In Silico Data
Miracles by the KY-methods

The Chem-Bio Informatics Society 情報計算化学生物学会

CBI学会2019年大会、2019年10月21日（月）、13：00－17：00

〈チュートリアル〉

計算毒性学と化学データサイエンスの基本

株式会社　インシリコデータ
湯田　浩太郎

本日のプログラム：

1．13:00-13:05:（5分） 挨拶：株式会社　インシリコデータ　湯田浩太郎

2．13:05-13:20（15分） ◆導入　計算毒性学と「化学データサイエンス」
計算毒性学でのコンピューター導入原理、二大毒性評価関連技術（化学多変量解析／パターン認識アプローチ、人工知能アプローチ）、データサイエンスから「化学データサイエンス」へ

3．13:20-13:50（30分） ◇第一部　計算機化学（Computer Chemistry）関連
化合物保存形式、化合物命名法、化合物検索（完全一致、部分構造、2・3次元構造検索、他）手法、一元一項対応串刺し検索、化合物の扱い（縮合多環、互変異性、立体／幾何異性）、化合物表記（ケトエノール、ニトロニトロソ、他）

4．13:50-15:20（90分） ◇第二部　化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連
化学パラメーター、2／3次元パラメーター、種々データ解析手法、過剰適合、偶然相関、線形／非線形性、特徴抽出、最少サンプル数、最少パラメーター数、クラスポピュレーション、次元変換／圧縮／縮小、分類率／予測率、要因解析、オートスケーリング、アウトライヤー／インライヤー、解析信頼性指標（サンプル数／パラメーター数）、KY（K-step Yard sampling）法、パーセプトロン、バックプロパゲーション、遺伝的アルゴリズム、ファジー理論、内挿／外挿問題、他

＜15:20-15:40　休憩 20分＞

5．15:40-16:20（40分） ◇第三部　人工知能（Artificial Intelligence）関連
人工知能の歴史、ルールベース型人工知能、ニューラルネットワーク型人工知能、深層学習、サンプル数問題、要因説明問題、ルールのコンピューターへの組み込み、ネットワーク構造、LISP、FORTRAN、PYTHON、

6．16:50-17:00（30分） ◇第四部　計算機科学（Computer Science）関連
データベース理論、プログラミング言語、クラスター、クラウド、スーパーコンピューター、ネットワーク、WEB、他

7．16:50-17:00（10分） ◇討論および名刺交換会

4．解析に使うパラメーター

◆化合物関連パラメーター

■　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

4．解析に使うパラメーター

◆化合物関連パラメーター:トポロジカルパラメーター

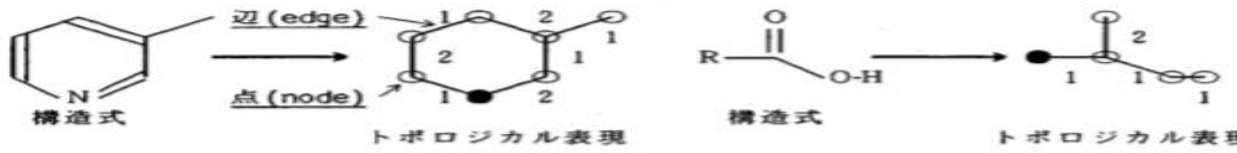
□　トポロジカルデータの特徴

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。

4．解析に使うパラメーター

◆化合物関連パラメーター:トポロジカルパラメーター

□　ＭＣＩ値の算出法

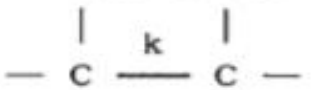
まず化合物を構成している個々の結合についてＣκ値を求める。　続いて、このＣκ値を化合物中の総ての結合について総和した値が分子に対するＭＣＩ値となる。

$$MCI = \sum_{k=1}^{m} C_\kappa = \sum_{k=1}^{m} \frac{1}{[\,L_i \cdot L_j\,]^{1/2}}$$

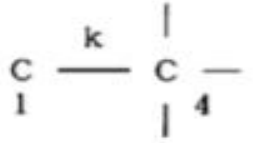
k：　ある一つの結合のＩＤ番号
i：　結合ｋを形成する原子２個のうちの一つの原子に関するＩＤ
j：　結合ｋを形成する原子２個のうちｉ以外の原子に関するＩＤ

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）　Ｃκ値の求め方



$$C_\kappa = \frac{1}{[\,3 \cdot 3\,]^{1/2}} = \frac{1}{3}$$

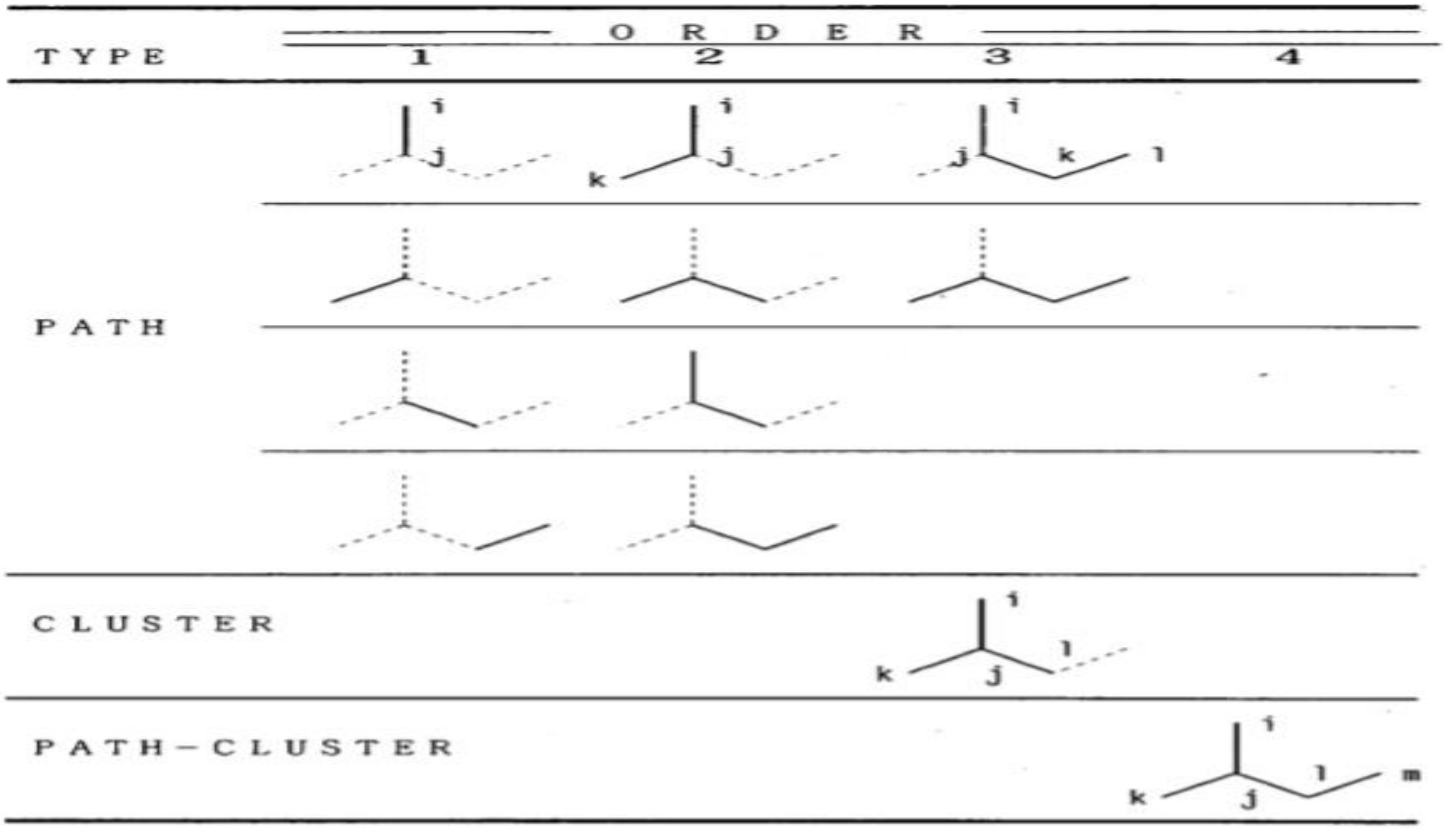


$$C_\kappa = \frac{1}{[\,1 \cdot 4\,]^{1/2}} = \frac{1}{2}$$

4．解析に使うパラメーター

◆化合物関連パラメーター:トポロジカルパラメーター

例）ＭＣＩにおける次数と結合タイプの概念及びＣｘ計算式



4．解析に使うパラメーター

◆化合物関連パラメーター:トポロジカルパラメーター

□　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_S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_Sとなる。　従って、次数が３以上で現れる
・ＰＡＴＨ－ＣＬＵＳＴＥＲはC_S内部にＰＡＴＨ部分とＣＬＵＳＴＥＲ部分を持つ。

4．解析に使うパラメーター

◆化合物関連パラメーター:トポロジカルパラメーター

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

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

□　次数（ORDER）が異なる時のC_{x}の計算式（次数1～4について）

$$\text{次数 1} = \sum_{s=1}^{N_{a}} (\delta_{i}\ \delta_{j})_{s}^{1/2}$$

$$\text{次数 2} = \sum_{s=1}^{N_{a}} (\delta_{i}\ \delta_{j}\ \delta_{a})_{s}^{1/2}$$

$$\text{次数 3} = \sum_{s=1}^{N_{a}} (\delta_{i}\ \delta_{j}\ \delta_{k}\ \delta_{l})_{s}^{1/2}$$

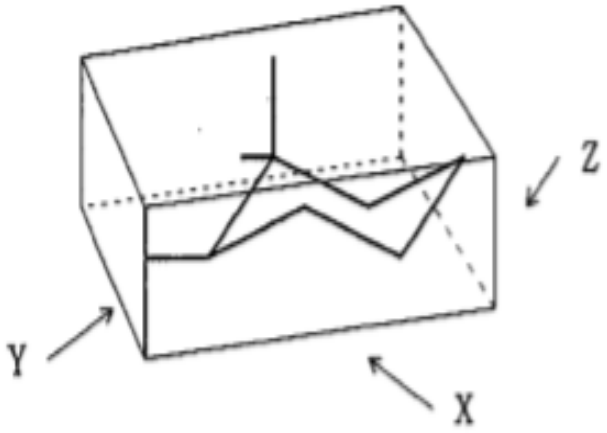
$$\text{次数 4} = \sum_{s=1}^{N_{a}} (\delta_{i}\ \delta_{j}\ \delta_{k}\ \delta_{l}\ \delta_{m})_{s}^{1/2}$$

4．解析に使うパラメーター

◆化合物関連パラメーター:トポグラフィカルパラメーター

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

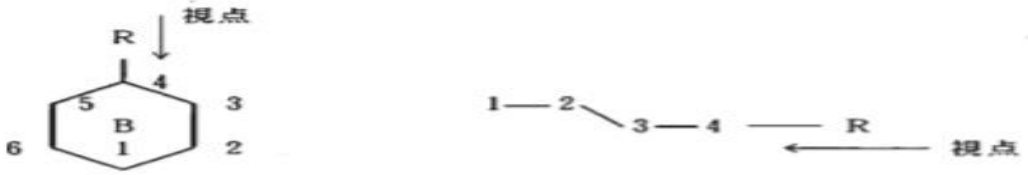
このパラメータにより、分子全体の立体的な形状についての情報がえられる。 例えば、分子が平面に近い、細長い、立方体に近い等の情報である。

4．解析に使うパラメーター

◆化合物関連パラメーター:トポグラフィカルパラメーター

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）

４．解析に使うパラメーター

◆化合物関連パラメーター：物理化学パラメーター

◇種々の物性：
　分子量、 融点、 沸点、分子屈折率、 LogP、 Hammetσ、 その他

◇分子軌道法関連パラメーター：
　電子密度、 HOMO、 LUMO、 分極率、 双極子モーメント、その他

◇分子力学関連パラメーター：
　結合エネルギー、トーションエネルギー、 水素結合エネルギー、 その他

4．解析に使うパラメーター

◆化合物関連パラメーター：物理化学パラメーター　（LogP）

□ 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

4．解析に使うパラメーター

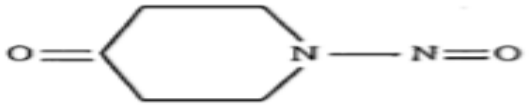
◆化合物関連パラメーター：物理化学パラメーター（LogP）

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

４．解析に使うパラメーター

◆化合物関連パラメーター：その他のパラメーター
◇部分構造パラメーター

部分構造を定義し、解析対象とする化合物中に定義した部分構造が含まれているかいないかを利用する

部分構造カウントの結果データの表記手法により以下の３種類に大別できる

１．**部分構造があるかないかのバイナリーデータ**
　１または０のバイナリーデータ
　＊MACCS，PubChem, Daylight等が提供している

２．**内包される部分構造の数をカウント**
　整数値のパラメーター
　＊フィンガープリントの拡張版

３．**内包された部分構造の隣接原子の情報を加味して数値化**
　連続変数パラメーター
　＊ADAPT(Automated Data Analysis by the Pattern recognition)で開発/採用

４．解析に使うパラメーター

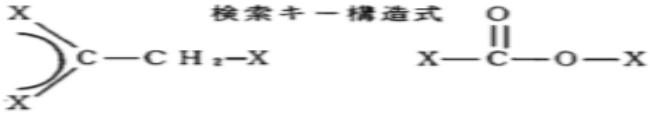
◆化合物関連パラメーター：その他のパラメーター
◇部分構造パラメーター

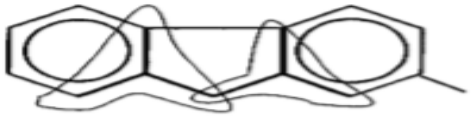
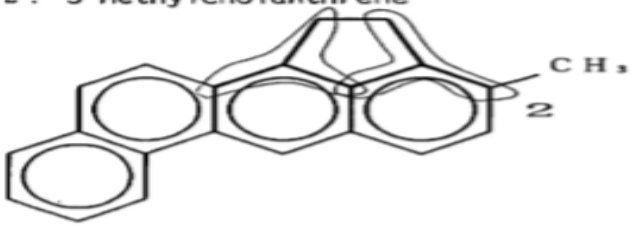
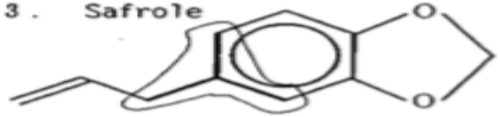
１．部分構造があるかないかのバイナリーデータ
１または0のバイナリーデータ
＊MACCS，PubChem, Daylight等が提供
＊この種のパラメーターは一般的に「フィンガープリント」と呼ばれる
＊パラメーター数は数百から千種類提供される
＊検索される部分構造は提供元の適用目的等により変化する

□適用目的
＊多変量解析／パターン認識による解析のパラメーターとして適用
＊化合物の類似度評価等に利用される事が多い

4．解析に使うパラメーター

◆化合物関連パラメーター:
◇部分構造パラメーター

検索キー構造式 

被検索化合物	部分構造数	MCI	部分構造数	MCI
1．2-Aminofluorene 	2	3．166	0	0．0
2．3-Methylcholanthrene 	2	3．301	0	0．0
3．Safrole 	1	2．690	0	0．0
4．Ethyl acetate 	0	0．0	1	1．904

４．解析に使うパラメーター

◆化合物関連パラメーター：その他のパラメーター
◇部分構造パラメーターの特徴

特徴：
①部分構造パラメーターは化合物構造式に直結するパラメーターのため、
　他のパラメーターと比較して要因解析がしやすい
②部分構造パラメーター情報は、薬理活性や毒性等のコントロールに必要な情報を
　化合物の構造情報として捉えることが出来る
③部分構造パラメーターの情報は化合物合成等に反映しやすい
④データ解析のみならず、化合物検索等にも適用できる

