で得られる解析情報は判別関数として得られる。情報の種類としては、・分類に重要な働きをするパラメータの種類、・各パラメータがどちらのクラスに寄与するかの情報、・Ｙ値の所属クラス情報。これら三種類の情報が判別関数より得られる情報である。

実際の解析手法）線形学習機械法（Linear Learning Machine)

線形学習機械法はパターン認識に属するアプローチであり、パーセプトロンとも呼ばれる。この解析手法の特徴は、多変量解析と異なり実際の計算時に種々の統計量に関するデータを必要としない事である。この線形学習機械法は最近注目をあびたニューラルネットワークの原型であり、二手法間の大きな差異は線形学習機械法は超平面による線形分類手法であるのに対し、ニューラルネットワークは超曲面による非線形分類手法であることである。

分類に必要となる判別関数はサンプルデータを用いたエラーフィードバックトレーニング(Error Feedback Training)と言われる繰り返し学習により求められる。このプロセスは、・現時点での判別関数を用いて全サンプル群を分類、・この時誤分類されたサンプル群が正しく分類されるように判別関数を修正、・この修正された判別関数を用いて再び・のプロセスに戻る。この繰り返しを全サンプル群が正しく分類されるまで行うものである。・の修正により前回は正しく分類されたパターンが誤分類されるようになるため、全サンプルが正しく分類されるまで何回もの繰り返し学習（判別関数の修正）が必要になる。

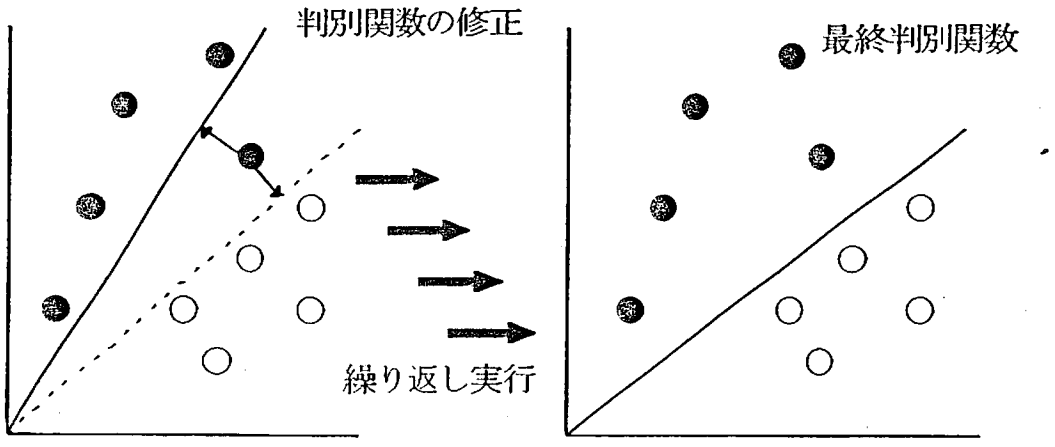


図　．エラーフィードバックトレーニングによる判別関数の修正

留意事項)

判別関数から得られる各パターンのＹ（目的変数）値の絶対値情報は利用出来ない。利用出来るのは単に符号（クラス）情報だけで、所属クラスの差異のみである。同様に、各パラメータの係数の絶対値（大小関係）情報も利用出来ない。これは分類過程で各パターン間の順位情報を用いないで判別関数が求められる為である。最初から存在しない情報が解析により得られる道理はない。パターン認識や多変量解析はデータ間に潜在する共通情報を顕在化するものであり、無から情報を創出するものではない。

その他の分類手法)

現在、構造一活性相関分野で多用される分類手法として様々なものが提唱されている。以下に構造一活性相関分野で利用される代表的な手法をリストアップする。個々の解析手法の詳細については専門書を参考されたい。

・判別分析　（最小二乗アルゴリズム）
・判別分析　（シンプレックスアルゴリズム）
・BAYES 線形判別分析
・BAYES 非線形判別分析
・ALS(Adaptive Least Squares)法

森口（現北里大学名誉教授）が提唱した解析手法[*1)]で、多クラスデータの扱いが可能でありながら一つの判別関数で活性を説明出来るため、Hansch-Fujita 法で扱えない多クラスデータの解析に多用されている。現在はファジイ理論を取り入れた Fuzzy ALS が良く利用されている[*2)]。

＊１：Moriguchi I. & Komatsu K., Chem. Pharm. Bull., 25, 2800（1977）.

＊２：Schaper, K.-Jurgen,'Quantitative Analysis of Structure-Activity-Class Relationships by (Fuzzy) Adaptive Least Squares',pp.244-280, in : Advanced Computer-Assisted Techniques in Drug Discovery (Methods and Principles in

Medicinal Chemistry, vol. 3), Van de Waterbeemd, H.,(ed.),VCH, Weinheim, 1994.

・最近隣法（K-NN 法）

パターン間の距離情報（通常はユークリッド距離）を用いて帰属クラスを決定するアプローチである。各パターンの帰属は、該当パターンに近在するK個のパターンが帰属するクラスの多数決により決定される。計算が簡単で、二クラス分類のみならず多クラス分類にも適用可能である。また、クラスパターンが互いにオーバーラップしているデータも良好に分類出来る特徴を持つ。このため、他の線形分類手法ではうまく分類出来ないデータ等の分類に利用される。

・ SIMCA(Soft Independent Modelling ofClass Analogy)法

本質的には④で述べるマッピング手法の一種であるが、クラス分類も可能であり、構造ー活性相関にはしばしば利用されてきた。最大の特徴としてはクラス単位で判別関数が得られるので詳細な議論が可能なことがある。また、本手法は PLS 手法へと発展し、3-D QSAR では次元圧縮手法として利用されている。

・ニューラルネットワーク

最近分野を問わず多用されている手法で、分類手法としては二クラス分類にも多クラス分類にも適用可能である。また、利用の形態により以下に述べるフィッティング手法にも利用可能であり、さらには次元の減少や増加等にも利用可能というオールマイティな解析手法である。但し、高い分類率を示すという特徴を持つが、予測率は低くなるという傾向があるので、利用目的によっては注意が必要である。

②フィッティング

サンプル群が最も誤差無く分布する直線（線形）や曲線（非線形：関数型、および非関数型）を求めるアプローチで、Hansch-Fujita 法で用いられる線形重回帰および非線形重回帰手法が代表的なものである。

この時用いられる回帰式は以下のような線形結合形式となり、形式上は前項の判別関数と変わらない。 ここで、d_iは i 番目のパラメータを示し、a_iは i 番目のパラメータの係数である。

$$Y = a_1 d_1 + a_2 d_2 + \cdots + a_i d_i + \cdots a_n d_n + C$$

この回帰式の評価は相関係数、標準偏差、F 検定等が用いられる。

得られる情報）

本解析で得られる情報は線形および非線形重回帰式として得られる。形式上重回帰式と判別関数とで差がないので各々から得られる情報の種類と数に差は無いが、議論出来るレベルは重回帰式の方がより細かなレベルまで可能となる。これらの情報は、・フィッティングを行うのに重要なパラメータの種類（d_i）、・各パラメータの Y 値に対する寄与（a_i）の大小関係、・各パラメータの寄与の方向（係数の符号）、・各サンプル間の Y 値の大小関係。これら四種

類の情報が重回帰式より得られる情報である。

図　に線形フィッティングと二種類の非線形フィッティングの様子を示す。この図中、非関数型のフィッティングはニューラルネットワークにより実現される。ニューラルネットワークではこのような重回帰式が存在しないため、原則として前記・データ示されたサンプルのＹ値の大小関係情報のみ得られ、その他の種類の情報入手には特殊な技術／手続きが必要となる。

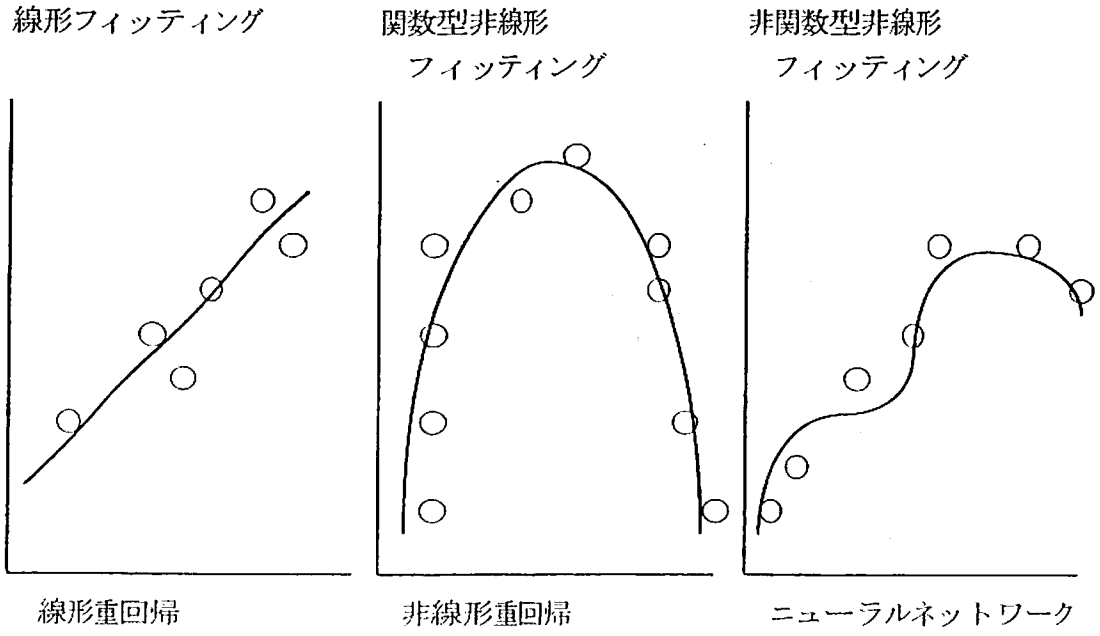


図　．三種類のフィッティング

留意事項)

判別関数と重回帰式との形式が同じ為、初心者は判別分析とフィッティング手法とで同じ解析精度を前提として議論しがちであるが、これは間違いであり注意が必要である。判別関数は定性的（クラス）レベルでの議論しか出来ないのに対し、重回帰式は定量的（大小関係）レベルでの議論が可能である。従って、判別関数の結果を用いてＹ値（目的変数）の相対的な大小（順位）を論じることは無意味である。また、各パラメータの係数について回帰式と同様のレベルで大小を論じることは危険である。定量的レベルでの議論をするためには回帰式を求め、その回帰式について議論することが基本である。