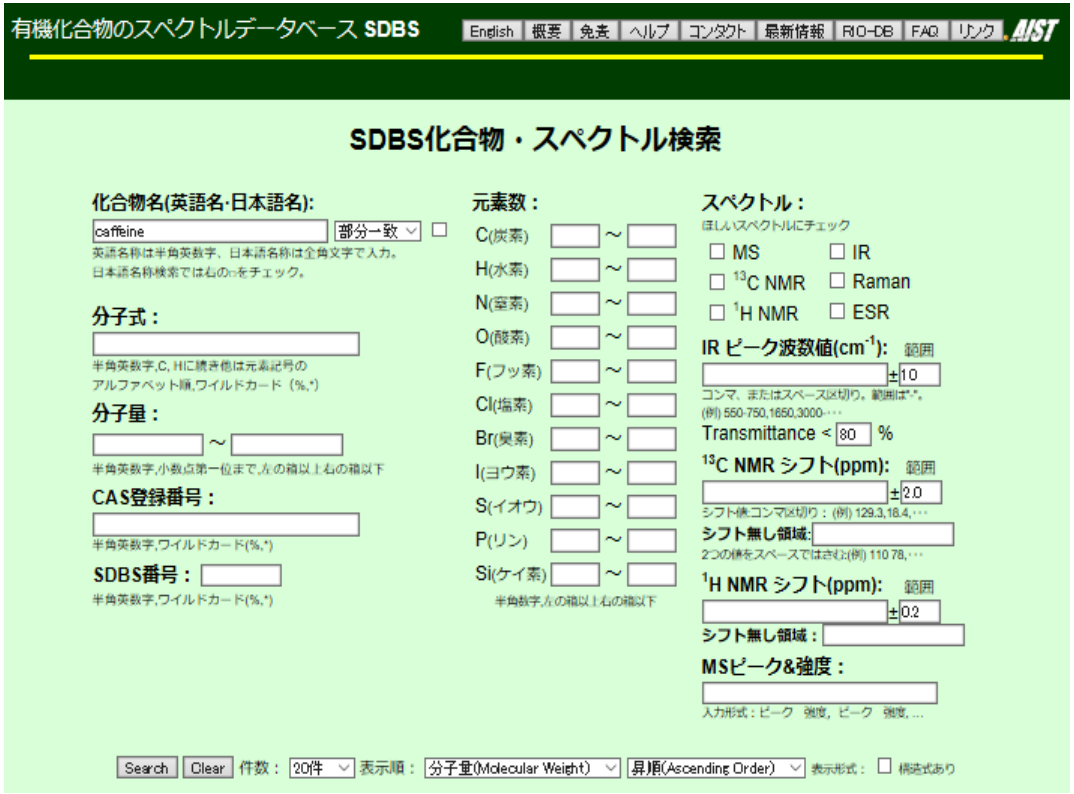
留意点：
①設定する部分構造の内容が実施目的の実現性に大きな影響を及ぼす
例：薬理活性では、ファーマコフォア等に留意した部分構造
　化合物毒性では、毒性要因に関連する部分構造
②部分構造の設定にノウハウや経験がある程度実施結果に影響することがある

4．解析に使うパラメーター

◇　機器スペクトルパラメーター

■有機化合物の
　スペクトルデータベース SDBS

トップ画面



https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

4．解析に使うパラメーター

◇　機器スペクトルパラメーター

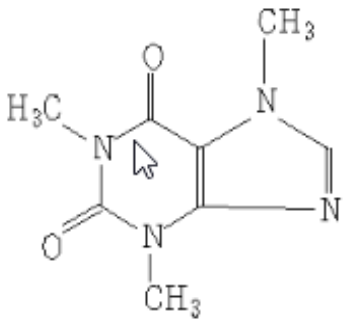
■有機化合物のスペクトルデータベース SDBS

SDBS No: 1898 **CAS Registry No.:** 58-08-2

DOI:

Molecular Formula: $C_8H_{10}N_4O_2$ **Molecular Weight:** 194.2

SDBS-NO= 1898
CAFFEINE



Compound Name:

caffeine
1,3,7-trimethyl-3,7-dihydro-1H-purine-2,6-dione
1,3,7-trimethyl-3,7-dihydro-purin-2,6-dion, kaffein
1,3,7-trimethyl-3,7-dihydro-purine-2,6-dione
1,3,7-trimethylxanthine
1H-purine-2,6-dione, 3,7-dihydro-1,3,7-trimethyl-
3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione
theine

InChI:

InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

InChIKey:

RYYVLZVUVIJVGH-UHFFFAOYSA-N

Publisher:

National Institute of Advanced Industrial Science and Technology (AIST)

https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

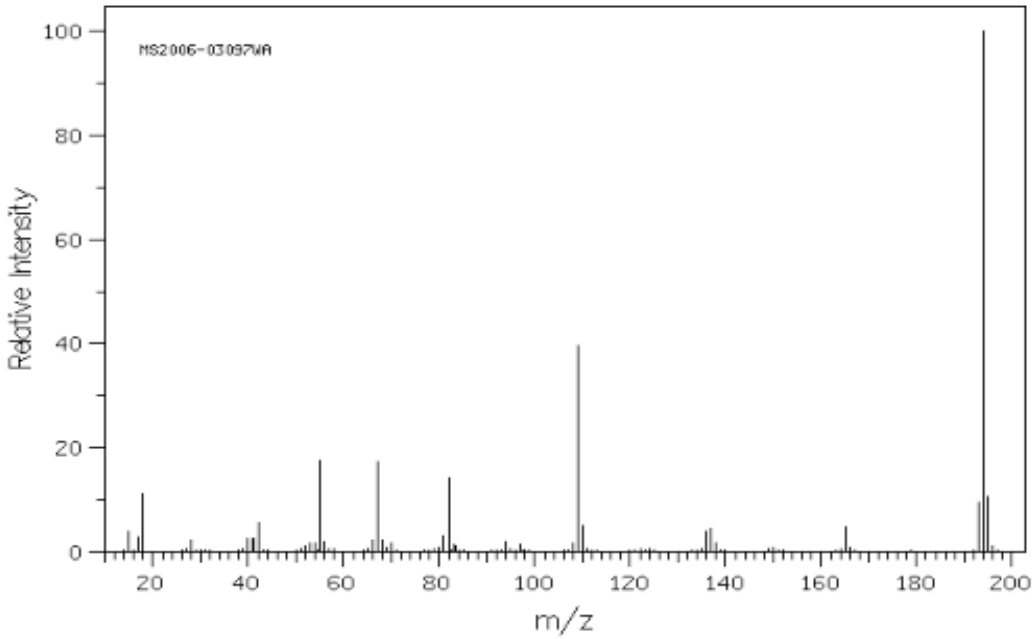
４．解析に使うパラメーター

◇　機器スペクトルパラメーター

SDBS-Mass

MS2006-03097WA　　SDBS NO.　1898
caffeine
C8H10N4O2　　(Mass of molecular ion:　194)

Mass



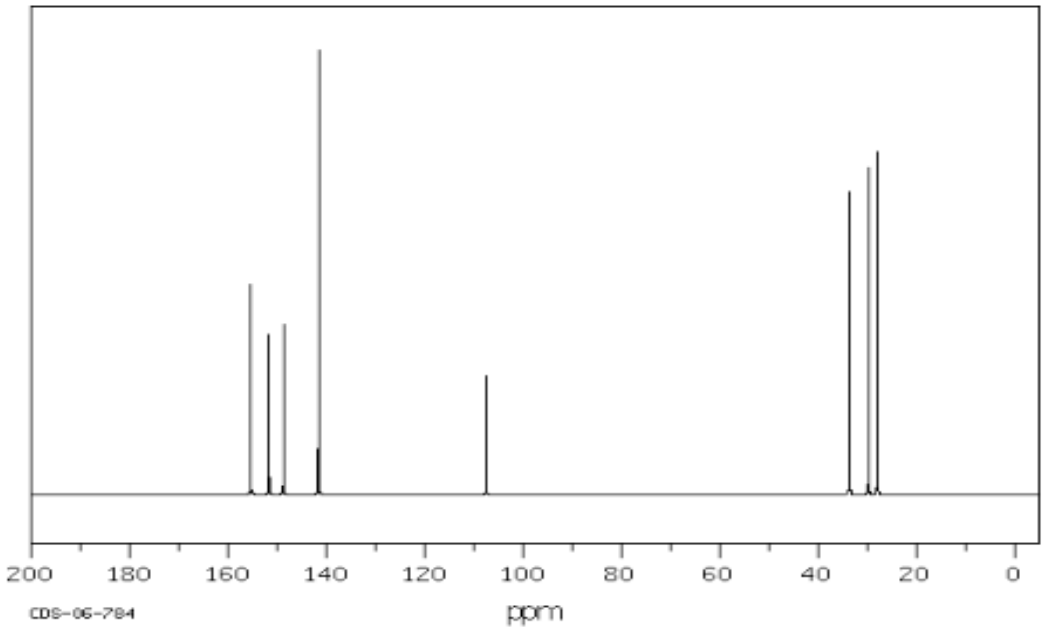
https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

4．解析に使うパラメーター

◇ 機器スペクトルパラメーター：

SDBS-^{13}C NMRSDBS No. 1898CDS-06-784
$C_8H_{10}N_4O_2$
caffeine

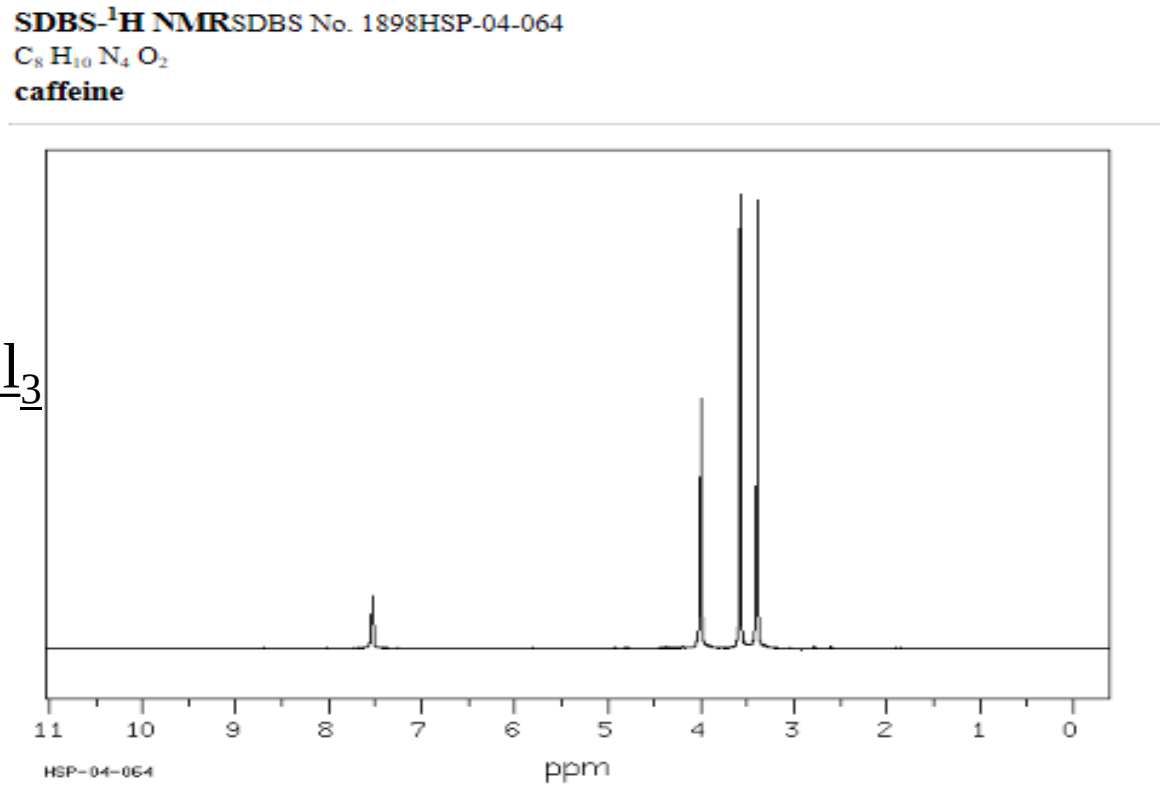
^{13}C NMR : in $CDCl_3$



https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

４．解析に使うパラメーター

◇　機器スペクトルパラメーター



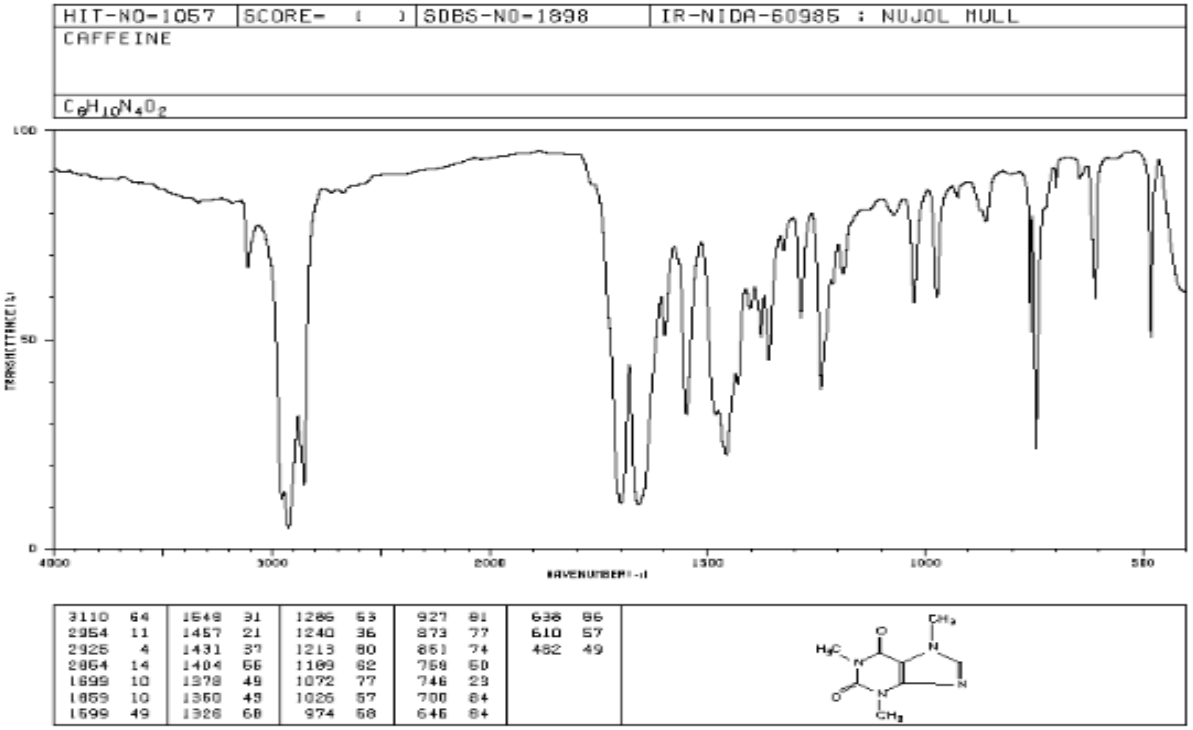
^{1}H NMR : 90 MHz in $CDCl_3$

https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

4．解析に使うパラメーター

◇ 機器スペクトルパラメーター

IR：nujol mull

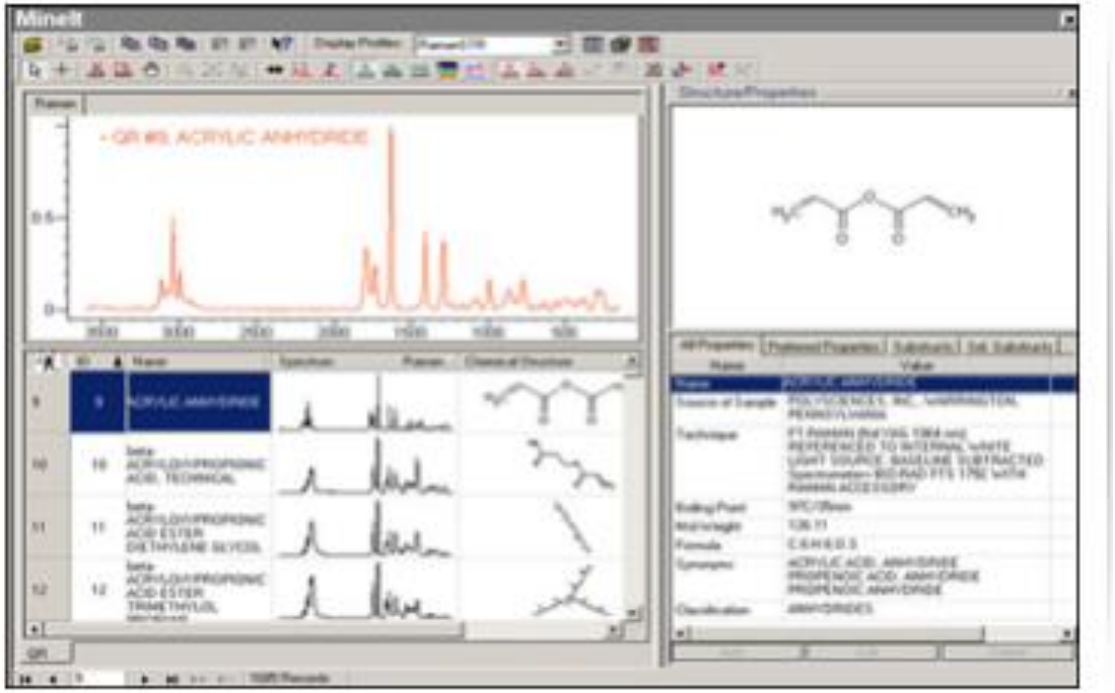


https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

４．解析に使うパラメーター

◇　機器スペクトルパラメーター

Ramanデータ



BIO-RADのスペクトルデータベースより

http://www.bio-rad.com/

http://www.bio-rad.com/ja-jp/product/raman-spectral-databases?ID=N0ZXPS4VY

４．解析に使うパラメーター

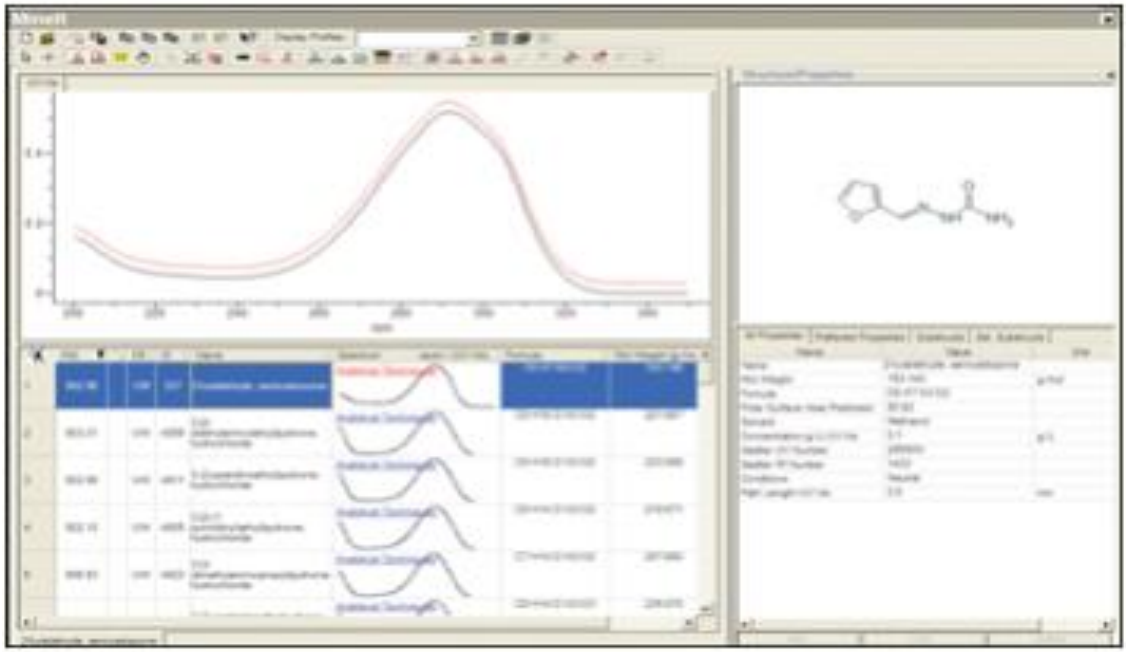
◇　機器スペクトルパラメーター

UV-Visスペクトル

BIO-RADのスペクトルデータベースより

http://www.bio-rad.com/

紫外可視データベース

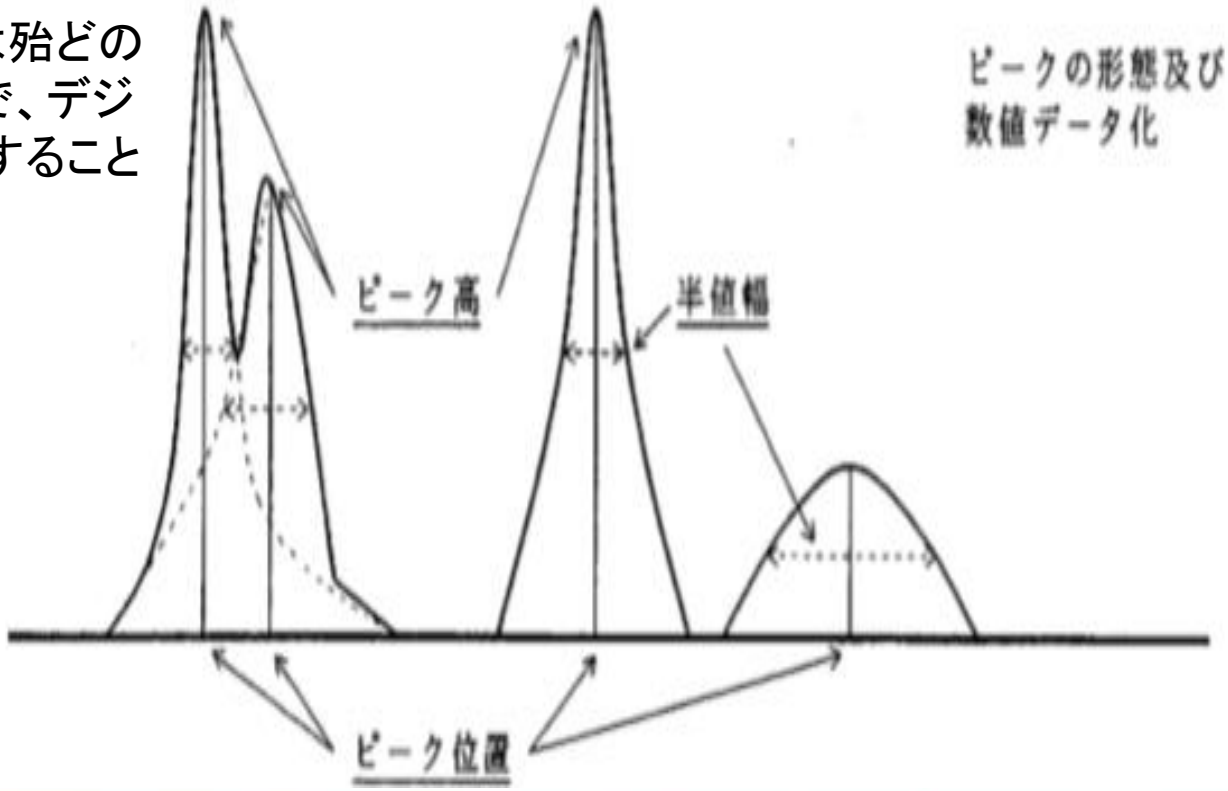


http://www.bio-rad.com/ja-jp/product/uv-vis-spectral-databases?ID=NH262L4VY

4．解析に使うパラメーター

◆ 機器スペクトルパラメーター

＊機器スペクトルデータは殆どの場合アナログデータなので、デジタルの数値データに変換することが必要である。



４．解析に使うパラメーター

◆ 機器スペクトルパラメーター

特徴：

①機器があれば数値データとして簡単に蓄積できる
②スペクトルチャートは様々な実験過程で種々蓄積される

留意点：

①一般的にパラメーター数が極めて大きくなりやすい
結果として、データ解析手法が限定、適用不可となる可能性が高くなる
②スペクトルデータは多重共線性が極めて高い
①と②の特徴により、データ解析実施においては次元圧縮・統合等が必要で、
解析手法もPLSやPCAと制限されることが多い
③スペクトルチャートの測定条件等統一が必要
出来れば測定機器メーカーや機種も統一
例：60MのH-NMRデータと90MのH-NMRデータを混在してのデータ解析は無意味
・測定条件等が統一されないとデータ解析の精度が保証されにくくなる

4．解析に使うパラメーター

◆ 機器スペクトルパラメーター

＊例え同じスペクトルで、測定環境条件を同じにしたとしても測定機器やソフトウエアの
違いにより多種多様のフォーマトがある。

＊スペクトルデータベース間でのデータのやり取りにはファイルフォーマト変換が重要

NMR

Vendor	File Format	Required Parameter Files	Optional Parameter Files
ACD/Labs	*.spectrus, *.esp, *.txt		
Acorn NMR, Inc.	*.fid, *.nmr, *.2d		
Agilent (Varian, Inc.)	data, *.fdf, fid0001.fdf, *.txt, fid, phasefile	acq, proc, procpar	acq_2, text
ASCII†	*.txt; *.prn, *.csv, *.asc		
Bruker Corporation	ser, rr, .fid, *r, 1i, 2rr, *.* (DISNMR)	acqus, procs, acqu2, proc2s, *.fqs, *.fa1, *.fa2, *.fp1, *.fp2	title, intrng, *.tit, *.ti2
GE	*.raw, *.* (Nicolet)		
JCAMP†	*.dx; *.jdx		
JEOL Ltd.	*.als, *.jdf, *.nmfid, *.nmf, *.nmdata, *.nmd, *.gxd, *.bin, *.* (Delta)	*.gxp, *.hdr	exp.param, exp.par
Lybrics	*.*		
MSI Felix	*.*		
Tecmag	*.tnt, *.* (MacNMR)		
Thermo Scientific†	*.spc		

https://www.acdlabs.com/products/fileformats/

4．解析に使うパラメーター

- 機器スペクトルパラメーター

Chromatography

Vendor	Data Format	Extension	Comments
Agilent Technologies	1100 Series LC/MSD Quad and Ion Trap Systems	*.ms, *.yep	UV, LC-UV and LC-MS
	ChemStation Rev. B.02.01, B.03.01, B.04.01, B.04.02, B.04.03, Rev. C.01.04	*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	Open Lab C v.1.04	*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	Open Lab Rev. C.01.07	*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	EZChrom	*.dat	UV traces only.
AB SCIEX		*.wiff	LC-UV and LC-MS
Bruker Daltonics and Agilent Technologies			LC-UC and LC-MS
Bruker	Compass (accurate mass data)	*.D	LC-MS, LC-UV, UV Entire *.D folder structure should be used.
Shimadzu Corporation	LCMS-IT-TOF	*.lcd	LC-MS and LC-UV. Requires vendor software on same computer.
	LCMSsolution	*.qld	LC-MS, LC-UV and UV traces May require vendor software on same computer
Thermo Scientific	Xcalibur	*.raw	LC-MS, LC-UV and UV traces
	Chromeleon® 6		UV and LC-UV, via Connect to

Mass Spectrometry

Vendor	Data Format	Import	Export	Extension	Comments
ACD/Labs	ACD/Labs	✔	✔	*.spectrus, *.esp	
Agilent Technologies	1100 Series LC/MSD Quad and Ion Trap Systems	✔		*.ms, *.yep	DAD data and single chromatogram curve are imported also. Splitter available
	ChemStation	✔		*.ms	Splitter available
	LC TOF	✔		*.wiff	
	MassHunter (6000 Series)	✔		*.bin	Entire *.D folder structure should be used. Agilent component requires Microsoft .NET version 2. DAD can be imported (V12) and MS/MS split controlled in newer versions
	Open Lab C v.1.04	✔		*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	Open Lab Rev. C.01.07	✔		*.D	UV, LC-UV and LC-MS Entire *.D folder should be used
AB SCIEX	Analyst	✔		*.wiff	Single mass spectra, LC-MS and most LC-MS" imported. Splitter available. UV data not currently imported
		✔		See above	LightSight—Spectra are 'pushed' via ACD/Labs (NetCDF)
	Analyst QS	✔		*.wiff	Single mass spectra, LC-MS and most LC-MS" imported. Splitter available
	Analyst TF	✔		*.wiff	
Applied Biosystems	Mariner Data Explorer ASCII LC/MS	✔		*.txt	LC-MS data only
Bruker Daltonics and Agilent Technologies	Agilent or Bruker LC/MS Ion Trap	✔		*.yep	LC-MS and DAD data

Optical Spectropy

Vendor	Data Format	File Format	Comments
ACD/Labs	ACD/Labs	*.spectrus, *.esp	
Agilent Technologies	HP 84552A	*.wav	
	ChemStation†	*.uv	
ASCII single, dual and multicolumn		*.txt; *.prn; *.csv; *.asc	
Bruker	OPUS	*.*	
DeltaNu		*.spc	
Dionex	Chromeleon†		"Connect to" ability available
Foss NIRSystems		*.da	
JASCO Corporation	J-700	*.js	
JCAMP, JCAMP multispectra		*.dx; *.jdx	
LabControl		*.uvd; *.irs	
MATLAB DSO†		*.mat†	
Ocean Optics			
PerkinElmer Instruments			
Shimadzu	IR	*.irs	
Thermo Scientific	Galactic	*.spc	
	Mattson	*.*	
	Nicolet OMNIC	*.spa; *.spg	
Varian	Cary UV	*.b*; *.d*	
Waters Corporation†	Empower and Empower 2		"Connect to" ability available
	MassLynx	*.inf	
	Millennium®†		"Connect to" ability available