その他のフィッティング手法)

ニューラルネットワーク（非線形：非関数型フィッティング）は従来の線形及び非線形重回帰では扱えない非関数型のフィッティングを行う手法として注目を浴びている。非関数型なのでどのような複雑な問題に対しても対応可能という便利な側面を持つが、前項でも述べたように予測能力は弱いので既知データの扱いに止まる。未知／新規データの扱いには危険が伴う事を承知で行う

事が必要である。

③クラスタリング

クラスタリング手法はサンプル群を様々な基準に従って個々のクラスター（グループ）へと分類する手法であり、このグループ分けにおいて参照すべきクラスターデータ（教師データ）を必要としないのが特徴である。

クラスタリング手法は解析結果の出力形態の差異により、・階層的クラスタリング手法と・非階層的クラスタリングの二種類に分類される。階層的クラスタリングでの計算結果はデンドログラム（ゲームではトーナメント表と呼ばれる）というグラフが利用される。これに対し、非階層的クラスタリングでは単にクラスター単位でメンバーリストだけが出力される。

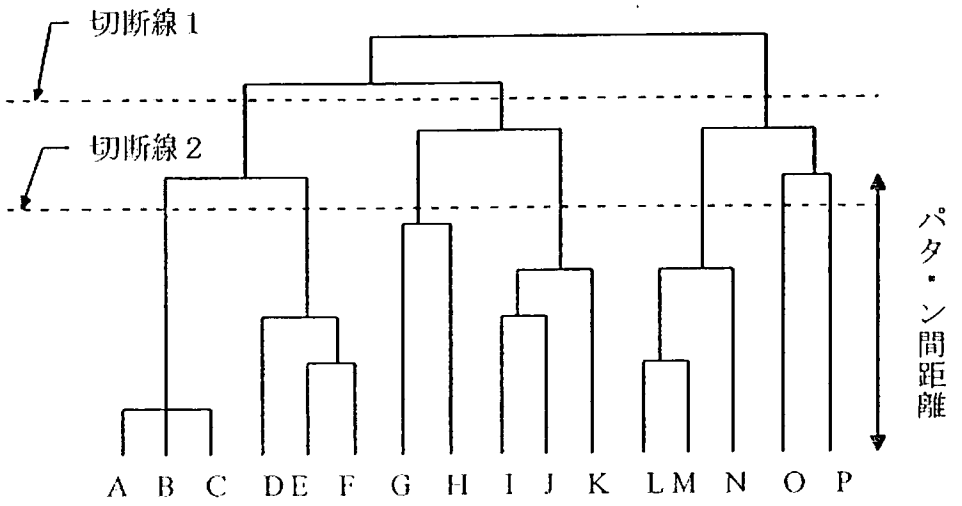


図　．クラスタリング手法のデンドログラム

このデンドログラムの解釈は以下のようになる。デンドログラムをＹ軸上の任意の位置でＸ軸方向にデンドログラムを切断した時、切断される縦線の数がクラスター数を示し、切断された縦線につながる総てのパターンが同一クラスターに属する。従って、同一のデンドログラムを用いても、切断位置が異なれば多種多様のクラスタリング結果が得られる。例えば図　において、切断線として１を採用した時はＡからＰの１６個のパターンは３個のクラスターに分類され、それぞれのクラスターはＡ～Ｆ、Ｇ～Ｋ、Ｌ～Ｐで構成される。また、切断線として２を採用した時は全体で７個のクラスターへと細分類される。なおこのデンドログラムにおいて縦線の長さは、各パターン（クラスター）間の距離を示す。

クラスタリング手法はさらに、パターン間の距離を求めるための距離基準、クラスターを形成するためのパターン同士の融合、あるいは分割方法の差異により更に細分化される。

得られる情報）

クラスタリングからは各パターンのグループ（クラスター）に関する情報が得られる。この情報から取り出すべき二次情報は、各グループがどのような要因でグループ分けされたかの原因を明確にすることから得られる。これには各グループを構成する個々のパターンの構造式情報に戻ること、用いたパラメータの特性を考慮すること、各グループ間やパターン間の距離関係を吟味するといった手続きが必要である。

留意事項)

クラスタリングの結果は計算過程（パターンの融合や、用いる距離尺度）により変わりやすい。また、デンドログラムの図の視認性は高いがその分結果の差異が大きく誇張されやすい点を考慮しておく必要がある。

その他のクラスタリング手法)

- ISODATA
- C-MEANS 法
- Minimal Spannning Tree

計算結果が樹上図として表示されるので、単なるクラスター情報のみならず、各パターン間の相対的な関係をグラフィカルに捕らえることが出来る。

④マッピング

マッピング手法はN次元（＞三次元）上での各パターン間の相互位置関係情報を保ちつつ、より少ない次元での位置情報へと変換したうえで、人間が認識可能な二次元あるいは三次元図として提供するものである。ここではパターン認識分野で広く利用されている非線形写像法（NLM : Non-Linear Mapping)を例に取り簡単に説明する。

非線形写像法はＮ次元空間上での各パターン間の相互位置関係をそのまま保ちつつ、人間が認識可能な二次元平面上におけるパターン間の相互位置関係へと置き換える（次元減少する）手法である。従って、二次元図を見ることでN次元における各パターンの相互距離関係を知ることが出来る。

計算はＮ次元空間上での各パターン間の相互距離と二次元上での相互距離との差をエラー関数とし、このエラー関数を最小化するように行われる。

$$E_{ij} = |\ (\text{N次元上でのiおよびjパターン間の距離}) - (\text{N次元上でのiおよびjパターン間の距離})\ |$$

$$E = \sum_{i}\sum_{j}\ |\ \sum_{k}(D_{ki} - D_{kj})^2$$

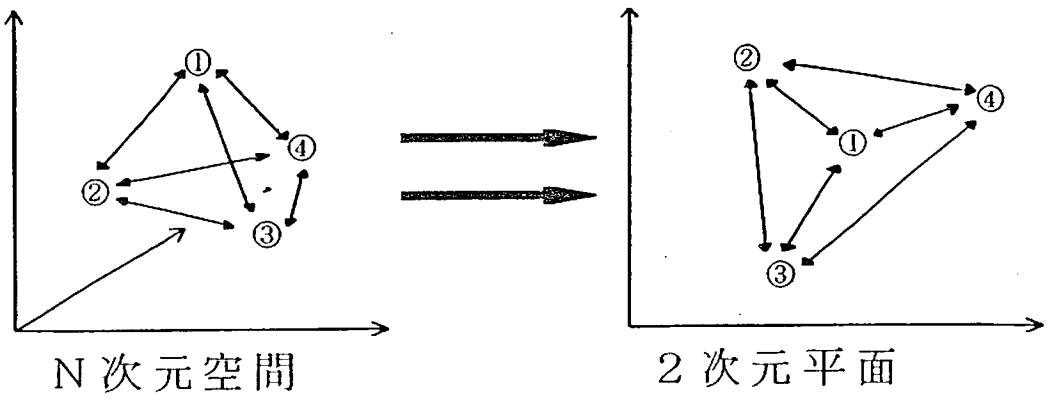


図　．非線形写像法による次元減少

得られる情報）

N次元パターン空間上の各パターン間の相互位置関係に関する情報。この情報によりパターン群のグループ化と位置関係、各パターン間の位置関係、用いたパラメータ群の善し悪し等に関する情報が見えてくる。

留意事項）

非線形写像法では、誤差計算の成否とその結果のチャート図が二次元座標計算に用いた初期座標データにより大きく変化することを意識することが必要である。また、最終的なエラーチャートがどの程度あるかも重要なチェックポイントとなる。

その他のマッピング手法）主成分分析：次元減少機能の利用

細かく分類するならば、パターン間の相互位置関係を表示する手法は大きく二種類に分類される。一つはマッピングであり、この詳細については述べてある。もう一つがプロジェクション（投影）と呼ばれるもので、N次元空間上のパターンを同じN次元上ではあるが異なる視点から眺めようとするものである。

この代表的手法としては多変量解析分野で頻繁に利用されている主成分分析がある。本著ではプロジェクション手法もマッピング手法に含めて説明する。

主成分分析により次元変換された新たな次元の各軸は主成分と呼ばれる。この次元変換過程ではサンプル群の分散が大きい軸（次元）から順に第一主成分、第二主成分のように並べ変えられ、この時全データの分散量に対する各主成分の寄与の割合が寄与率と呼ばれる。この寄与率を全主成分について求め、その総和を取ったものが累積寄与率と呼ばれる。　詳細は専門書を参照されたい。

この主成分分析が構造一活性相関に利用される時は要因解析手法としての目的の他に、次元減少という目的で利用されることが多い。これは、構造一活性相関で頻繁に利用される線形・非線形重回帰や判別分析手法が統計上の理由（偶然性の回避）から使用パラメータ数に制限があるからである。パラメータ

数が多い時、手っ取り早い次元減少手法の一つとしてこの主成分分析が利用される。なお、この次元減少手法として最近はPLS(Partial Least Squares)法が利用されることが多い。このPLSについては3-D QSARの節で説明する。なお、解析上での主成分分析とPLSの最大の違いは、主成分分析に用いられるパラメータは互いに独立性が高いことが前提であるが、PLSでは相関の高いパラメータでも適用が可能という点である。現にPLSは、比較的相関の高いデータを扱うことの多いスペクトル解析分野での利用例が多い。

⑤グラフ解析法

グラフ解析法としていろいろなアプローチがなされているが、最も広範囲に利用されているのはレーダーチャートであろう。これは別名クモの巣チャートとも呼ばれる。これらのグラフ解析法は各パターン単位で表示される。その表示された図を眺めることで、各パターン間に存在する一般的傾向や特殊要因等を人間が取り出す。このようにグラフ解析法では前述までの解析手法群と異なり、人間にパターン認識させる事を最大の特徴とするアプローチである。

図　に六次元データで示される三個のパターンのレーダーチャートの例が示されている。各パターンの次元データが相対的な大きさで表示されているために各パターン間の傾向が容易に読みとれる。