https://www.acdlabs.com/products/fileformats/

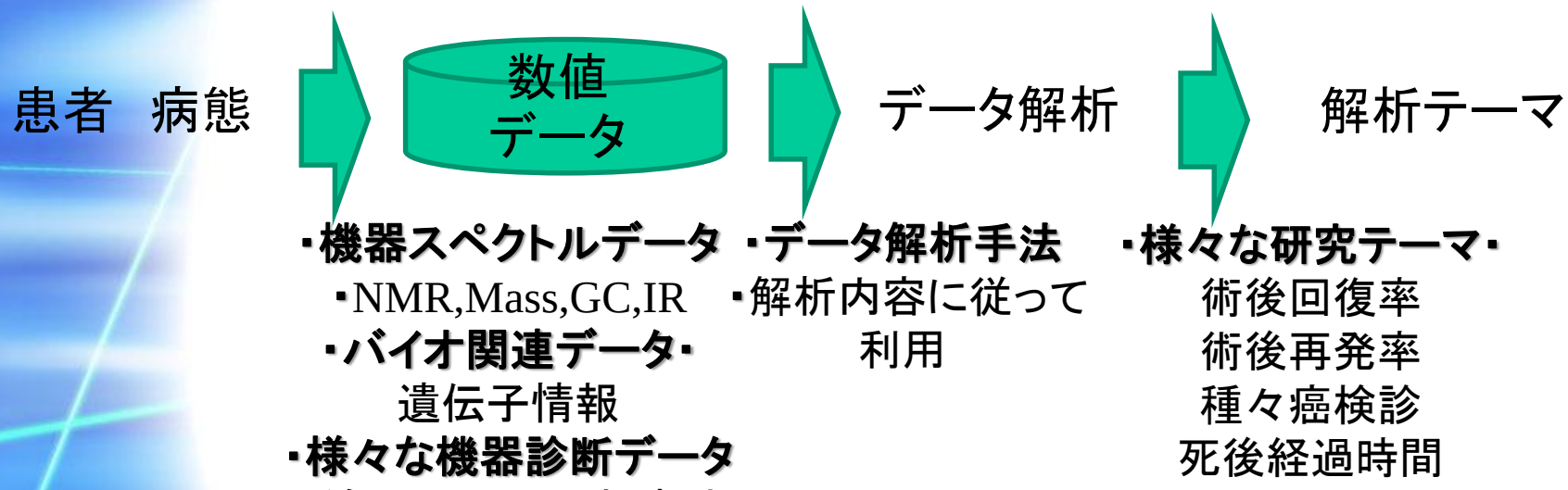
４．解析に使うパラメーター

◆　医療関連データ

＊医療分野でのデータ解析はデータのとり方で様々な解析を実施できる

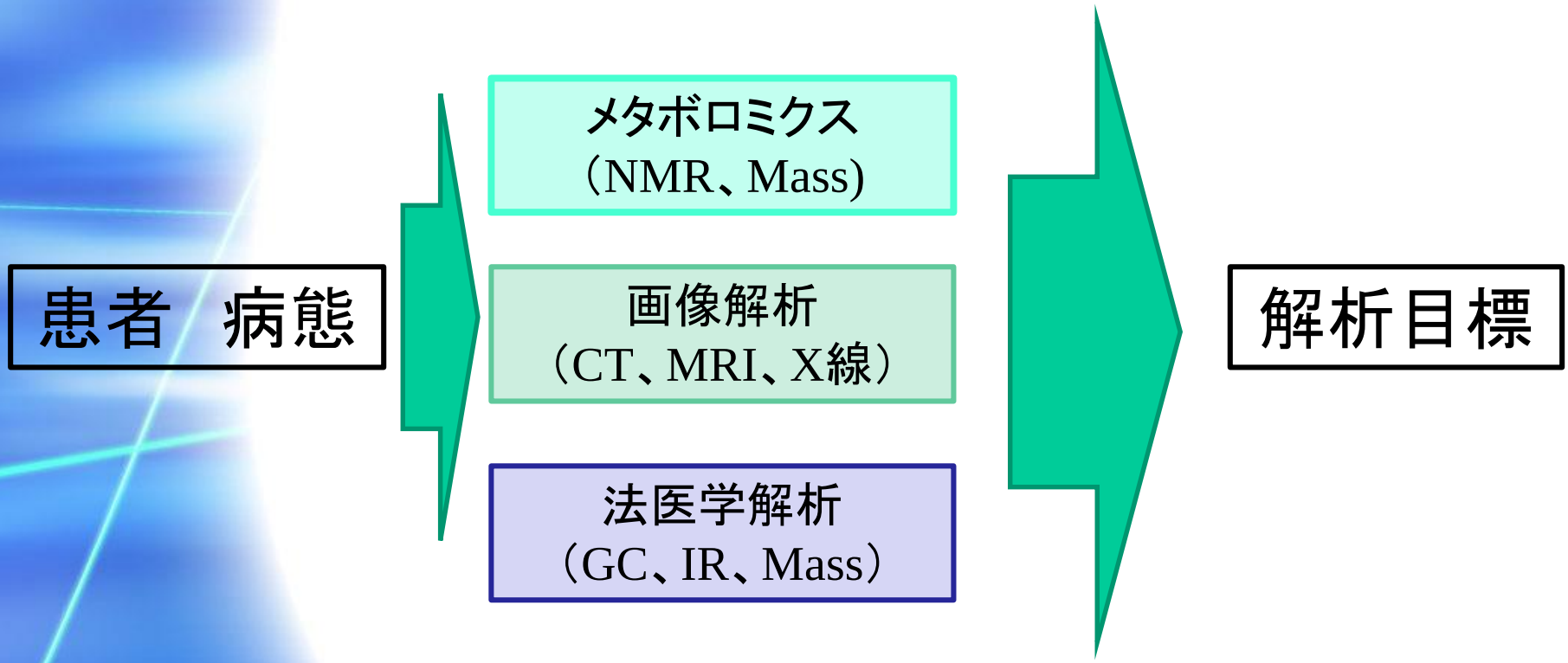
＊現在は医療関連は統計解析を適用し、様々な病気との因果関係や薬効検証等が中心

＊今後は**アイディア次第で様々な解析**を実施できるようになる



４．解析に使うパラメーター

◆ 医療関連データ:
研究分野により名前や分析機器の種類が変わる



□パラメーターの操作や選択に関連する事項

1. パラメーターの単位等に関する留意事項

2. パラメーター選択について

3. サンプルの選択について

4. データ解析を保証するための制限事項等

◇パラメーターの正規化

桁数が大きく異なるパラメーターをデータ解析時に混在すると、要因解析等を困難にする要因となる。
このようなパラメーター群を利用する場合は、通常オートスケーリングを実施する。
オートスケーリングは数値データを平均0で標準偏差1のパラメーターに変換する技術である。

■桁数の異なるパラメーター事例

以下には値の桁数が異なるパラメーターや、正／負の両方を取る値等列挙する

①桁数の大きくなるパラメーター；分子量（数百から千）、
②値が一桁単位から二けた単位：原子数や環の数等
③値が一桁程度：バイナリーデータ、フラグメントデータ、
④値が小数点以下となる：分子軌道法関連パラメーター
（電子密度、HOMO/LUMO、自由エネルギー、超分極率、ヤング率、その他）
⑤値が生と負の両方を取る：LogPパラメーター、

◇パラメーターの正規化：オートスケーリング

①解析時に用いるパラメーターは、桁数が揃っていることが望ましい
②同じ種類のデータを用いる時は桁数がそろっている場合が多いので問題ないが、
化学分野で扱うデータは内容が多種多様だけでなく、その単位（桁数）も大きく
異なる場合が多い

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

$$W_i' = \frac{W_i - \bar{W}}{Q} \qquad (2)$$

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

◇パラメーターの選択
（特徴抽出：Feature selection or Parameter selection）

＊データ解析の信頼性を保つためにはサンプル数をパラメーター数で割った値である、信頼性指標を守ることが求められる。
このため、少ないサンプル数の場合はパラメーター数を減少させることが必要。
*現在のケモメトリックス解析では、先に述べた化学関連パラメーターはプログラムにより1サンプル（化合物）あたり数千パラメーター発生することが出来る。
＊化学関連研究ではサンプル数が少ないことが多い。このため、解析精度を保つためにもパラメーター数を減少する特徴抽出が極めて重要である。
＊解析に重要なパラメーターは、Intrinsic Parameter、反対がNon-intrinsic parameter。

＊データ解析手法によってはPLSのようにパラメーター数を形式的に減少させることが出来る手法があり、パラメーターを多数発生できる化学分野ではよく利用される。
しかし、PLSは一種の次元変換／圧縮手法で、パラメーターが多いとき緊急避難的に適用される手法。データ解析で重要な要因解析は出来なくなり、分類率や予測率もキレが良くないことになる。

◇パラメーターの選択：特徴抽出

特徴抽出手法は大きく以下に示される4種類に分類される

①パラメーターを構成する値の特徴と、統計的特性を利用した特徴抽出
　0値チエック、 同値データ出現率、フィッシャー比
＊二クラス分類および重回帰（フィッテイング）の両方で利用可能
②パラメーター間の相互関係に注目した特徴抽出（相関係数によるアプローチ）
　単相関、多重相関
＊二クラス分類および重回帰（フィッテイング）の両方で利用可能
③個々のデータ解析手法の特徴を利用した特徴抽出
　個々のデータ解析手法の特徴や機能を用いることで特徴抽出を行う
＊データ解析の種類により二クラス分類や重回帰に適用される
④最適化等の手法を利用することで特徴抽出を行う
　遺伝的アルゴリズム等が利用され、二クラス分類や重回帰に適用される

◇パラメーターの選択：フィッシャー比

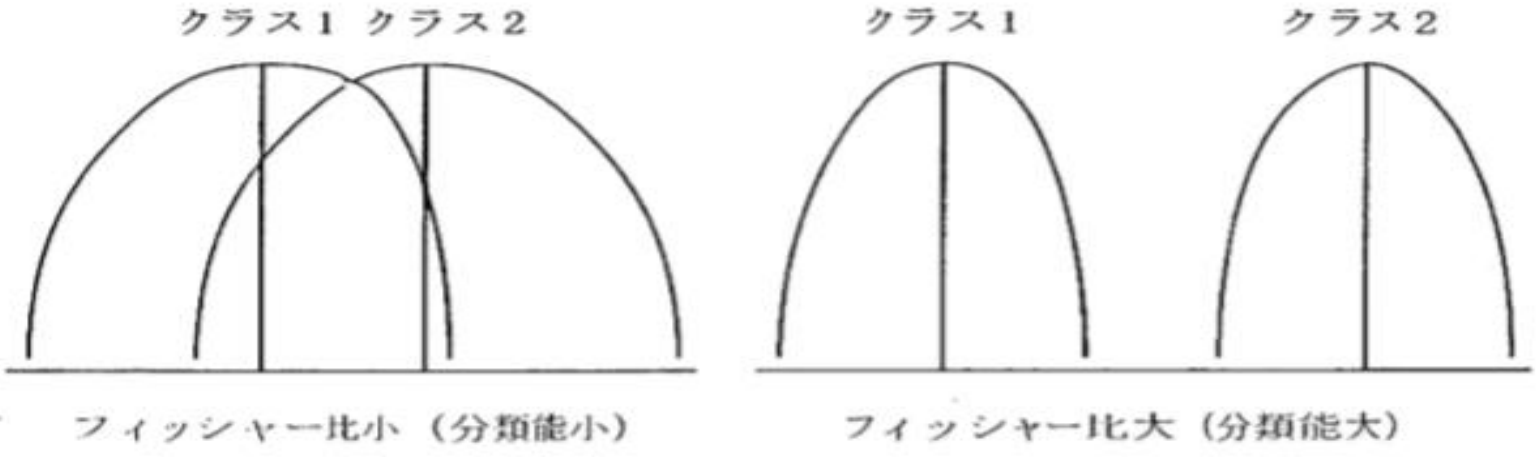


図　．フィッシャー比の値と二クラス分類に対するパラメータの関係
右図は一個のパラメータのみで100%分類が可能な例である

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

$\overline{X1}$：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

◇パラメーターの選択：バリアンスウエイト法

パーセプトロンの特性を利用した特徴抽出手法

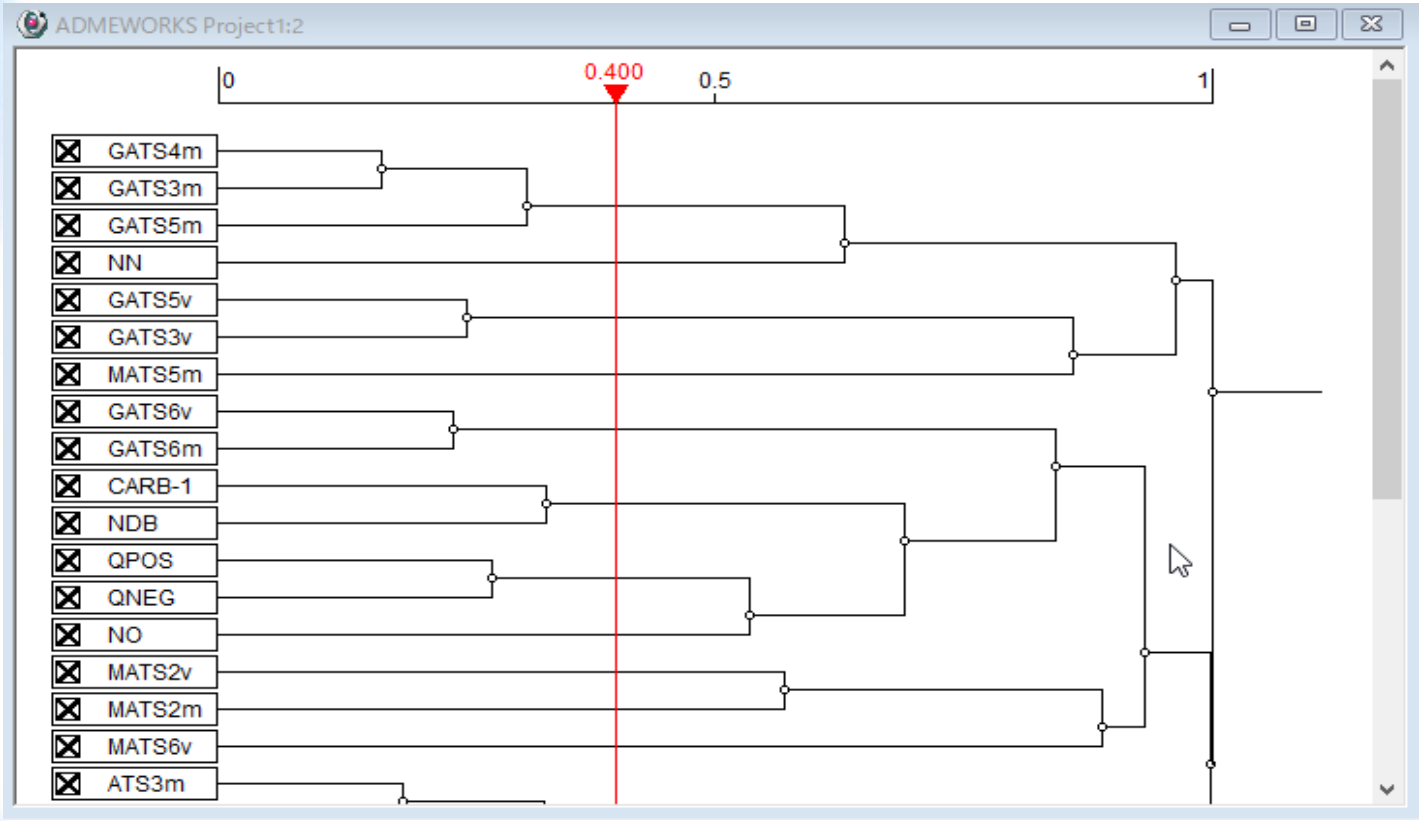
$$VW_j = \frac{V_j}{\overline{W_j}} \quad .$$

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

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。

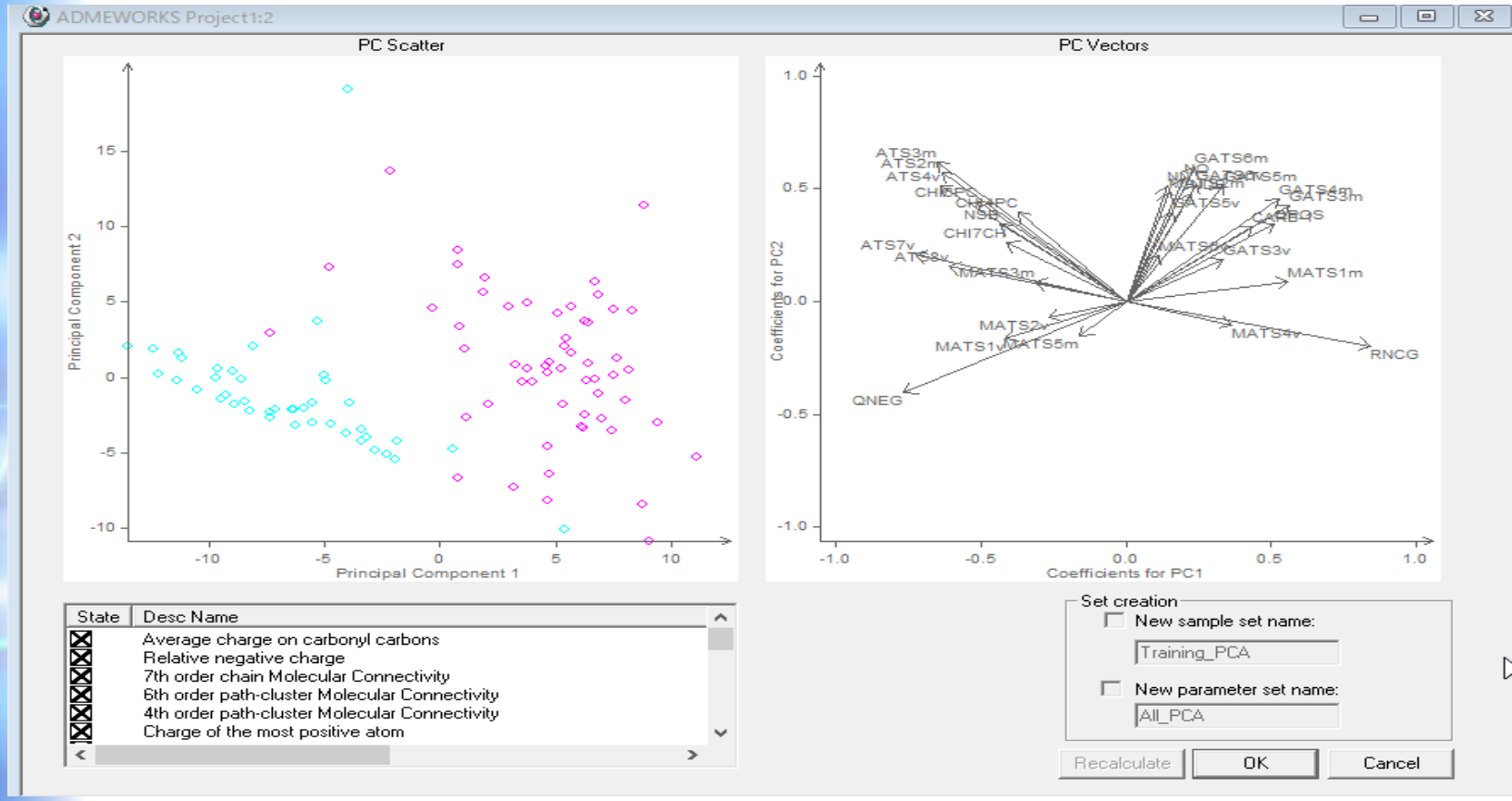
＊1）Jurs P.C. et al., J.C.I.C.S,

◇パラメーターの選択：クラスタリング



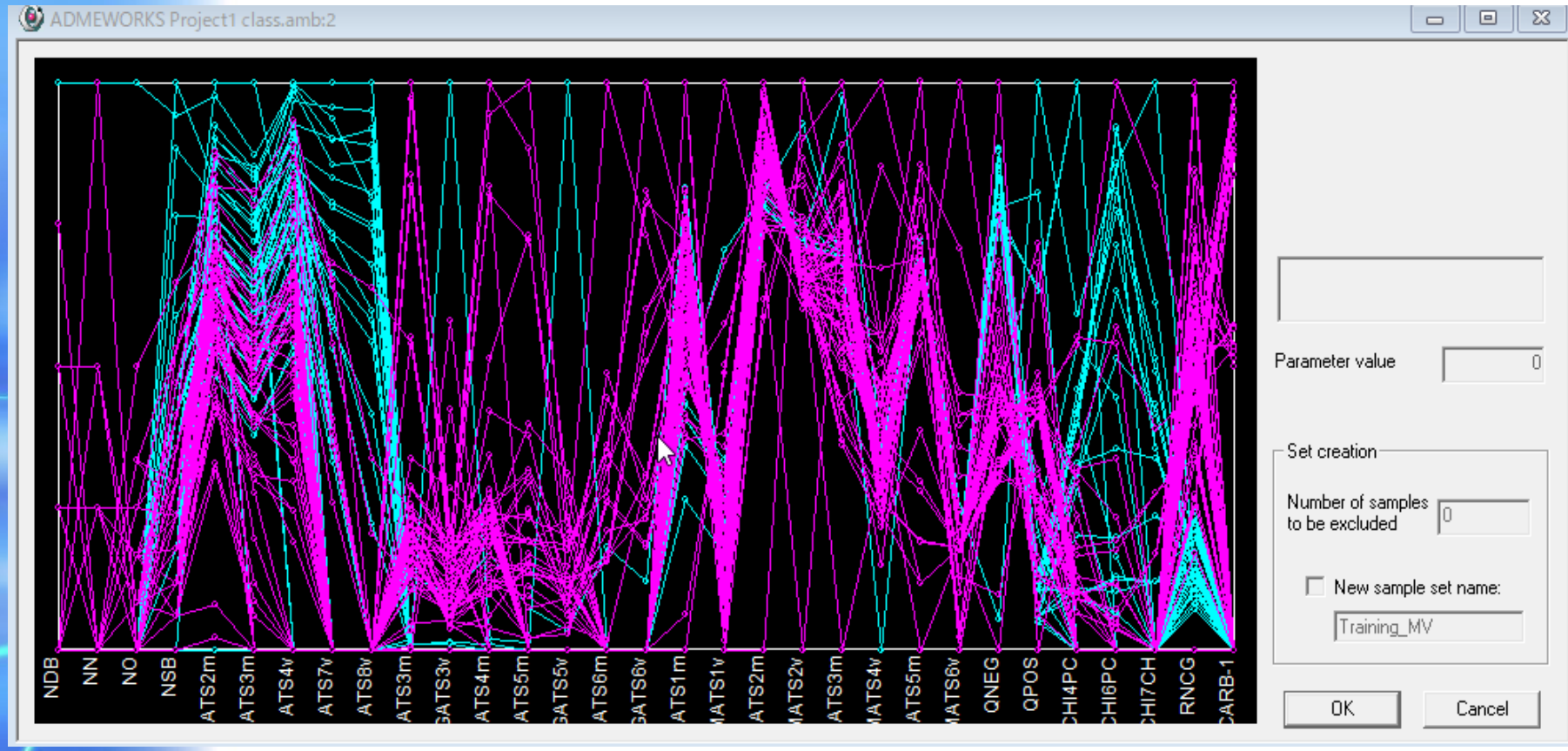
ModelBuilderの画面より

◇パラメーターの選択：主成分分析法（ＰＣＡ）



ModelBuilderの画面より

◇パラメーターの選択：ラインチャート

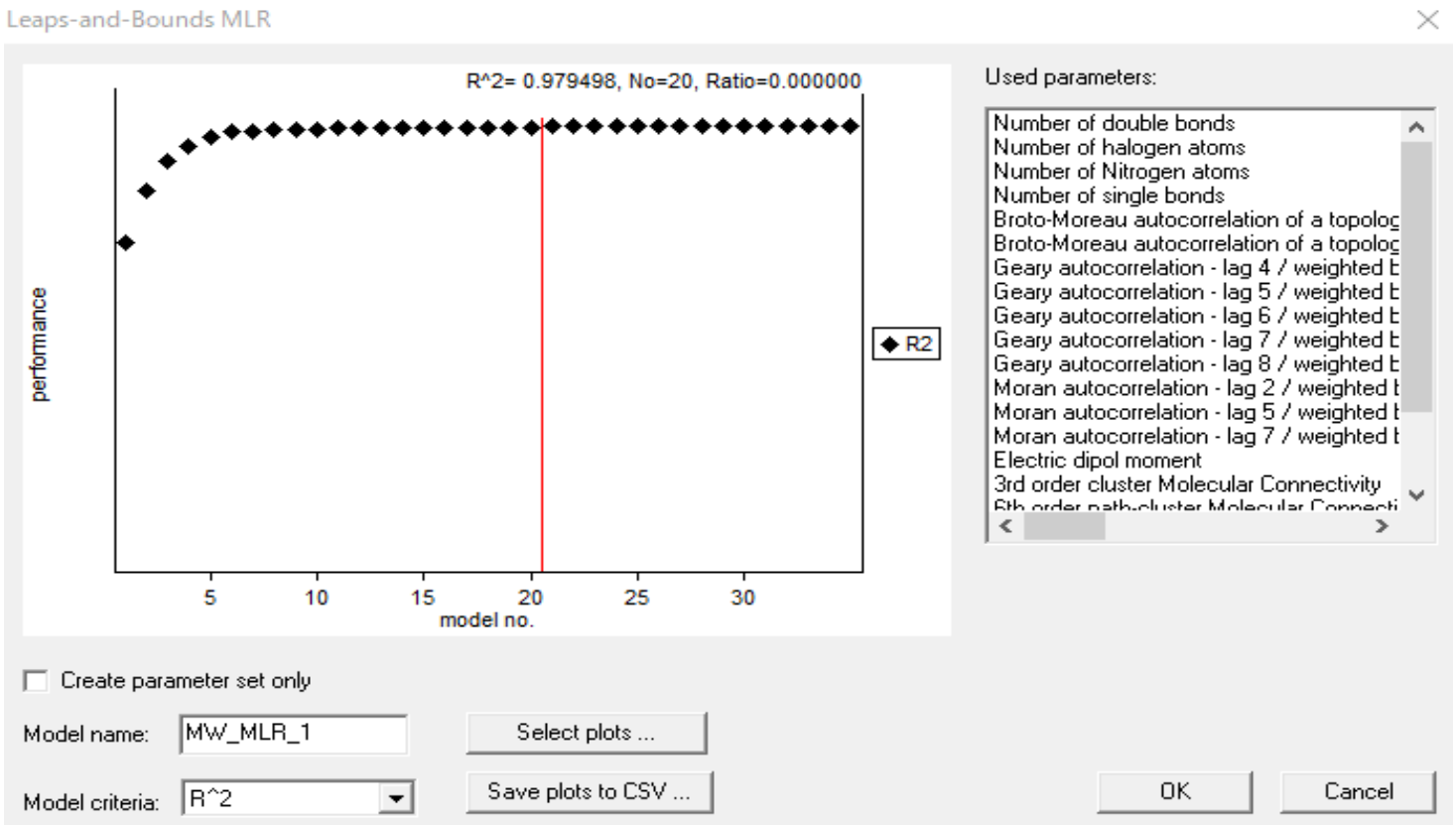


ModelBuilderの画面より

◇パラメーターの選択：クラスディビジョンマップ

ModelBuilderの画面より

◇パラメーターの選択：重回帰



ModelBuilderの画面より

◇パラメーターの選択：データ解析手法の特徴を利用した特徴抽出

クラス分類と重回帰の両方に適用

■二クラス分類
パーセプトロンによる特徴抽出
①ウエイトサイン法
②バリアンスウエイトサイン法

■多クラス分類
SIMCA法の指標を用いた特徴抽出
①モデリングパワー
②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
①忘却学習法
②消却学習法

□PCA（主成分分析）
①因子負荷量の適用

□遺伝的アルゴリズム適用
種々の解析手法と
組み合わせて利用される

■重回帰（フィッテイング）
①前進選択法
②後進選択法
③T検定適用
④総当たり法

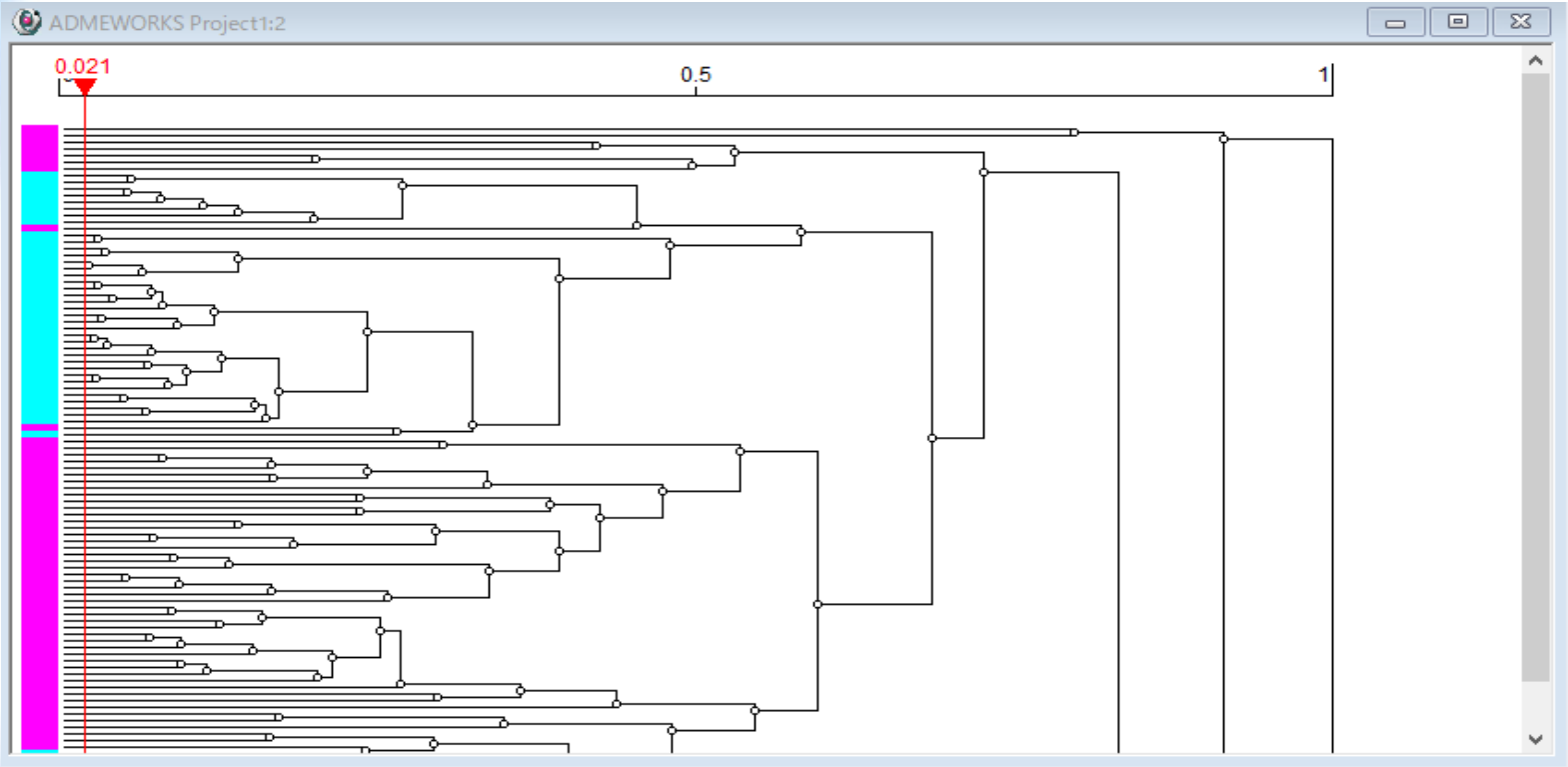
◇サンプルの選択

データ解析で判別関数や重回帰式をリファイニングしてゆく過程で
サンプルも抽出する必要がある

■二クラス分類では誤分類の原因となる「インライヤー」の除去
繰り返し最適化の過程で分類できずに残ったサンプル群の除去
「インライヤー」が特定されたことは、重要な情報源となる

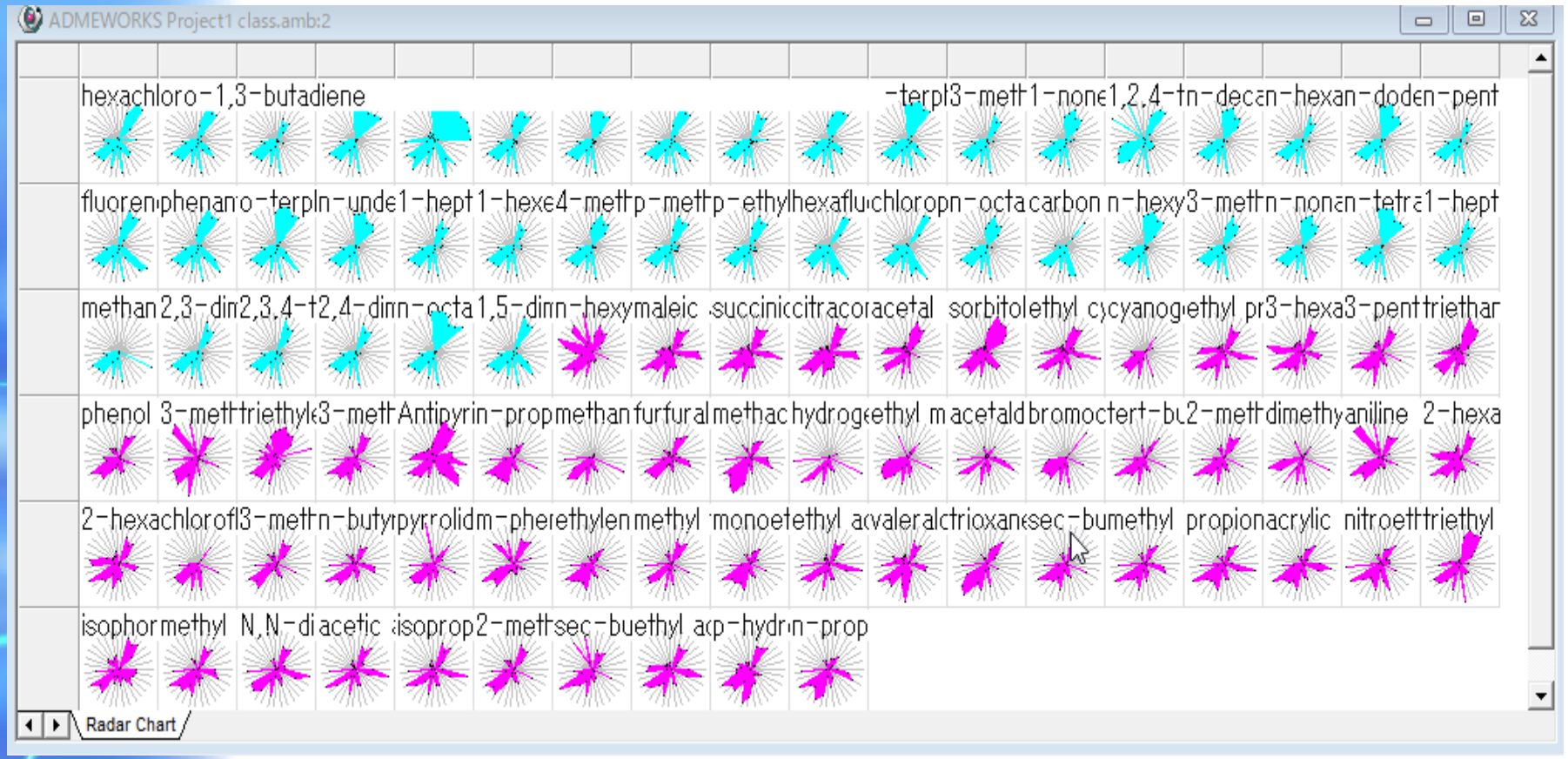
■重回帰では相関係数を下げる原因となる「アウトライヤー」の除去
回帰式をプロットし、アウトライヤーを確認して取り除く
「アウトライヤー」が特定されたことは、重要な情報源となる