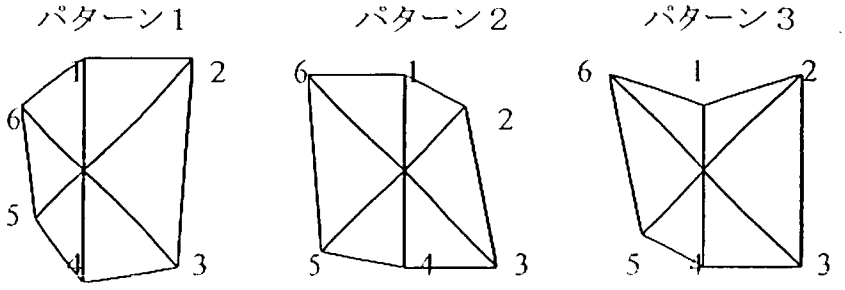


図　. 六次元データを用いたレーダーチャート

得られる情報)

このレーダーチャートからは個々のパターンのパラメータ情報が視覚的に得られる。この情報と解析目的である薬理活性との相対的な関係は人間が認識しながら捕らえることが必要となる。

その他のグラフ解析法)

・チャーノフの顔チャート
・三角多項式グラフ

1.4.3 データの種類と解析手法との関係

パターン認識法（Hansch-Fujita法、3-D QSARも含む）による解析を行う時には二種類の数値データが必要である。一つは解析目標となる‘目的変数’で

薬理活性データが該当し、判別関数や回帰式の左辺となる。もう一つは薬理活性データを説明するのに用いられる‘説明変数’で、パラメータとして呼んできたものである。この目的変数と説明変数とにそれぞれ多様な形式の数値データが利用されている。ここではこの目的変数と説明変数とにわけて利用される数値データについて簡単にまとめる。

①数値データの種類

解析で利用されるデータは解析の目標となる目的変数と、解析過程で利用される説明変数がある。これらのデータはその特性からさらに細かく分類される。総ての解析手法はこれらのデータの特性に応じて適用できる解析手法に差異が出る。実際の解析ではこのようなデータ特性を意識することが必要となる。

- 目的変数（薬理活性データ）
  量的変数（連続データ）
  等級変数（順序付きクラスデータ、バイナリデータ）
  カテゴリー変数（順序のないクラスデータ、バイナリデータ）
- 説明変数（パラメータ）
  量的変数（連続データ）
  質的変数（クラスデータ、バイナリデータ等）

②数値データと解析手法との関係

解析手法は総てのデータに適用されるものではなく、データの形態や特性等の違いにより適用される解析手法は変化する。ここではクラスデータに関する解析手法との関係についてまとめる。

クラスデータには大きく、二クラス（バイナリ）データと多クラスデータがある。一般的に二クラスデータの分類には総ての分類手法が適用出来るが、多クラス分類に適用出来る解析手法は限定される。

表　．二クラス分類および多クラス分類への適用マップ

解析手法	二クラス分類	多クラス分類
線形学習機械法	●	×
判別分析（最小アルゴリズム）	●	×
判別分析（シンプレックスアルゴリズム）	●	×
ＢＡＹＥＳ線形判別分析	●	×
ＢＡＹＥＳ非線形判別分析	●	×
ＡＬＳ法	●	●
最近隣法（Ｋ－ＮＮ法）	●	●
ＳＩＭＣＡ法	●	●
ニューラルネットワーク	●	●

③データのパターン分布特性と解析手法との関係

二クラス分類においてはデータセットが二分割可能か否かで適用される解析手法が変化する。基本的にデータセットが二分割可能な時は総ての判別分析手法が適用可能である。しかし、二分割が不可能な場合は解析手法の特性により
適、不適の問題が生じる。

例えば、二分割不可能なデータセットに線形学習機械法を適用したならば、最終的に得られる判別関数は信頼性の低いものとなる。これは、線形学習機械法で得られる判別関数は学習過程の後半部分で修正されたパターンに大きな影響（場所決め）を受けるからである。

表　．データセットの特性と解析手法

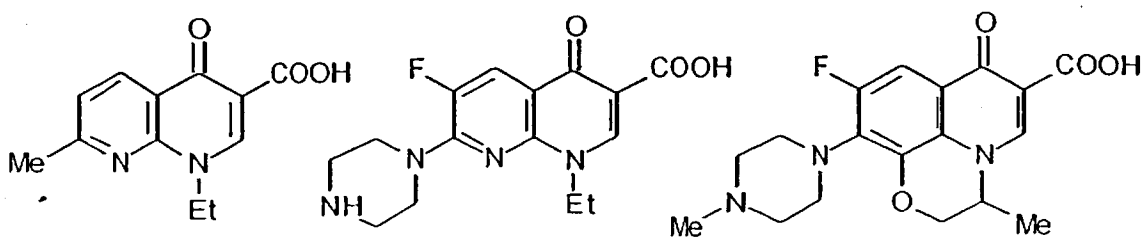
解析手法	線形二分割可能	線形二分割不可
線形学習機械法	●	×
判別分析（最小アルゴリズム）	●	○
判別分析（シンプレックスアルゴリズム）	●	○
ＢＡＹＥＳ線形判別分析	●	○
ＢＡＹＥＳ非線形判別分析	●	○
ＡＬＳ法	●	○
最近隣法（Ｋ－ＮＮ法）	●	●
ＳＩＭＣＡ法	●	●
ニューラルネットワーク	●	●

2. パターン認識法による構造一活性相関解析事例（1）

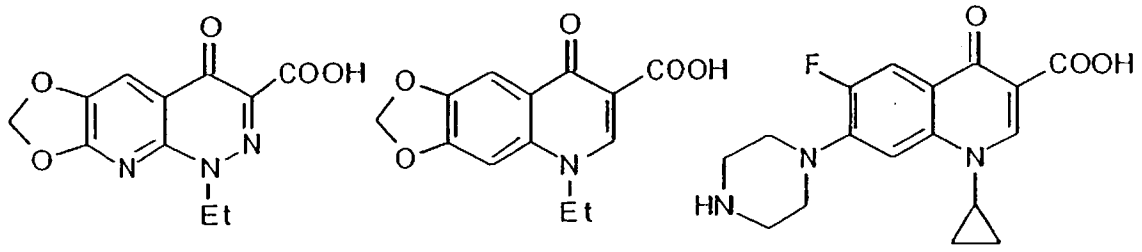
パターン認識法による構造一活性相関の特徴は前節で述べたように、解析の自由度が高く、数多くのバリエーションを生み出す事が可能となる点である。つまり、用いる薬理データ、パラメータの種類、解析手法等において様々な組み合わせが考えられる。ここでは著者が行ったパターン認識法による構造一活性相関の典型的な事例を示す。

2.1 ジヒドロキノロンカルボン酸系化合物の抗菌活性概論

ジヒドロキノロンカルボン酸誘導体、はグラム陰性菌にも有効で広範囲な抗菌スペクトルを持つ抗菌活性化合物として注目を浴びている化合物である。



Nalidixic acid　　Norfloxacin　　Ofloxacin



Cinoxacin　　Oxolinic acid　　Ciprofloxacin

1962年にNalidixic acidに抗菌活性がある事が報告され[*1)]、以下γ-pyridone-β-carboxylic acid基本骨格を有する一連の薬物のリード化合物となった。その後多数の誘導体が合成され、Norfloxacinやジヒドロキノロン骨格の1、8位間をブリッジした特有の構造を持つOfloxacin、メチレンジオキシ基を持つCinoxacinやOxolinic acid、一位置換基に三員環を持つCiprofloxacin等多数の薬物が実用化されてきた。現在も多くの製薬会社がこの系列の化合物群の改良を試みている。

＊1：Lesher, G.Y., Froelich, E.J., Gruett, M.D., Bailey, J.H., and Brundage, R.P., (1962), J.Med. Pharm. Chem., 5, 1063-1065.

2.2.1 ジヒドロキノロンカルボン酸系化合物のパターン認識による解析

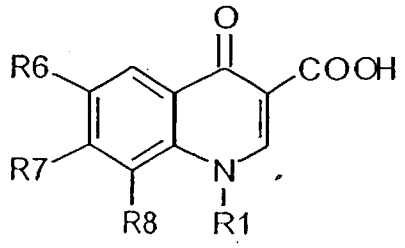
①解析目的

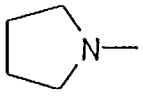
ジヒドロキノロンカルボン酸系化合物57化合物を活性および非活性の二クラスに分割する。このサンプル化合物群を活性情報に従って完全に二クラス分類出来るパラメータセットを決定し、抗菌活性を支配する（活性及び非活性化合物とを識別できる）要因の取り出しとその検討を行う。

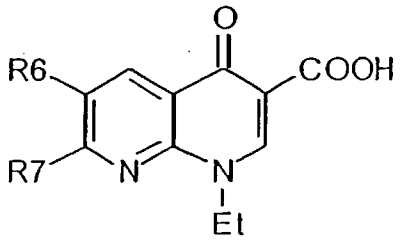
②入力データセット（化合物構造式および薬理活性）

入力データはDomagraらが1986年にJMC上に発表したデータ[*1)]を用いた。

＊1：John M. Domagra, Lori D. Hanna, Carl L. Heifetz, Marland P. Hutt, Thomas F. Mich, Joseph P. Sanchez and Marjorie Solomon, J.Med. Chem., 29, 394 (1986).