◇サンプルの選択：クラスタリング



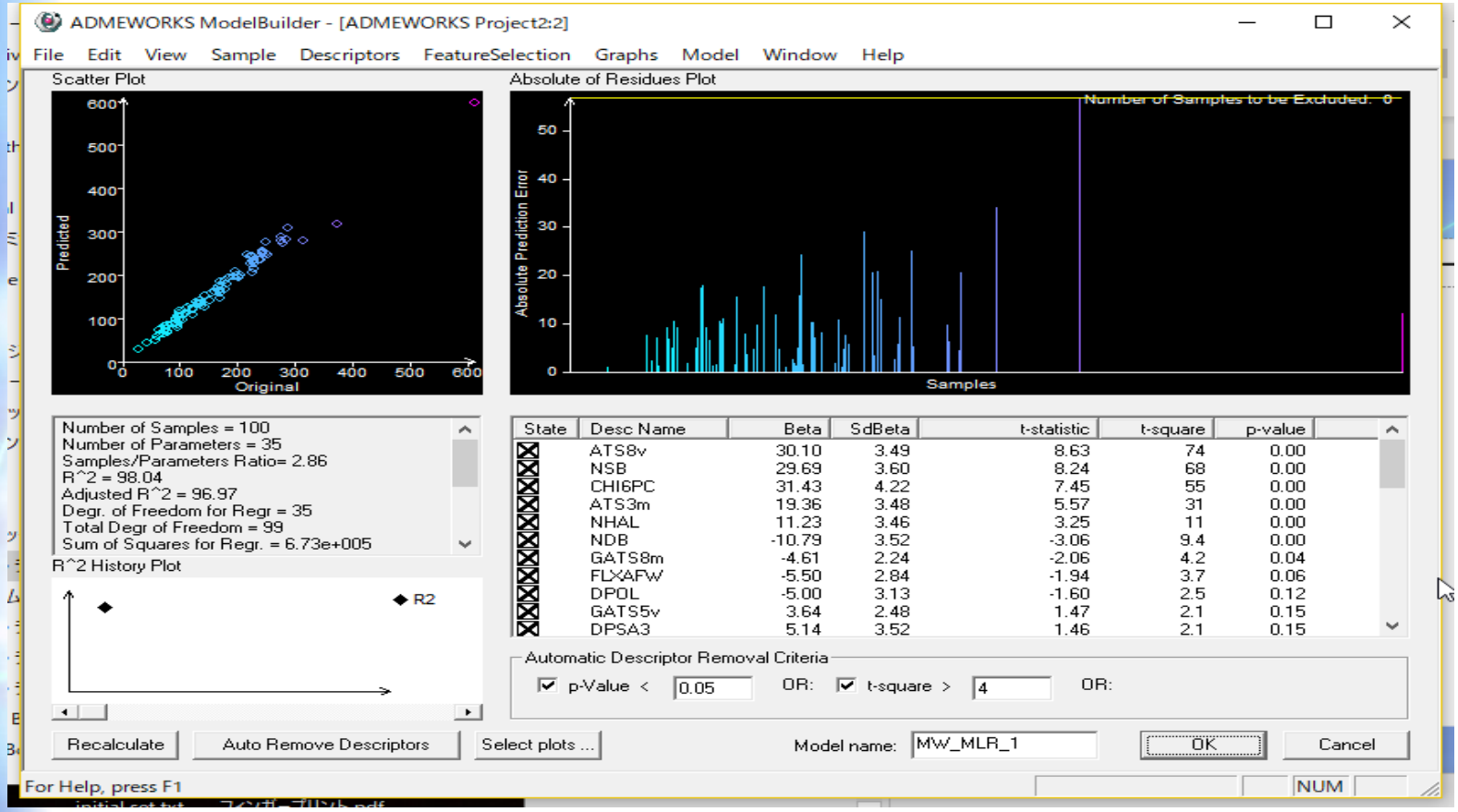
ModelBuilderの画面より

◇サンプルおよびパラメーター選択：レーダーチャート



ModelBuilderの画面より

◇サンプルおよびパラメーター選択：重回帰



ModelBuilderの画面より

□データ解析関連技術の展開

■解析原理・化合物操作・データ解析での留意事項

多変量解析／パターン認識によるデータ解析⇒宝箱・発見型アプローチ
化合物関連：一元一項対応、一次元表記、二次元表記、三次元表記
三次元の扱い問題：ローカル・グローバルミニマ対応
確率問題⇒偶然相関
フィッテイング⇒オーバーフィッテイング（過剰適合）
線形/非線形問題；
特徴抽出（パラメーター手法）；種類、特徴、限界
サンプル関連；総サンプル数、最小サンプル数、ポピュレーション（絶対数、クラス比率）
パラメーター関連：種類、単位の違い、オートスケーリング、最少パラメーター数
ニューラルネットワーク⇒中間層の数　⇒　縦の重なり
・問題点；チャンスコリレーション、非線形性
パラメーター数が多いとき：パラメーター圧縮　PCA　PLS
パラメーター数が少ない時；成功率低下、解析不能、物性式等
要因解析：パラメーターの読解力、分類力が強いパラメーター
KY法：サンプル数フリー、ポピュレーションフリー
　　二クラス分類；完全分類、　重回帰；高相関係数、高絶対係数
分類/予測⇒クロスバリデーション
外挿と内挿の違い

◇ケモメトリックス解析を保証するための最低限の制限事項

データ解析結果を保証する3段階＋1項目の保証手続き

1. 実施したデータ解析自体が正しい状態にあるか否かのチェック
   データ解析をする前の前提条件
2. データ解析手法自体が有する制限事項や適用限界等
   データ解析手法自体が適用対象や適用限界を有する
3. データ解析の結果が解析的に良好であるか、否か
   一般的には、分類率、予測率、相関係数、絶対係数等の指標を用いる
4. その他
   プログラムの制限事項、化合物を扱うときの特殊事項、その他
   個々のプログラム上での制限事項、バージョンの違い、化合物に起因する特殊問題等への注意が必要

◇ケモメトリックス解析を保証するための最低限の制限事項

1．実施したデータ解析自体が正しい状態にあるか否かのチェック
　データ解析をする前の前提条件

＊サンプル数と解析に用いるパラメーター数との比が指針となる
＊データ解析で避けなければならない以下の問題を避けることを
　保証するパラメーターである

- **過剰適合**：Over Fitting
- **偶然相関**：Chance Correlation

＊上記二問題はクラス分類や重回帰（フィッテイング）手法の総ての
　データ解析手法に適用される問題である

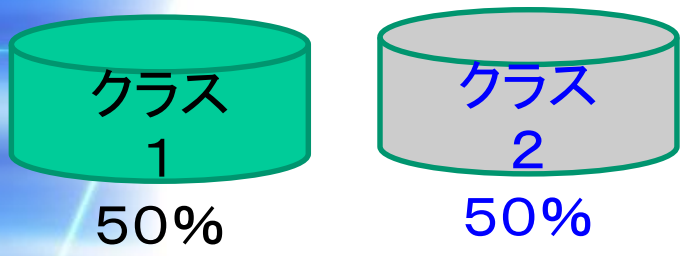
□サンプルポピュレーション

□サンプルポピュレーションが問題になるのは判別分析等を適用する場合

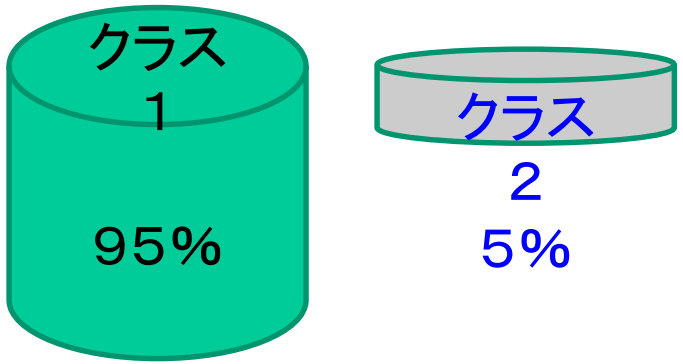
＊サンプルポピュレーションがクラス間で大きな差異がない時はデータ解析を問題なく実施できる

＊サンプルポピュレーションがクラス間で大きな違いがある場合は、データ解析結果がクラスポピュレーションの影響を強く受ける

＊クラス分類では、ポピュレーションが多いクラスの分類率が高くなる



分類率に関する寄与は略同等となる



創出される判別関数は総てクラス1に判定
しかし、分類率は95%と高い値になる

◇総サンプル数／クラスサンプル数での制限事項

In Silico Data
Miracles by the KY-methods

＊データ解析を行う時はデータ解析結果の信頼性を保証することが必要である
＊一般的にサンプル分布を基本とする統計解析ではサンプル数を多くそろえることが必要なことはよく知られている
＊多変量解析／パターン認識を行う時の最小サンプル数はどの程度必要なのだろうか
＊他の分野と異なり、薬理活性や毒性等の分野ではサンプルを多数集めることは殆どできない
＊この点で、信頼性を保ちながら多変量解析／パターン認識を行うための最小サンプル数が重要になる

＊信頼性の高いデータ解析を実施するための**最少サンプル数**は、解析に用いた**パラメーターの数**との関係で決まる
＊データ解析信頼性は以下のパラメーターを基準として設定される

2クラス分類の信頼性

$$\frac{\text{総サンプル数}}{\text{総パラメーター数}} \geqq 4$$

重回帰の信頼性

$$\frac{\text{総サンプル数}}{\text{総パラメーター数}} \geqq 5$$

◇ケモメトリックス解析を保証するための最低限の制限事項

□ 「偶然性」問題における次元数とサンプル数との関係（一般化）

次元（記述子）が一つ増える毎に分類可能な場合の数は2倍ずつ増加する。
従って、次元数dにより定まる分割可能な場合の数Rは以下の式で示される。

$$R = 2^{d} \qquad (1)$$

この結果、次元数がNで、サンプル数が2^{N}以下の時には必ず分類出来、この分類結果は偶然により支配されている事は明白である。

一方、サンプル数がnの時、このサンプルを2クラスに分類出来る場合の数Cは単なる組み合わせ問題であり、以下の式で示される。

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

これらの項目を考慮し、与えられた記述子（次元数）dでサンプルnを2分割出来る可能性Pは

$$P = \frac{\text{サンプルnに対する2分割の場合の数}}{\text{記述子dによる2分割の場合の数}} = \frac{C}{R}$$

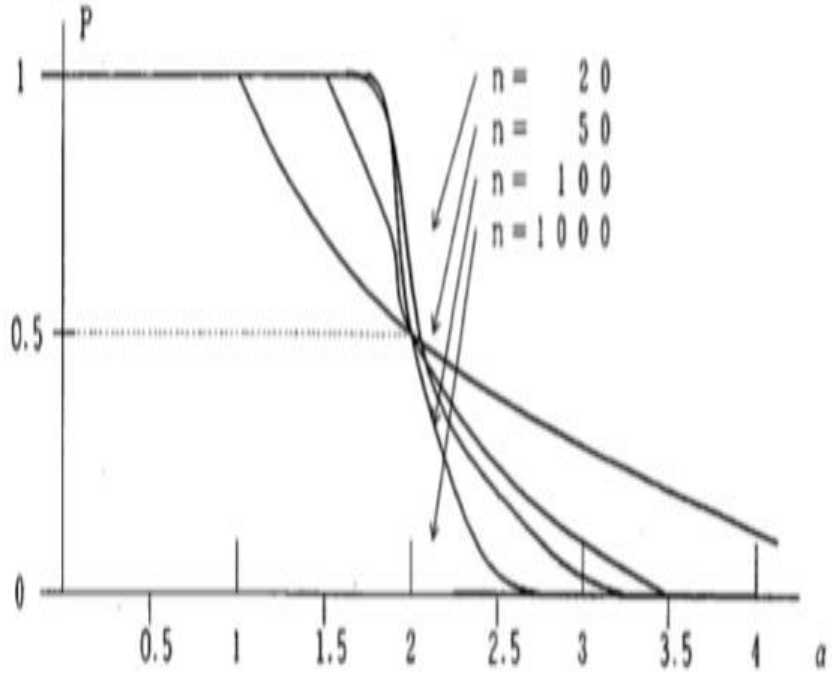


図4． 2分割の可能性に対するα（サンプル数n／次元数d）とPの関係

◇ケモメトリックス解析を保証するための最低限の制限事項

■解析信頼性の簡単な事例

□二クラス分類

例1：100サンプルを1パラメーターで100%分類 ⇒ **信頼性指標＝100**
このパラメーターはクラス分類に極めて**重要な情報を持つ**

例2：100サンプルを1000パラメーターで100%分類 ⇒ **信頼性指標＝0．1**
1000個のパラメーターはクラス分類に**重要な情報を持たない**

□重回帰（フィッテイング）

例1：100サンプルを1パラメーターで100%分類 ⇒ **信頼性指標＝100**
このパラメーターはクラス分類に極めて**重要な情報を持つ**

例2：100サンプルを1000パラメーターで100%分類 ⇒ **信頼性指標＝0．1**
1000個のパラメーターはクラス分類に**重要な情報を持たない**

◇ケモメトリックス解析を保証するための最低限の制限事項

■解析信頼性の簡単な事例

□二クラス分類

例1：100サンプルを1パラメーターで100%分類

ポジかネガの100サンプルの可能な組み合わせの場合の数は2^{100} となる。
パラメーターが二値パラメーターであれば、表現できる場合の数は2。
従って、1パラメーターで100サンプルを二分割できる確率は、
$P = 2 / 2^{100}$ で、殆ど0であり、**チャンスコリレーション（偶然相関）**はない。

例2：100サンプルを1000パラメーターで100%分類

ポジかネガの100サンプルの可能な組み合わせの場合の数は2^{100} となる。
パラメーターが二値パラメーターであれば、表現できる場合の数は2^{1000} である。
従って、1パラメーターで100サンプルを二分割できる確率は、
$P = 2^{1000} / 2^{100}$ で、Pは極めて大きい値となる。
例2の条件下では、100サンプルの100%分類は確実に実現する。即ち、
チャンスコリレーション（偶然相関）が発生する。

■本日のプログラム

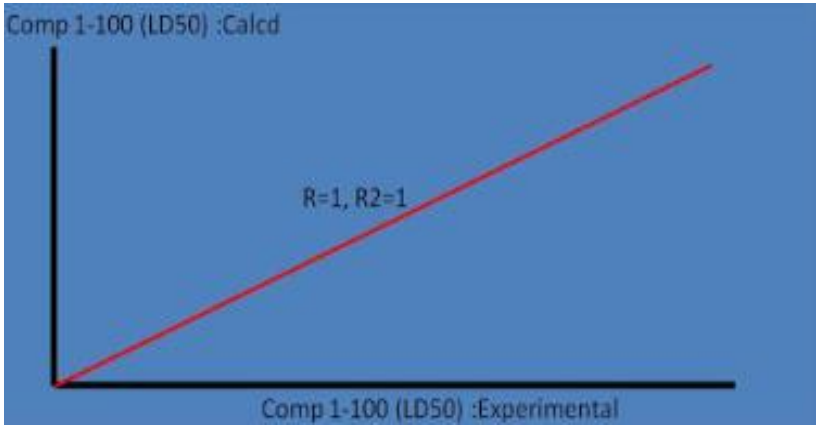
◇重回帰で相関係数や絶対係数を１にする方法

　目的変数として薬理活性のED50を用いて、100個のサンプルを用意。なお、これら100個のサンプルにはあらかじめ
1から100番までの任意のID番号を付ける。

1. **サンプルデータとして薬理活性のED50値を持つ100個の化合物を用意。**
2. **使用するパラメータとしてサンプル数と同じ100パラメータを用意します。**
3. **各パラメータは化合物のID番号の部分を1とし、残りはすべて0とします。**
4. **100サンプルのED50を目的変数、100個のパラメータを説明変数として重回帰を実行します。**