R6	R7
F	H
NO_2	H
F	
H	$H_2NCH_2CH_2NH$



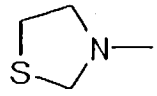
R6	R7
F	HON=CH-
NH_2	H_3C
F	

図　．解析に用いた化合物構造式例（Quinolone および 1,8-Naphthylidin e 誘導体）

Domagra らはヂヒドロキノロンおよび 1,8-ナフチリヂン骨格を有する一連の化合物群を合成し、その抗菌活性を求めて既存の薬物との活性比較／検討を行っている。この解析に用いられた化合物構造式のうち代表的なものを図　に示す。今回の解析に用いた化合物は、Domagra らにより合成された前記二種類の基本骨格を有する誘導体 43 化合物群に既存のジヒドロキノロン化合物 14 種類を加えた総数 57 化合物である。

これらの化合物群は E.coli H560 に対する最小阻止濃度 MIC の値の大小により、活性および非活性の二クラスに分割した。MIC が 0.1～6.3 μ g／mL の範囲にある化合物を活性化合物とし、12.5～100.0 μ g／mL にある化合物を非活性化合物とした。活性化合物（クラス 1）として 27 化合物、非活性化合物（クラス 2）として 30 化合物が帰属された（表　）。

表　．E.coli H560 に対する活性および非活性化合物の構成

活性化合物（クラス 1）	MIC　0.1～6.3 μ g／m L	２７化合物
非活性化合物（クラス 2）	MIC　12.5～100.0 μ g／m L	３０化合物
	総計	５７化合物

サンプル化合物群のクラス分類は統計的な観点を考慮しながら以下の二点に留意しながら行われた。①活性および非活性クラスとでサンプル数に大きな差異がない事。②活性および非活性グループ間にある程度の活性ギャップがある事。この二点を満足するサンプルセットとして前記二クラスが設定された。

解析に用いられた全化合物の構造式（三次元に正規化済み）を図　に示す。図中点線部分は芳香族結合を示す。化合物名が固有名詞で書かれているものが

市販の化合物群で、JMC-という接頭語が付いている化合物は Domagra らにより合成された化合物群である。化合物群は活性の低い順に右方向および下に行くに従って活性が高くなる。従って、最下段の右端にある Ofloxacin が今回用いた化合物中最も活性が高い。化合物は単環性のピリドン骨格を持つ化合物群から 7 位の置換基が無環のものから二環性のものまで広範囲に構造式が変化している。なお、非活性クラスは左上 637 JMC 7A の化合物から第 4 行 3 カラム目の 647 JMC 2H までの 30 化合物、続く 620 JMC 1T 移行の化合物群は活性クラスに帰属する。

図　. 全化合物構造式

③最終パラメータセットと判別関数

前記三次元構造式を基本として創出されたパラメータ群（83 種類）を初期パラメータセットとし、種々の特徴抽出を経て最終的に 5 個のパラメータが選択された。これら 5 個のパラメータが持つ情報の内容と判別関数の係数と符号を表　に示す。なお、この特徴抽出過程で一つの化合物がサンプル化合物群（非活性クラス）から取り除かれ、全体で 56 化合物が解析対象となった。なお、用いたパラメータの値は桁数や平均等が大きく異なるが、解析前に平均 0、標準偏差 1 に正規化されている。

各パラメータの符号を見た時、＋の符号を持つパラメータは活性クラスに寄与し、－の符号を持つパラメータは非活性クラスに寄与する。従って、1、3、および 5 番目のパラメータの値が大きい化合物は活性クラスに、また 2 および 4 のパラメータが大きい化合物は非活性クラスに分類される。従って、活性化合物を設計する時は 1、3、5 のパラメータ値が大きくなり、同時に 2 と 3 のパラメータ値が小さくなるような化合物を合成すると良い。

表　．最終パラメータセット

	符号	係数	パラメータ内容
1.	＋	0.1876	パスの総数／原子数
2.	－	0.6912	窒素原子数
3.	＋	0.3183	基本リング数
4.	－	0.3004	最小求核スーパーデローカライザビリティ
5.	＋	0.5428	トーション歪みエネルギー
6.	＋	0.0302	定数項

最終解析化合物データセット

活性クラス	27 化合物
非活性クラス	29 化合物
総計	56 化合物

最小求核スーパーデローカライザビリテイはフロンティア電子密度に関連するパラメータで通常はこの値が大きい所で反応が起こる。この場合は最小値なので最も求核反応の起こりにくい場所の値を意味する。パスの総数を原子数で割った値はイメージ的に構造と結び付きにくいが、簡単には化合物のサイズにある程度比例すると考えても間違いはない。その他の項目はそのまま素直に解釈出来るはずである。

④種々の解析の実施

最終的に選択された五個のパラメータを用いて種々のパターン認識法を実行する。

・クラス分類率の算出

種々の二クラス分類手法を用いてその分類率を求める。表　に求められた分類率の結果を示す。

表　．様々な二クラス分類手法による分類率

分類手法名	分類率 クラス1（活性）	クラス2（非活性）	全　体
線形学習機械法	100.0%	100.0%	100.0%
線形判別分析[1)]	100.0%	100.0%	100.0%
線形判別分析[2)]	100.0%	100.0%	100.0%
BAYES 線形判別分析	96.3%	93.1%	94.6%
BAYES 線形判別分析	88.9%	96.6%	92.9%
最近隣法（K=1）	85.2%	86.2%	85.7%
最近隣法（K=3）	81.5%	79.3%	80.4%

＊1）最小二乗アルゴリズムによる判別分析

＊2）シンプレックスアルゴリズムによる判別分析

表から分かるように同じ二クラス分類手法であっても分類率に差異がある事がわかる。この場合、線形学習機械法からシンプレックスアルゴリズムによる判別分析までは 100%の分類率を示し、その他の分類手法においても高い分類率を達成している。これからも選択された五個のパラメータは今回の 56 化合物の分類に重要な働きをする事がわかる。なお、最近隣法を除く二クラス分類から得られた判別関数の各パラメータの符号判別関数総て一致しており、係数の値にも大きな差異は見られない。

・グラフ解析（レーダーチャート）

五個のパラメータを用いたレーダーチャートを図　に示す。図中五個のパラメータは各パターンの中点からX軸方向の線を1番目のパラメータとし、以下反時計回りに2～5の順で各パラメータの値の大きさに比例して描かれている。

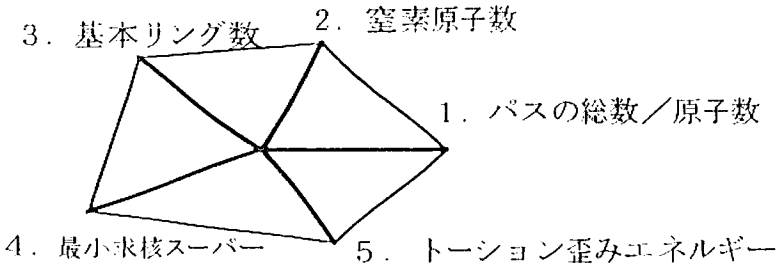


図　．レーダーチャートと各パラメータとの関係

図中上位の三行は活性化合物群（クラス1）を示している。図を眺めると上位三行と、下位四行とでレーダーチャートの図の傾向が異なっている事が分かる。上位三行の活性化合物のレーダーチャートは全体的に左上から右下に延びた構造を持ち、一方下位の非活性化合物のレーダーチャートは反対に右上から左下の方向に伸びた構造をしている事がわかる。特に活性化合物群で個の方向性が強い DAN640 は Ofloxacin であり、DAN606 の化合物は Ciplofloxacin である。これは先の判別関数の係数を考えると理解出来る。つまり、判別関数では 1、3 および 5 の係数

が正で、2および4の係数が負である。図　からも分かるように活性化合物は1、3、5が大きくなるので左上から右下に伸びた形特徴抽出なり、一方で非活性化合物は2と4の値が相対的に大きくなるために右上から左下に伸びた構造となる。このように、判別関数から得られる情報とレーダーチャートから得られる情報とでは同質の情報であるがその表現形態が異なっている事がわかる。

なお、非活性化合物（クラス2）でDAN637はその形が他の化合物群と比較して大きく異なるが、この化合物は単環のピリドン誘導体で構造的にも他の化合物群とは大きく異なっている。

図　．全化合物のレーダーチャート

・マッピングによる解析（主成分分析）

図　に主成分分析における第一主成分（X軸）、および第二主成分（Y軸）を用いたバイプロット図を示す。図の中央から引き出し線が出ていて数字番号が四角の中に書かれているが、個の番号はパラメータのID番号である。46が5番目のパラメータ、18が2番、25が3番、4が1番、62が5番目のパラメータに該当する。この引き出し線の長さと方向は各パラメータの第一主成分、および第二主成分に対する寄与の程度を表している。この図から分かることは、

・第一主成分と第二主成分とで80.99%の変量が説明されている、

・パラメータの1、3、5は殆ど第一主成分のみに寄与しており、2と5のパラ

メータが第二主成分に寄与する、

・活性化合物（クラス1）が画面左部分にコンパクトにまとまり、非活性化合物（クラス2）は画面右半分に左下から右上方向に分散して分布する、

等の事実である。これらに情報を基本としてさらに幾つかの情報を得る。

◎バイプロット図の考察（全体概要）

まず、第一主成分の持つ情報であるがこの場合、化合物の分散を考えるならば活性および非活性化合物群が第一主成分上で比較的綺麗に分類出来る事がわかる。またこの図から、第一主成分を支配するパラメータは1、3および5の三種類であり、これらのパラメータは左に行くほど大きな値を取り、第二主成分には殆ど寄与しない事がわかる。これら三個のパラメータは、判別関数で総て正の符号を持つパラメータである。先の判別関数では活性化合物に寄与するパラメータと判定したが、このバイプロット図により非活性化合物の分散にも重要な働きをしている事がわかる。

第二主成分に関しては化合物の分散が活性化合物群に対して大きくなく、一方で非活性化合物群は上下に大きく分散している。この事実から第二主成分はクラス2の化合物群の差異説明に重要な要素となる。この第二主成分には窒素

の数と求核スーパーデローカライザビリテイが寄与しており、窒素数は左側が大きく、求核スーパーデローカライザビリテイはその反対である。

◎バイプロット図の考察（各サンプルレベル）

各化合物の構造レベルに戻って個のチャートを見るならば、基本リング数の大小に従って単環化合物（図中最右端：DAN637のピリドン誘導体）から、二環、三環そして四環性化合物（図中左端側4化合物：DAN640 Ofloxacin およびDAN606 Ciplofloxacin を含む）へと並んでいる事がわかる。

従って、化合物群を活性／非活性に分類する重要なパラメータは1、3 および5の三種類である事がわかる。これらのパラメータは判別関数で総て正の符号を持つパラメータである。このバイプロット図で明確になった情報は基本リングの数に従って第一主成分上に化合物が並んでいる事である。キノロンおよび1,8-ナフチリヂン誘導体で活性がある化合物は極一部を除き、環数が3以上あることがわかる。

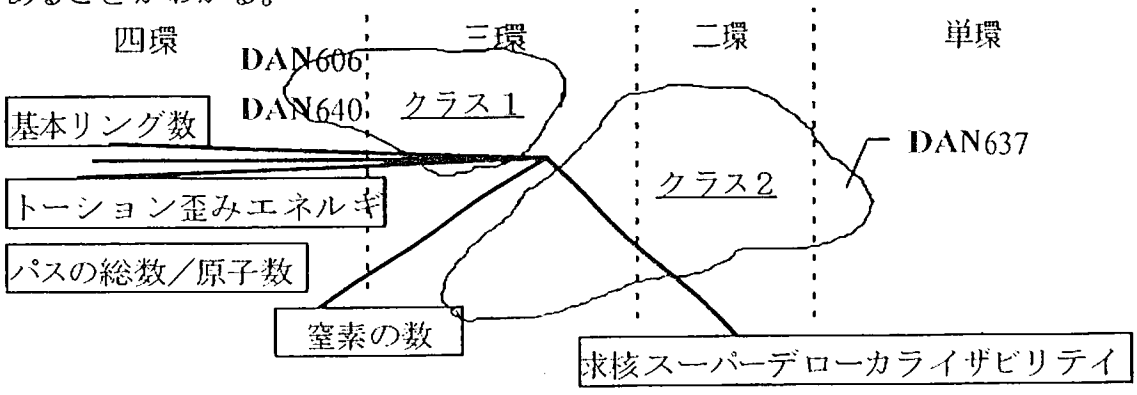


図　．バイプロット図の概略図

このバイプロット図を化合物の分散状況に応じてより小さなグループに分割したものが図　に示される。化合物群は第一主成分の右の方からグループ1～5へと分布している。グループ1、2、5はクラスが同じ化合物だけであるが、グループ3と4は他クラスの化合物が若干数含まれている。

グループ1と2にまたがる領域④にはピリドン基本骨格を持つ三化合物がある。グループ3に属する領域①、②、③はそれぞれ異なった特徴を持つ化合物群が帰属する。キノロンおよび1,8-ナフチリヂン誘導体の6位と7位の置換基に注目するならば、領域①は一原子置換基を持つ化合物が帰属している。領域①と②の間にはOxolinic acid やMiloxacin等のメチレンヂオキシ基を持つ化合物が分散する。領域②はピペラヂンを置換基として持ち、領域③は 1,6,8-Naphchilidine基本骨格を持つ化合物群が帰属している。また、グループ2の領域④以外の部分にはNalidixic acid が帰属し、その他に6、7位置換基としてニトロおよびニトロソ基を含む化合物がある。

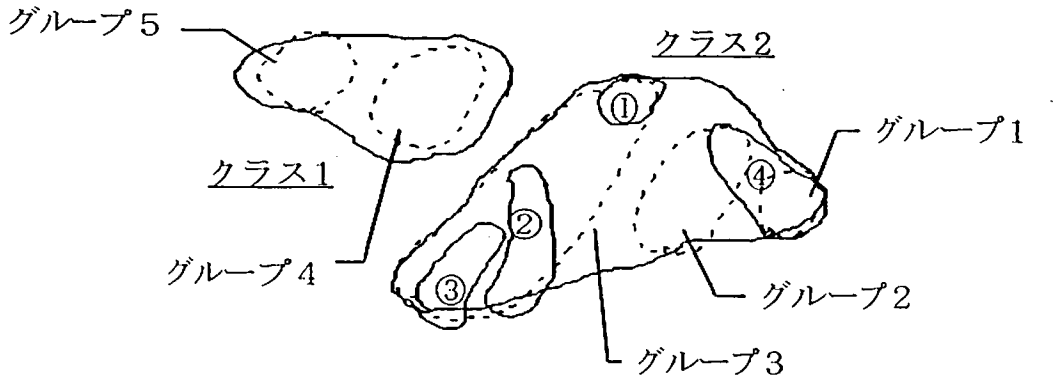


図　.グループや領域に分けられたバイプロット図

2.2.2　化合物データベースを用いた“リード候補化合物検索”の実施

パターン認識法による構造ー活性相関の最大の特徴は、対象とする化合物群に対する構造上の制限が他の解析法と比較して極端に小さい事である。この特徴を生かすことで広範な化合物群を対象としたプレスクリーニングを実施する事が出来る。この機能は、最近急速に展開されつつあるコンビナトリアルケミストリ／HTSにも利用出来る。このコンビナトリアルケミストリ／HTSでは多数で構造式の異なる化合物群を短時間で合成する事が要求されるが、この要求を満たす事は困難である。そこで、実際に化合物を合成する前にプレスクリーニングが出来れば、リード候補化合物の発見効率を高めると同時に合成の手間も軽くする事が可能である。この意味でパターン認識法が持つ多種多様な化合物群を対象としたスクリーニング機能は非常に重要なものとなりつつある。

①“リード候補化合物検索”概要

このアプローチはリード候補化合物群を既存の化合物データベースから検索（取り出す）するアプローチである。検索する時の手法はパターン認識法で得られた判別関数を用いて行う。このアプローチの特徴は検索対象とする化合物の構造式が多岐にわたり、構造変化量が極めて大きいことである。前にも述べたように構造ー活性相関手法は解析対象を限定する事でその解析精度を高くしてきた。従って、構造式の変化の大きい化合物群を対象とした解析は物理的に実行不可能となる。現時点で、この特徴に耐えるアプローチはパターン認識法を除いて存在しない。

②“リード候補化合物検索”による新規抗菌活性リード候補化合物の選択[*1)]

＊１：湯田　浩太郎、“ＬＥＡＤ化合物検索概念及びシステム（１）”、第１４回構造ー活性相関シンポジウム講演要旨集、Ｐ.２９４、名古屋、１９８６年.

実験の目的：抗菌活性リード候補（既知）化合物ライブラリを構築する。

解析の流れ：化合物データベースから無作為に２００化合物を取り出す。こ

の取り出された化合物群について抗菌活性を予測し、抗菌活性有りと予測された化合物をライブラリに加える。

実験手続き：・キノロンカルボン酸化合物の抗菌活性データを用いて、ADAPT
システム上で判別関数を求めておく。
・化合物データベース（Aldrich）から無作為に 200 化合物の取り
出し、ADAPT システムに入力する。
・判別関数で利用されたパラメータと同じ種類のパラメータ群を、取り出された 200 化合物について創出する。
・予め用意された判別関数と、・で創出されたパラメータを用いて、取り出された化合物群の抗菌活性を予測する。活性有りとされた化合物群を抗菌活性リード候補化合物ライブラリとする。

リード候補化合物検索を実行する時の流れを図 に示す。

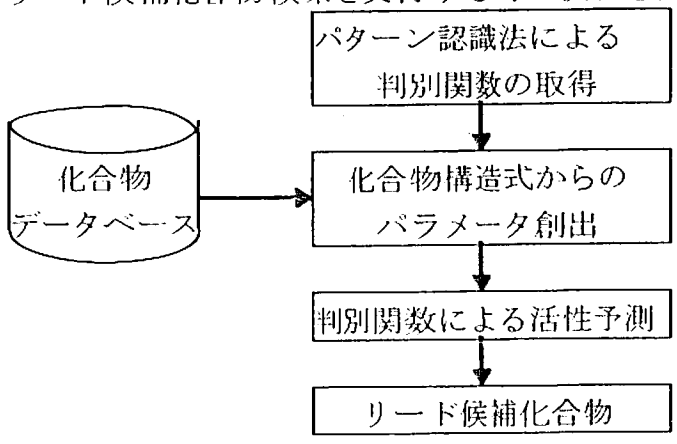


図 . リード候補化合物作業の流れ図

最初に目標とする薬理活性に関する判別関数を求める。このリード候補化合物検索で利用される判別関数を構成するパラメータは解析の特性上、化合物構造式に依存する事の少ないパラメータを用いる事が必要である。例えば各原子の数、分子容積、様々なトポロジカル関連パラメータ等であり、部分構造パラメータのような特別な部分構造と強く相関するパラメータは用いない。このように化合物構造式と直結しないパラメータを用いる事で、より構造に変化のある化合物群を取り出す事が可能となる。

このリード候補化合物検索で利用される判別関数を構成するパラメータは、解析の特性上化合物構造依存度の少ないパラメータを用いる事が必要である。例えば各原子の数、分子容積、様々なトポロジカル関連パラメータ等であり、部分構造パラメータのような特別な部分構造と強く相関するパラメータは用いない。このように化合物構造式と直結しないパラメータを用いる事で、より構造に変化のある化合物群を取り出す事が可能となる。

・化合物データベース

今回用いた化合物データベースは Aldrich のデータベースを用いて、200 化合物をランダムに取り出した。この取り出された化合物を図　に示す。図からも分かるように取り出された化合物は、単純な直鎖状の炭化水素化合物からステロイド系、芳香族多環縮合化合物やポリオキシン系の化合物群まで多種多様な化合物となっていることがわかる。

図　．Aldrich 化合物データベースより取り出された化合物構造式

・判別関数

リード候補化合物の選択に利用する判別関数は前節で述べたキノロンおよび 1,8-ナフチリヂン誘導体の抗菌活性解析で取り出された物を用いる。この判別関数は当初から意識的に構造と直結するパラメータ群は除いてあり、今回のリード候補化合物検索にもそのまま利用する事が出来る。

③リード候補化合物検索の実行結果と考察

判別関数による選択により最終的に取り出された化合物群を図　に示す。

図　．抗菌活性有りとして取り出されたリード候補化合物群

総計 26 化合物が取り出された。当初の 200 化合物が約一割強の化合物へと減少したことになる。しかも、これらの化合物群は総て判別関数により活性有りと判断されている。このように既存の化合物群から効率良く、高速にリード候補化合物群を取り出す事が出来るのは非常に有用な事である。

これら 26 化合物をみると、判別関数の取得に用いたキノロンおよび 1,8-ナフチリヂン誘導体とは基本構造式が全く異なる化合物が取り出されており、新規のリード候補化合物群を取り出すという目的は十分達成されている。構造式を詳しく見るとさらに化合物群は 3〜4 グループに限定されており、幾つかの傾向を読みとることが出来る。以下にこれらの特徴を簡単にまとめる。

- ステロイド誘導体（七化合物）が比較的多数含まれている。
- 芳香族スルホンヂアゾ化合物群（四化合物）も多数含まれる。
- トリフェニル系色素（七化合物）が多い。
- その他（ポリオキシン化合物、芳香族縮合多環化合物、他）

ステロイド基本骨格を持つ抗菌性化合物は既に多数報告されているし、スルホンヂアゾ化合物はサルファ剤の一種と考えられる。トリフェニル系色素は構造的に D.D.T 等の薬物を思い出させる。その他、ポリオキシン系化合物にも抗菌活性を持つ化合物がある。