実行結果：
相関係数R=1、 絶対係数R2=1

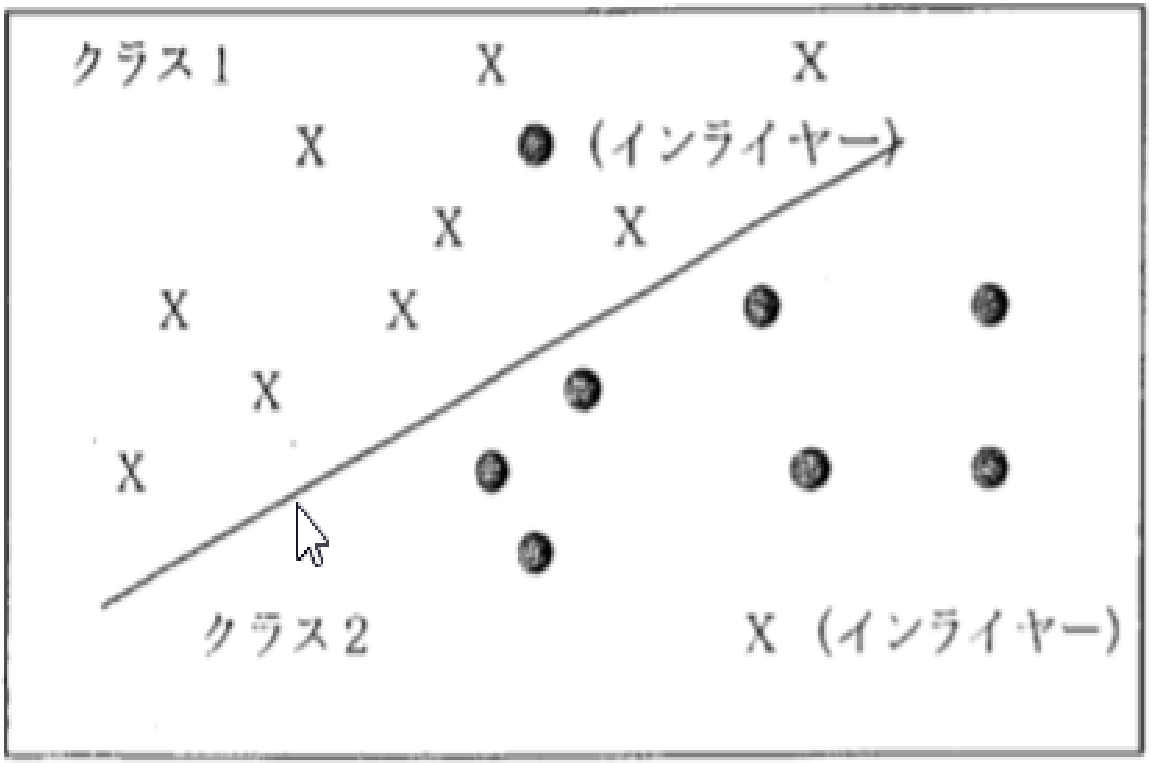
	LD50	param1	param2	xxx	xxx	param99	param100
comp1	58	1	0	0	0	0	0
comp2	132	0	1	0	0	0	0
comp3	12	0	0	1	0	0	0
comp99	728	0	0	0	0	1	0
comp100	311	0	0	0	0	0	1



◇取り出すサンプル数の限界

＊**ニクラス分類**ではご分類するサンプル（インライヤー）を取り除くことで
判別関数の予測精度を高める。
このサンプル取り出しは初期サンプル
数の約10%を上限にするとされている。

＊**重回帰（フィッテイング）**でもサンプル（アウトライヤー）を取り除くことで重回帰式の相関／絶対係数を高くする。
このサンプル取り出しはニクラス分類と同様に初期サンプル
数の約10%を上限にするとされている。



◇クラスサンプル数の限界

クラスポピュレーション　≧　パラメーター数

ポピュレーション例：
総サンプルすう100、クラス分類率100%、
でもどちらの結果を信用しますか

①クラス1；99サンプル、　クラス2；1サンプル
②クラス1；50サンプル、　クラス2；50サンプル

◇分類率と予測率

＊二クラス分類では分類率と予測率が作成された判別関数の
精度を示す指標として利用される

＊分類率はデータ解析に用いたサンプルを判定するものなので、
サンプルに困ることはない

＊予測率は予測項目の実測値が無いので、予測率を出すことは出来ない

＊予測率算出には、クラス既知のサンプルを使って仮の値を出すことが出来る

＊一般的にはクロスバリデーションと呼ばれており、様々な手法が展開されている

＊“Leave N Out”法が最も良く利用されている
クラス既知のサンプルのなかからN個のサンプルを取り出す。このNサンプルを
クラス未知とし、残る（T－N）個のサンプルを用いて予測モデルを構築し、この
予測モデルを用いて取り出されたNサンプルについて予測を行う。この手順を
総てのサンプルについて繰り返して、全体の予測値を出す。
この時、パラメーターセットは同じものを利用する。

◇分類率と予測率

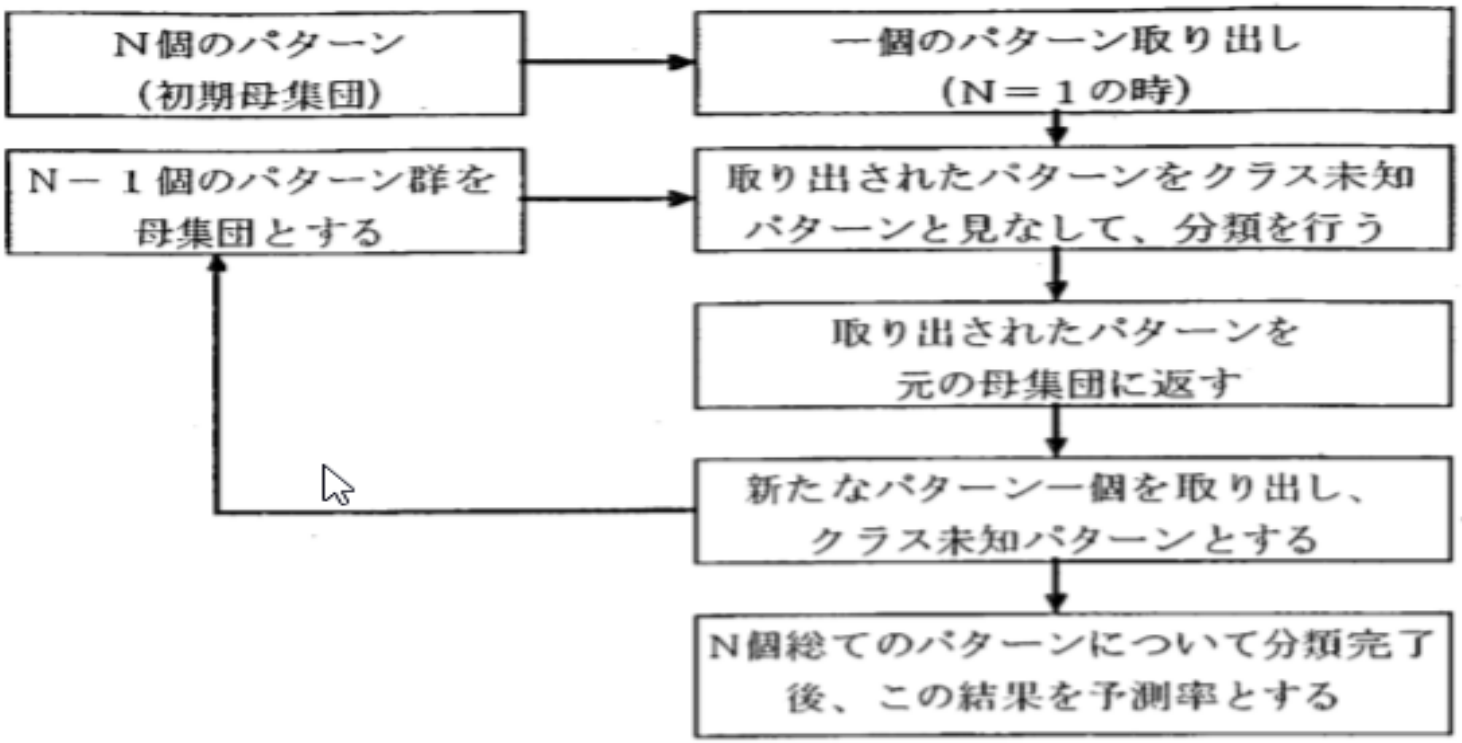


図　．　リーブワンアウト法による予測率計算の流れ図

◇総サンプル数／クラスサンプル数

＊総サンプル数は使ったパラメーターの数で決まる
＊データ解析時に用いるサンプル数を4（二クラス分類）か5（重回帰）で割った値のパラメーターを用いてデータ解析を行えば、データ解析の信頼性が保証される

100サンプル用いた場合、解析に用いたパラメーターが

2クラスタ分類	：	25**以下**	⇒	解析信頼性が保証
		25**以上**	⇒	解析信頼性が低い
重回帰	：	20**以下**	⇒	解析信頼性が保証
		20**以上**	⇒	解析信頼性が低い

結論：最小サンプル数の縛りはない
用いたサンプル数でパラメーター数の縛りが発生

◇総サンプル数／クラスサンプル数

＊クラスの最少サンプル数は使えるパラメーターの数と直結する
＊解析信頼性を保ってクラス分類を行う場合、総サンプル数よりも最小クラスのサンプル数がパラメーターの利用制限に強い影響を及ぼす