④リード候補化合物検索のまとめ

判別関数の創出に用いた解析母集団化合物とは構造的に全く異なる化合物群が取り出されている。さらにこれらの構造式を検討すると既存の抗菌活性剤

と同じ、あるいは似た基本骨格を有している。Aldrich 化合物データベースから取り出された化合物が多種多様な構造式を有していた事から考えるならば、かなり構造的に絞られ、かつ後付けではあるが抗菌活性化合物に関係の深い化合物群が選択されている事がわかる。

これらの化合物群がなぜ選択されたかの具体的な意味付けは難しい。ただ言えることは、判別関数の取り出しに用いた化合物群と、その判別関数より取り出された化合物群とは同じ基準をクリアしていることである。人間が構造の全く異なる化合物の構造式を見て、それらが同じ基準をクリアするか否かを判定する事は不可能である。少なくともコンピュータはその仕事を高速で成し遂げることは確かである。今回は一つの薬理活性に関する判別関数しか利用していないが、実際は多数の判別関数を用意することで複数の薬理活性に関するリード候補化合物の取り出しが可能である。さらに、毒性や副作用等の同時チェックも可能である。

⑤判別関数によるリード候補化合物取り出しの簡単な検証実験

判別関数を用いたリード候補化合物検索について簡単な検証を行っているので、以下に簡単にまとめる。

活性予測化合物群として抗菌活性を有する九個の化合物を用いた（図　）。これらの化合物群は意識的に構造が大きく異なる化合物群を取り出している。これら九個の化合物群について、前項で用いた判別関数を使って抗菌活性を予測した。予測結果は、全九化合物中活性有りと判断された化合物は六化合物（511、512、513、514、516、517）で、残る三化合物（510、515、518）が活性なしとされた。

正答率は67%であり、これが高いか低いかは判断する人により異なるはずである。少なくとも、構造式が全く異なる化合物から得られた判別関数を用いて、多種多様の構造を持つ化合物群から過半数の確率で活性化合物群を取り出した事は事実である。このように、構造の全く異なる化合物群から同じ基準を満たす化合物群を高速に取り出す事は人間には不可能な作業である。

これだけ構造式が異なれば薬理学的な考察や、構造一活性相関を議論することは困難である。ただ、今回取り出された化合物群はキノロンおよび1,8-ナフチリヂン誘導体を抗菌活性に従って100%完全に分類するのに用いられたのと同じ情報を用いて活性と判断されている。つまり、今回活性ありと判断された化合物群は、キノロンおよび1,8-ナフチリヂン誘導体の抗菌活性発現要因と同種の情報を共有していることは事実である。この共通要因と活性メカニズムとの相関を求める事が必要であるが、この作業には分野を越えた多くの研究者の強力が必要である。現時点で答えは出ていない。

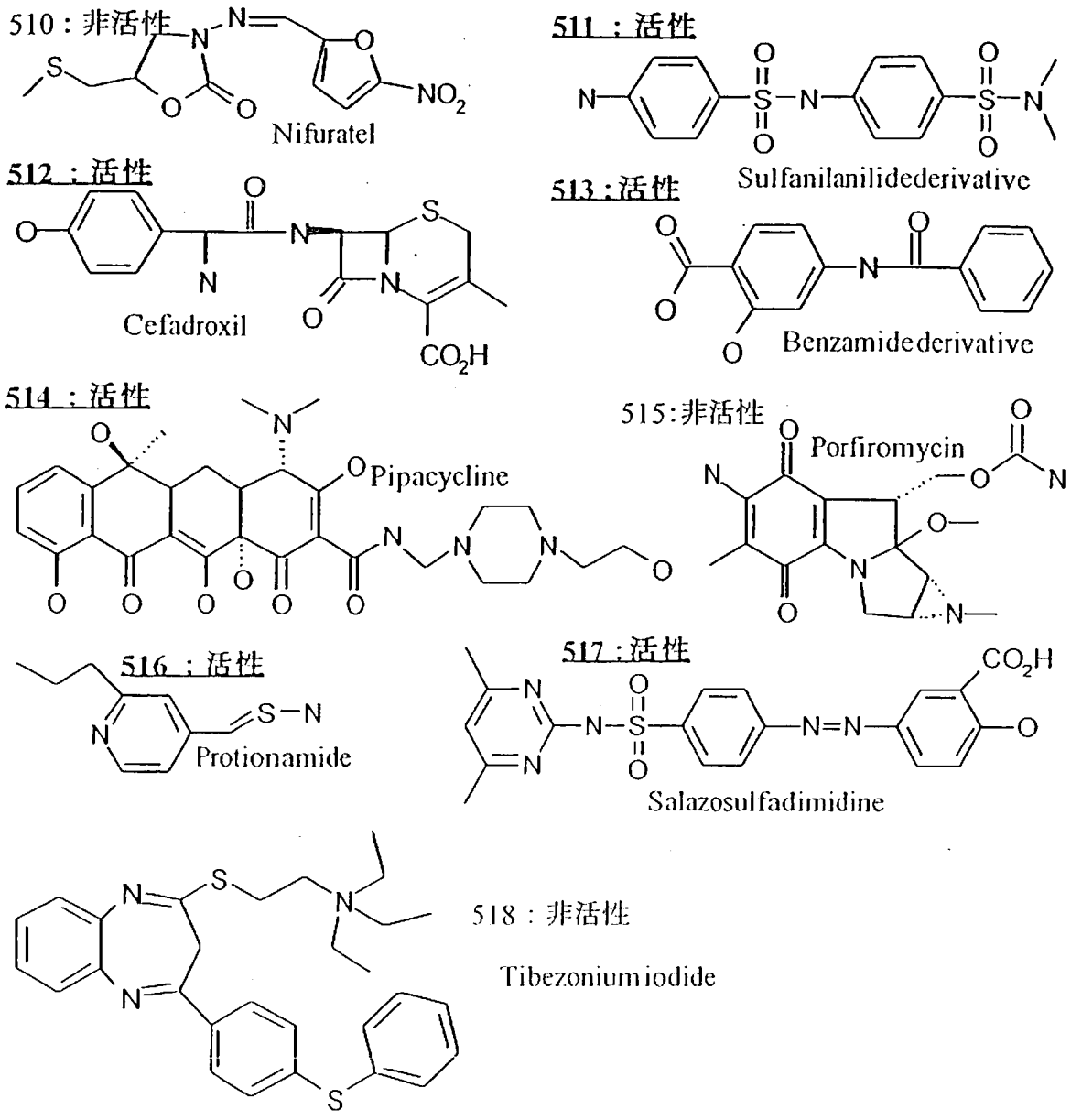


図　．既知抗菌活性化合物を用いた判別関数による活性予測テスト

3. パターン認識法による構造一活性相関実施上での統計的観点からの留意点

3.1　サンプル数とパラメータ数との関係

①解析における“偶然相関（Chance Correlation）”の問題

大量のサンプルとパラメータを扱う解析手法において常に意識するべき問題として“偶然による解析成功”の問題がある。この問題は解析の妨げとなるものであり、解析過程ではこの問題を回避する事が必要である。この問題は、Hansch-Fujita 法で用いられる線形重回帰、非線形重回帰においてパラメータの数を限りなく増やせば最終的に相関係数は 1 になること。また、判別分析においてもパラメータ数を増やせばいつかの時点で必ず 100%分類が実現できると

いう現象として実現される。相関係数が１、および分類率が100％は解析の最終目的ではあるが、パラメータ数を単純に増大して得られた結果は単なる偶然による成功でしかなく、解析の信頼性は全くない。

実際に解析を行う時には、信頼性の高い解析結果であるか否かを判定するための分かりやすい基準が必要である。この基準として一般的には解析に用いるサンプル数をその解析に用いたパラメータ数で割った値が利用される。パターン認識法による構造ー活性相関で最も頻繁に利用される判別分析ではこの値が４以上あれば偶然による問題を回避出来るとされている。この４という値はJursらのChance Correlationに関する研究[*1]により設定された。

＊１）Jurs. P.C.,　　J.C.I.C.S., ,　(197).

②判別分析（二クラス分類）に対するChance Correlationの研究

判別分析でも二クラス分類（例えば、活性グループおよび非活性グループの二群）は構造ー活性相関では利用されることの多い解析である。この二クラス分類についてJursらによりChance Correlationの研究が行われているので簡単に述べる。

二クラス分類の時、二クラスに分類される場合の数は一次元（一パラメータ）ごとに二種類発生する。従って、パラメータがＫ個（次元）における分類の場合の数Ｐは　$P=2^K$　で示される。一方、サンプル数がＮの時、これらのサンプルが二クラスに分類される場合の数Ｃは以下のようになる。

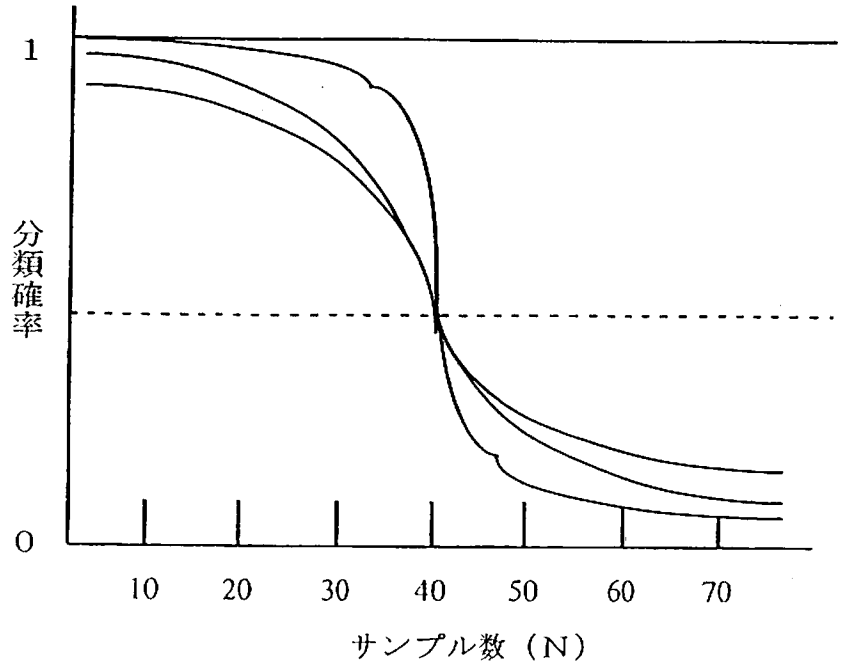
$$C=\frac{1}{2}\sum_{k=1}^{n}\frac{n!}{k\times(n-k)!}$$

従って、偶然により二クラス分類が成功する確率Ｐは前記二種類の計算式より以下のようにして求められる。

二クラス分類の成功確率　$P=\frac{R}{C}\times 100$

Ｒ：パラメータ数により分類可能な場合の数

Ｃ：サンプル数により実際に分類される場合の数



3.2 パラメータの桁数、平均値等の変動に関する問題

①パラメータの正規化

パターン認識法で利用されるパラメータは多種多様であるため、一般的にはその値の桁数やデータの正規性や等分散性をあらかじめ揃え定量的構造一活性相関おくことは困難である。このように桁数が異なるパラメータ群を混在させつつ解析を行うと解析結果として得られる判別関数や重回帰から得られた回帰式の係数の桁数にバラツキが見られる。また、極端に桁数の異なるパラメータを混在させたデータを用いた解析が、その解析結果に与える影響を予測することは困難である。このためにパターン認識による解析に先だって、利用する全パラメータを平均0、標準偏差1に揃える正規化(Normalization)がおこなわれる。

なお、この正規化はオートスケーリング（Autoscaling)とも呼ばれる。

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

$$W_k' = \frac{W_k - \overline{W}}{Q}$$

ここで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の値のうち、k番目のW_kの値、$\overline{W}$は用いるパラメータの平均値である。

②正規化されたパラメータを扱う時の留意点

信頼性の高い結果を得るべくパラメータを正規化した時は解析上いくつかの留意点が生じるので説明する。注意すべきポイントは解析で得られた判別関数や回帰式の取り扱いに関するもので、以下のようになる。

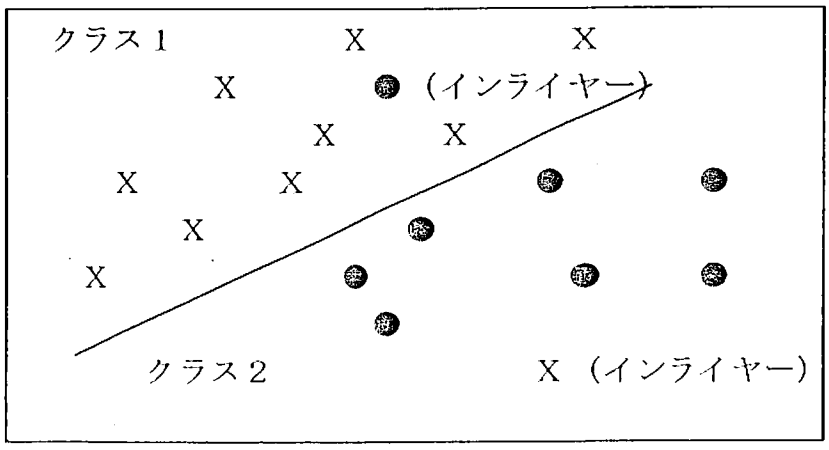
・判別関数や回帰式の係数同士の比較

解析結果得られた判別関数や回帰式の係数同士の大小による直接比較は行わない。係数同士の比較は、符号（＋、－）だけの比較にとどめる。

3.3 解析結果の信頼性を示す／向上させるための種々の指標

①取り出すサンプル（インライヤー）数の限界

解析過程でノイズの原因となるサンプルを取り出すことは非常に大事な作業となる。この効果の一つはノイズサンプルを取り出すことで解析の信頼度を高める事が出来ることがある。もう一つはノイズサンプルを明確にすることで要因解析に重要な情報（なぜそのサンプルが他と異なる特性を示すのか等）を提供することである。



このインライヤーの限界数を理論的に求めた発表はないが、著者は Hansch-Fujita 法のアウトライヤーにならい、総サンプル数の一割を限度としている。

②クラスを構成するサンプル数に関する留意点

サンプルの総数が大きくとも、個々のクラスを構成するサンプル数に偏りがある場合、たとえ解析上の諸条件（サンプル数／パラメータ数など）を満たしていても解析結果の信頼性に問題が生じることは明白である。この点で、解析母集団を収集する時にも注意が必要となる。例えば総数 100 サンプルあったとしても、50：50 で構成される二クラスデータを用いた解析と、99：1 で構成されるデータを用いた解析とでは、たとえ同じ 100%の分類結果を得たとしても、その信頼性に大きな差異があることは容易に想像がつく。

このクラスポピュレーションに関しても既に Jurs らによりその基準が示されている。ここでは簡単に結果だけ述べる。実際の解析時にはこの件を考慮し

つつ解析母集団のクラスポピュレーションを調整する事が必要である。守るべきルールは、“各クラスのサンプル数はその解析に用いるパラメータ数と同じか、それよりも大きくなければならない”というものである。

クラスポピュレーション　≧　パラメータ数

③予測率と分類率

線形重回帰や非線形重回帰等と異なり、パターン認識で頻繁に行われる判別分析の結果の評価は予測率と分類率の二つの指標が利用される。

・予測率

活性未知サンプル群を対象とした時の正答率を表すもので、判別関数の外挿性の目やすとなる。活性未知の化合物群を用いてこの予測率を出す事は困難であるので、実際には活性既知サンプル群を用いてこの予測率を出す手法が展開されている。この再評価法としてリーブワンアウト（Leave One Out)法、リーブNアウト（Leave N Out) 法等が用いられている。なお、多変量解析分野ではブートストラップ法が良く用いられるが、その手続き過程でパターンのコピーが行われるのでパターン認識分野ではあまり利用されない。

この予測法は解析母集団からN個のサンプルをとりだし、残ったサンプル群を用いて判別関数を求めて、予め取り出されていたサンプルの活性を予測して予測率を求めるもので、この手順を全化合物が必ず一回以上はサンプルとして取り出されるように設計する。。サンプル数が大きい時は一回の予測で取り出されるサンプル数は大きくなるが、サンプル数が小さい時はN=1として行うことが多い。さらに、サンプルの組み合わせによる予測率の変動を最小限度に抑える為には様々な組み合わせで予測サンプルを取り出す事が必要であり、従って一つの化合物が複数回サンプルとして取り出される事になる。図　にリーブワンアウト法の流れ図を示す。

＊最近展開されている 3-D QSAR では薬理活性予測に線形及び非線形重回帰手法を用いている。この場合にも活性予測を行う時の信頼性を示す指標が必要である。この指標としてCoMFA法ではクロスバリデーションなる指標を用いている。この詳細については3-D QSARの節にて説明する。このクロスバリデーションは次に述べる特徴抽出（PLS用）としても利用される。

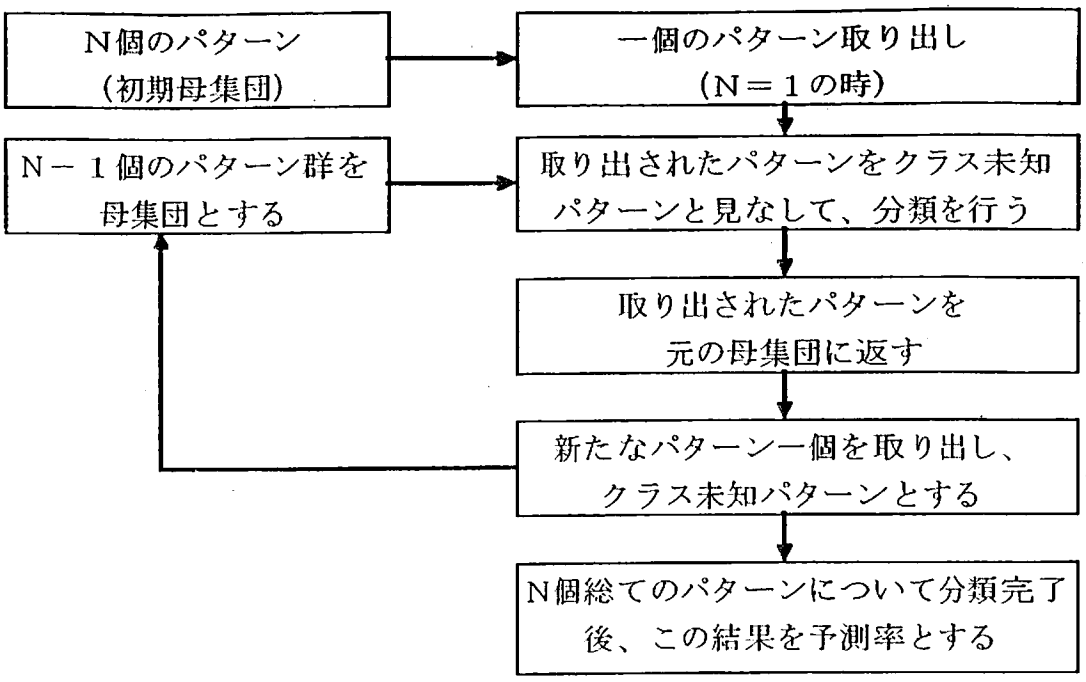


図　．リーブワンアウト法による予測率計算の流れ図

・分類率：解析母集団に対する分類率である。解析結果として直接得られる指標で、正解したパターン数を全パターン数で割って 100 を掛けた値である。

3.4 特徴抽出（解析の妨げとなるノイズパラメータの除去）

特徴抽出 (Feature Selection) 手法はパラメータ数が大きくなるにつれて重要となる。ひと昔前までは解析に用いるパラメータを揃えることが大変で、このパラメータの収集作業に多くの作業量を必要とした。このような環境下の解析ではパラメータを減少させる事よりも、収集されたパラメータを用い、その範囲内でいかに高い分類率や相関係数を得るかに多くの努力が払われた。このような解析を行っている場合、パラメータの減少を目指す特徴抽出の役割は大きくない。しかし、コンピュータの進歩により多数のパラメータを自動的、かつ簡単に創出出来るようになると特徴抽出の果たす役割は非常に重要となる。

パターン認識による構造ー活性相関では多数のパラメータを扱うので、この特徴抽出手法の果たす役割は非常に重要である。この特徴抽出機能が弱ければ解析に用いるパラメータ数を絞り込むことが出来ず、3.1 項で示した遵守項目等を満たすことが出来なくなるばかりか、解析自体も良好な結果が得られなくなり、要因解析が出来なくなる。解析に重要なパラメータを Intrinsic parameter、重要でないパラメータを Non-intrinsic parameter とよぶ。

①特徴抽出手法（パラメータ選択）

特徴抽出手法として様々な手法が存在するが、手法の基本的な特性から大きく三種類に分類する事が可能である。以下に各々の代表的な特徴抽出手法について簡単にのべる。

・パラメータを構成する個々の値や統計的特性を利用した特徴抽出

０値・同値データの出現率、フィッシャー比

これらの特徴抽出手法はパラメータ単位で、個別に善し悪しを評価できるアプローチである。例えば二クラス分類ではフィシャー比が利用される。このフィシャー比が示す情報は、各パラメータが持つクラス分類能力である。フィシャー比が小さい時、そのパラメータで二クラスのパターンを分類することは困難である。逆に、フィシャー比が大きい時は高い分類能力を示す。

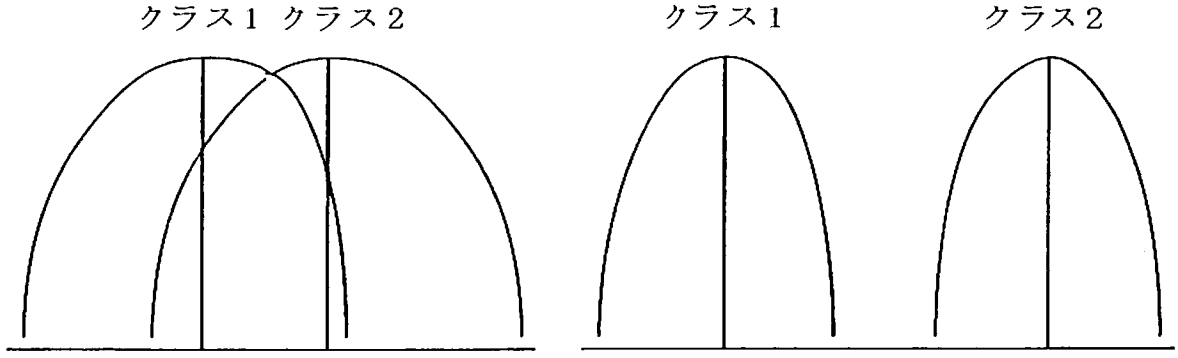


図　．フィッシャー比の値と二クラス分類に対するパラメータの関係
右図は一個のパラメータのみで100%分類が可能な例である

$$F = (\overline{X1} - \overline{X2})^2 \ / \ (\sigma 1^2 - \sigma 2^2)$$

$\overline{X1}$：パラメータのクラス１の平均値
$\overline{X2}$：パラメータのクラス２の平均値
$\sigma 1$：パラメータのクラス１の分散
$\sigma 2$：パラメータのクラス２の分散

なお、線形・非線形重回帰で用いられるＴ値やＦ値は統計量を基本とするが、本書では・の部類に分類し、Hansch-Fujita 法の節の中で説明する。

・パラメータ間の相互関係に注目した特徴抽出（相関係数による特徴抽出）

前項の特徴抽出手法は個々のパラメータ単位で評価するものであり、内容的にはパラメータの分類能力に関する評価が主体であった。この他に、パラメータ間の相互関係にも留意することが必要となる。この場合は、パラメータ同士の一対一関係を論じる場合（単相関）と、一対複数の関係を論じる（多重相関）二つのケースが存在する。多重相関は線形および非線形重回帰で利用されるこ

とが多い。ここではパターン認識や多変量解析の総てにわたって利用される単相関について述べる。

・単相関のチェック

単相関は簡単に考えるならば、パラメータ同士の類似度を判定すると言える。ふたつのパラメータが有している情報の内容（データの分散の形や程度）が似ているならばどちらか一方のパラメータがあれば十分であり、この場合残るパラメータはノイズと見なされる。

$$\text{相関係数} = \frac{\text{2個のパラメータの共分散}}{\text{パラメータ1の標準偏差} \times \text{パラメータ2の標準偏差}}$$

この相関係数が０の時、二個のパラメータ間に類似関係は全く存在しないことになる。係数が１の時は情報的にその二パラメータは全く同じとみなせる。統計量を扱う多変量解析では書くパラメータは直交関係（相関係数が０）にあることを前提として展開される。手続き上では相関係数が１のパラメータがある場合、マトリクスの扱い等で問題が生じてくる。また、パターン認識にしてもこのようなパラメータが存在する時の動きは傾向的には認識出来るが、正確に予測することは困難である。

実際に解析を行う時は相関係数が 0.9 以上のパラメータはどちらか一方を除いている。この時、二個あるパラメータのうちどちらのパラメータを取り除くかが問題となるが情報論的にはどちらのパラメータを取り除いてもかまわない。

しかし、構造ー活性相関では要因解析が非常に重要なプロセスとなる。この要因解析を考慮するならば、取り除くパラメータは要因解析に貢献する事の少ないパラメータ（1.2 の③参照）から優先的に取り除くことが必要である。

・多重相関のチェック

情報重複について考えるならばパラメータ間の１：１の関係の他に、１：多の関係も存在する。これは１個のパラメータと他の複数のパラメータ間における相関の目安である。クラス分類等ではあまりこの多重相関についてチェックする事はないが、線形重回帰手法等を適用する時は多重相関による特徴抽出を実施する。一般的にはこの多重相関が 0.95 以上ある場合はパラメータ群から取り除く。

・個々の解析手法の特徴を利用した特徴抽出

線形学習機械法（二クラス分類）：ウェイトサイン法、
バリアンスウエイト法

SIMCA 法（多クラス分類）：モデリングパワー、

ディスクリミネイティングパワー

ニューラルネットワーク（多クラス分類、フィッティング、他）：

忘却学習法*1）によるネット数の減少と特徴抽出の実施

＊1）Ichikawa et al., J.M.C.,

線形重回帰（フィッティング）：総当たり法、前進選択法、後進選択法、F値およびT値によるインタラクティブな特徴抽出

主成分分析（マッピング）：因子負荷量の利用

ここに示された様々な手法は解析手法の特性をうまく利用したアプローチである。一般的に、解析手法に特化した特徴抽出手法の能力は高いので、前記二ステップ・・を実施した後におこなう最終的な特徴抽出手法として利用されることが多い。ここでは、パターン認識法による構造一活性相関解析で最も良く利用される二クラス分類を行う線形学習機械法によるバリアンスウエイト法について簡単に述べる。

・バリアンスウエイト法*1）

本手法は二クラス分類手法である線形学習機械法の特性を利用して特徴抽出を行うもので、Jurs らにより開発された。一般的に線形学習機械法では全く同一の解析母集団を用いたとしても判別関数を算出する学習過程の条件の差異により、たとえ同じように100%分類したとしても、得られる判別関数の個々のパラメータの係数値には変動がでる。この変動量はノイズとなるパラメータの方が大きく、重要なパラメータは小さくなることをJurs らが証明した。従って、この変動量を算出し、変動量の大きなパターンから順にパラメータセットから取り除くことで簡単に特徴抽出を実行できる。

$$VW_j = \frac{V_j}{\overline{W_j}}$$

$$V_j^2 = \frac{1}{(n_k - 1)} \sum_{k=1}^{n_k} (W_{jk} - \overline{W_j})^2$$

式中jはパラメータの、kはウエイトベクトルのインデックスである。$\overline{W_j}$はj番目のウエイトベクトルの平均値、n_kは用いたウエイトベクトルの数である。

＊1）Jurs P.C. et al., J.C.I.C.S,

なお、ウエイトサイン法はバリアンスウエイト法と同様な原理を基本とするが、単にパラメータの係数の符号の変化をモニターすることで特徴抽出を行う

ものである。つまり、同一の母集団を用いているにかかわらず符号が変化するものはパラメータとしての安定性に問題がある、すなわちノイズパラメータであるとして取り出す手法である。

・その他の手続き的アプローチ（分類率を利用した特徴抽出）

以上の他にもパラメータを選択するアプローチがある。例えば、全パラメータから一つのパラメータだけを取り出した残りのパラメータセットを用いて分類を行う。この時、分類結果が悪化する時は取り出されたパラメータは重要であるが、結果に大きな差異が無いか向上する時は取り出されたパラメータはノイズパラメータとなる。この反対に、一パラメータだけを用いて分類し、その分類結果の善し悪しで判断することも可能である。しかし、一般的には一個だけのパラメータを用いたときの分類結果にパラメータ間の大きな差異は出にくいため、このアプローチはあまり行われていない。

また最近では、最適化手法として展開されている遺伝的アルゴリズムを特徴抽出手法に取り入れて効率良く特徴抽出を行う手法も試みられている。このアプローチでは遺伝子として何を選択し、その遺伝子の選択に用いる関数としてどのような関数を用いるかが大きなポイントとなる。

②パターン抽出について（パターン選択）

実際の解析では特徴抽出によるノイズパラメータ除去の他に、ノイズとなるサンプルパターンの除去が必要となる。クラス分類では間違ったクラスに飛び込んでしまうインライヤー（3.3 の①取り出すサンプル（インライヤー）数の限界参照）、また線形／非線形重回帰のフィッティング手法において線上から大きく離れて存在するアウトライヤーがノイズとして取り出される。

このパターン抽出についてのまとまった手法は存在しないが一般的には解析を繰り返す過程でこのようなパターンを絞り出すことが行われている。たとえばクラス分類（繰り返し学習によるクラス分類）では誤分類される回数の高いパターンを、線形重回帰では実測値との予測ズレが大きいパターンを手続き的に取り出している。

4. まとめ

パターン認識によるアプローチは第一章の Hansch-Fujita 法と比較すると、目的指向の高いアプローチである。これは、パターン認識自体が工学分野で展開された手法で、分からない部分は取りあえずブラックボックスとして解析を進め、結果を先取りするという手続きを持つ為である。この特性が故に、活性メカニズムが特定しにくく、仮説を立てることが困難な毒性や分解といった分野の解析にも利用できるという、他の構造ー活性相関手法には無いパターン認識

法特有のメリットを生み出している。

パターン認識によるアプローチの最大の特徴は解析に先立って仮説を持たない発見型あるいは絞り込み型のアプローチを取る事である。他の構造ー活性相関手法が総て種々の仮説に基づいて解析手法を展開する証明型のアプローチを取ることに比較するならばその特徴は際だったものとなる。

パターン認識法は解析後に原因／要因が特定されるので、全く新規で情報の無い（仮説等が立てられない）問題に関しても必要なデータさえあれば解析を実施出来る。むしろ、仮説を立てる前に行う要因解析手法の一つと考えた方が良い。この点が、パターン認識手法はあらゆる分野の基本手法と称される所以である。