100サンプルでクラス1が90でクラス2が10の場合
2クラスタ分類　：　10**以下**　⇒　解析信頼性が保証
10**以上**　⇒　解析信頼性が低い

結論：クラスポピュレーションの最小クラスのサンプル数
この値を超えた数のパラメーターは使えない

＊パラメーターを多くしたい時は、最小クラスのサンプル数を増やす＊

◇線形／非線形問題

＊線形および非線形に関する問題はデータ解析を行う時に常に考慮すべき事項である

＊データ解析の**外挿性や内挿性**に関係する

＊**過剰適合**を起こしている場合、非線形解析の方が発生しやすい

＊データ解析の簡易度で考える場合、線形解析よりも非線形解析の方が成功率は高い

＊判別分析や重回帰を行う場合、分類率や相関係数、絶対係数値は線形解析よりも非線形解析の方が高い／良好な結果を導きだす

＊N次元サンプル空間での問題を考えた場合、

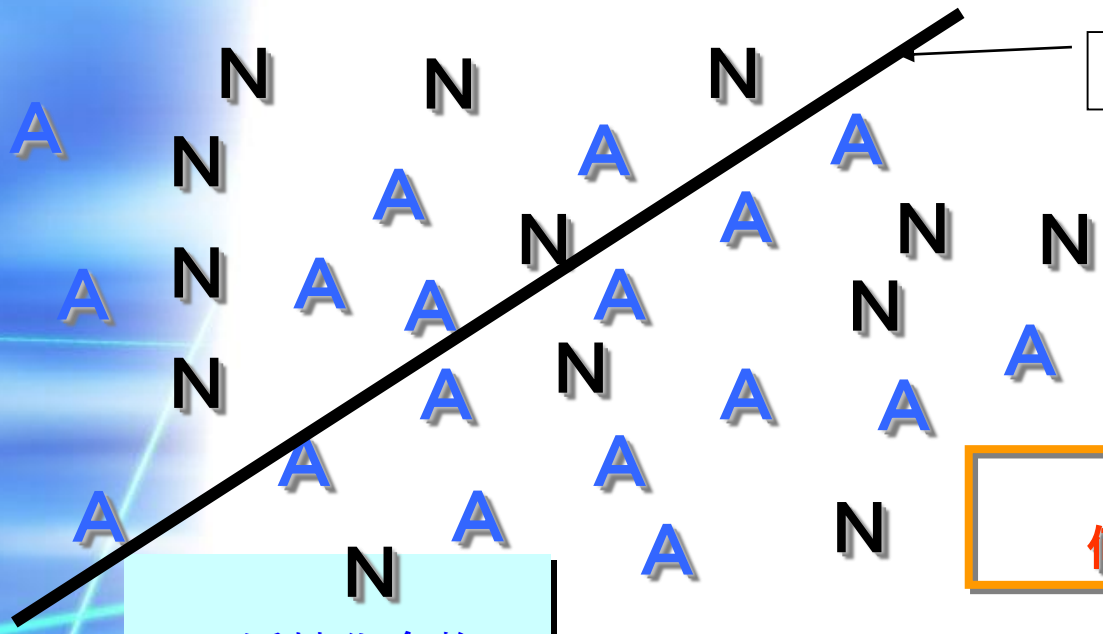
・線形での解析は**サンプル空間を作り直して**分類や重回帰を実施

・非線形解析の場合、**サンプル空間の形に合わせて**判別関数や重回帰式を算出

◆線形分類によるクラス分類

＊分類できない場合

■ノイズデータを含んだN次元パターン空間

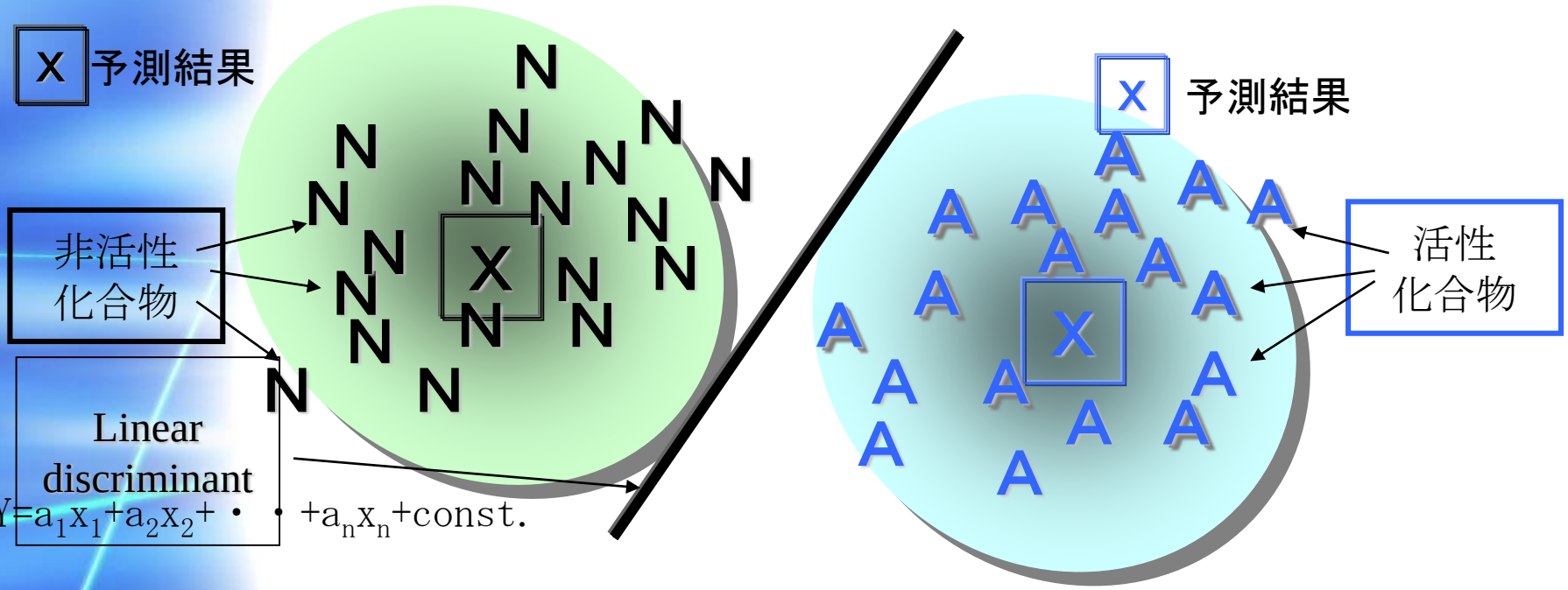


**目的活性とパターン空間に
何ら相関関係が見出せない**

◆線形分類によるクラス分類

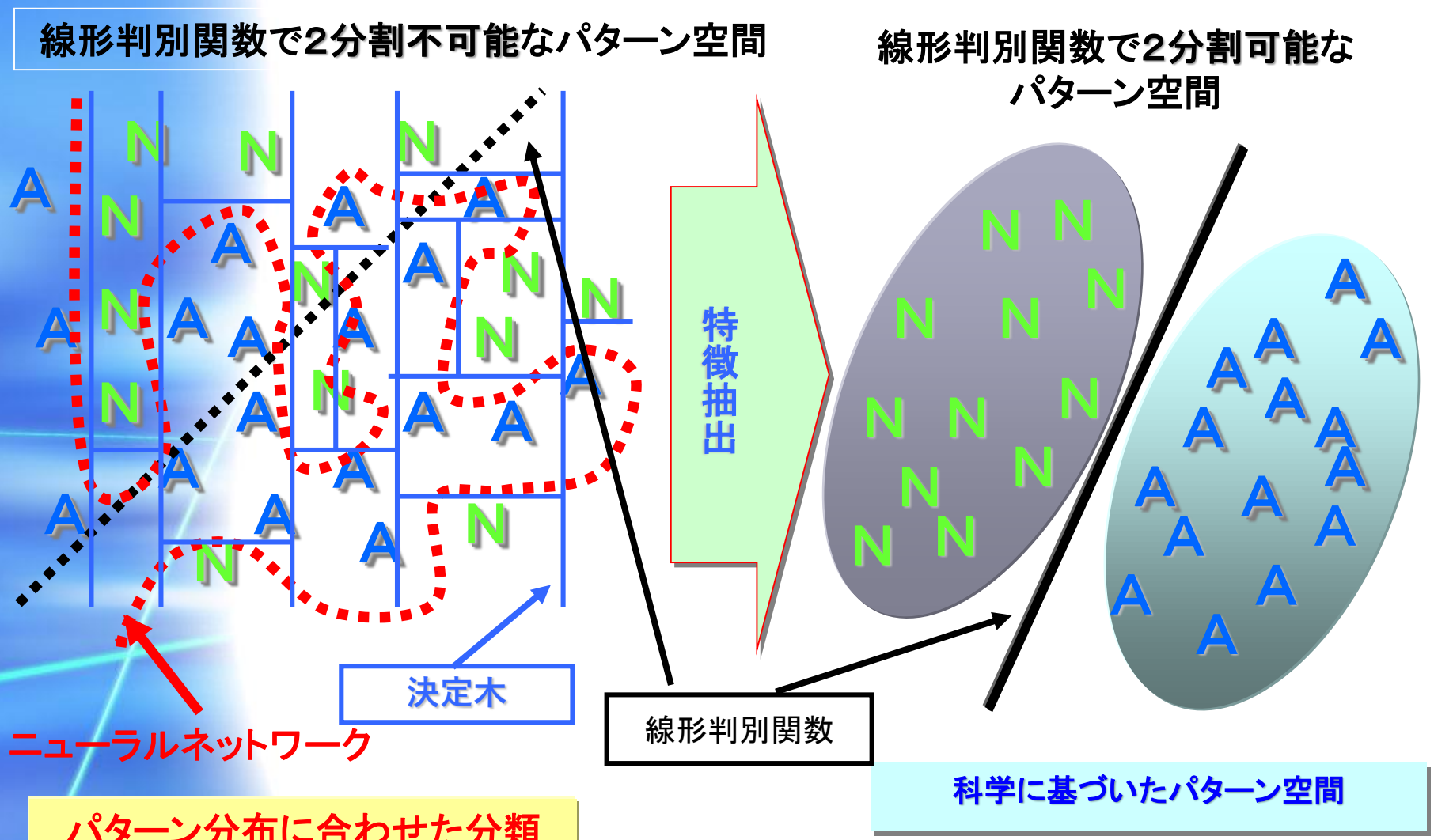
＊分類可能な場合

■目的活性と相関の高いパラメータ群により
形成されるN次元空間



＊活性および非活性化合物群と2分割可能なパターン空間
＊本空間を作るパラメータ群には、活性／非活性と分類するに重要な情報 （科学的根拠／相関）を持つ

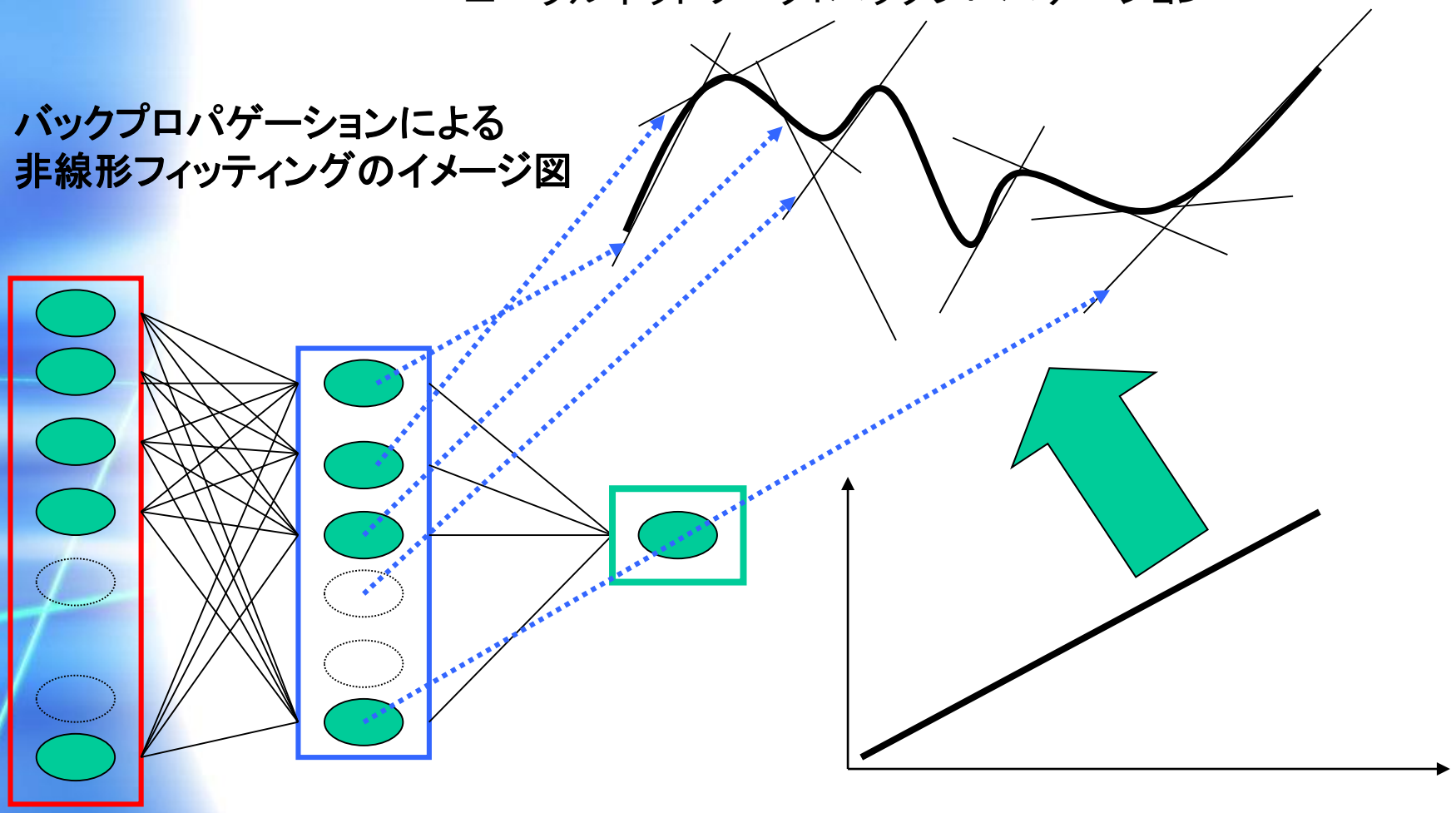
◆非線形分類によるクラス分類



◆ニューラルネットワークによる非線形分類

ニューラルネットワーク：バックプロパゲーション

バックプロパゲーションによる
非線形フィッティングのイメージ図



◇ケモメトリックス解析を保証するための最低限の制限事項

□ニューラルネットワークによる二クラス分類

ニューラルネットワークのネットワーク構造により、パラメーターが表現できる場合の数は単純パーセプトロンと比較して極端に大きい値となる。
例：100サンプルの二クラス分類で、ニューラルネットワークで、入力パラメーター数は10とし、中間層のユニット数も10とした場合の100%分類の可能性は以下となる。

ポジかネガの100サンプルの可能な組み合わせの場合の数は2^{100} となる。
一方、パラメーターが二値パラメーターであれば、ニューラルネットワークで表現できる場合の数は入力層で2^{10}。これに中間層で生起される場合の数は、$(2^{10})^{10}$
従って、1パラメーターで100サンプルを二分割できる確率は、

$P = (2^{10})^{10} / 2^{100}$ で、極めて大きな値となる。

従って、**チャンスコリレーション（偶然相関）**は必然的に発生する。

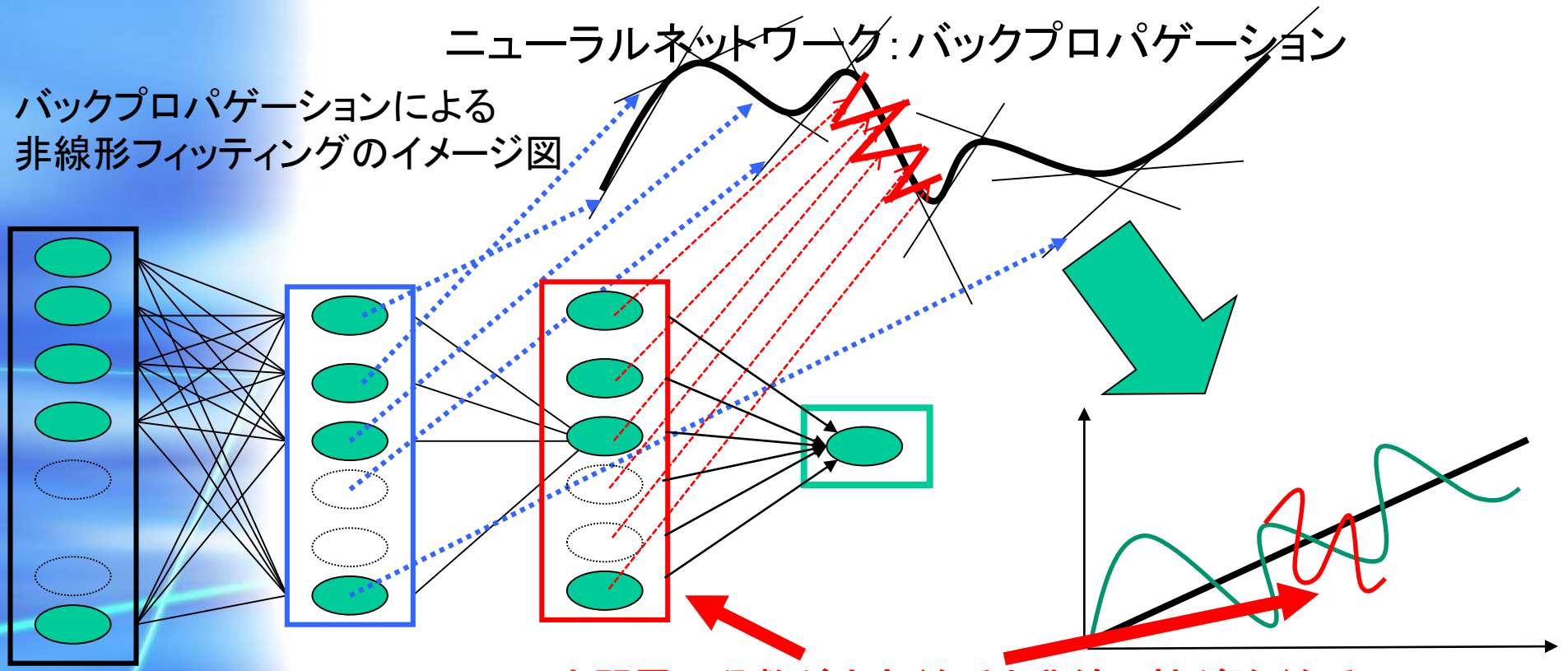
ニューラルネットワークの二クラス分類は**中間層のユニット数**が大きくなると場合の数が拡大し、**中間層の層数**が拡大するとさらに場合の数は急激に拡大する。
最近の**深層学習**を行う**多層のネットワーク構造**では、**チャンスコリレーション**の影響を少なくするべく、学習用サンプル数を極めて大きくすることが必要となる。

□機械学習型人工知能

・パーセプトロンとニューラルネットワーク

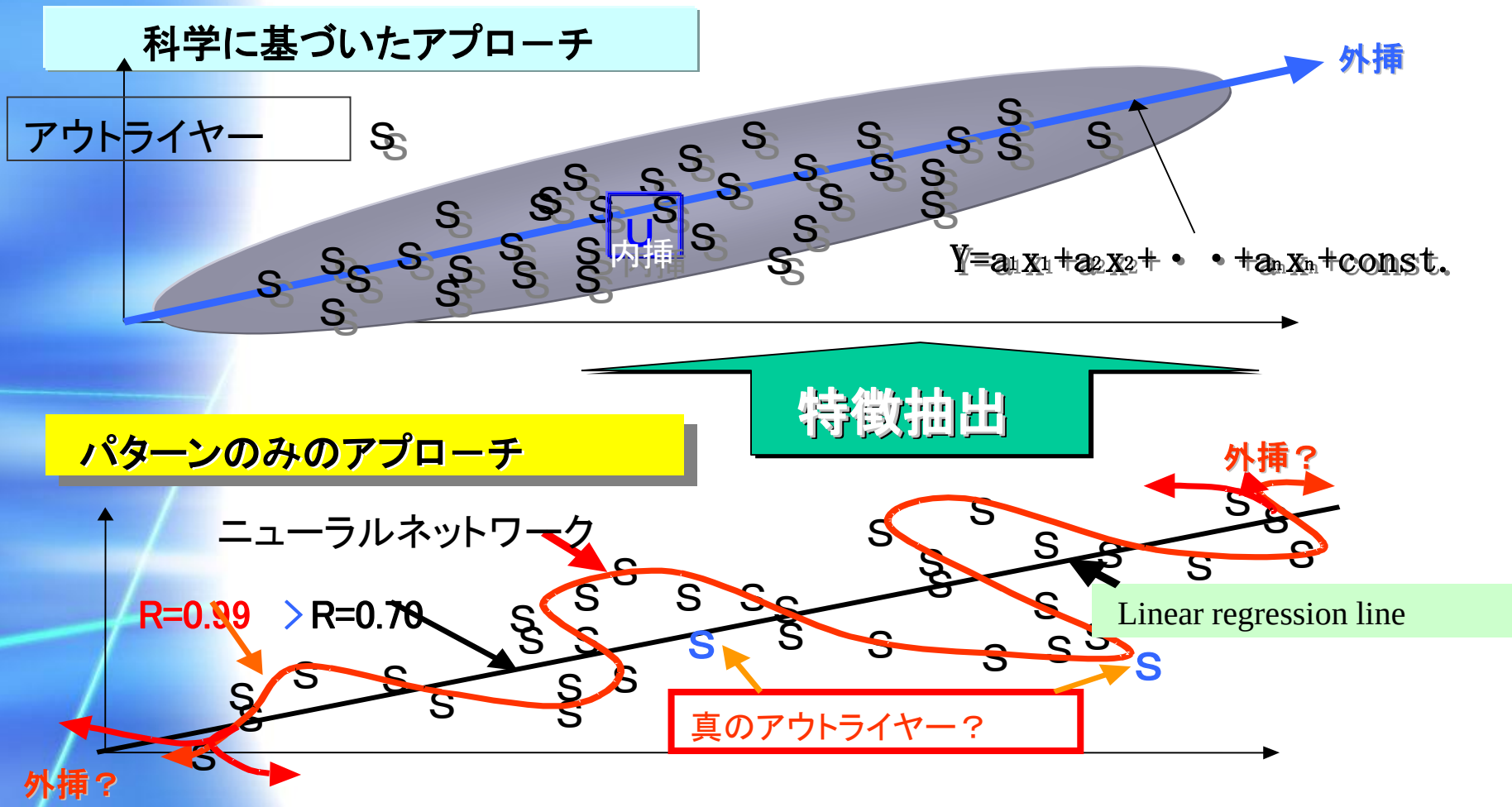
ニューラルネットワーク：バックプロパゲーション

バックプロパゲーションによる
非線形フィッティングのイメージ図



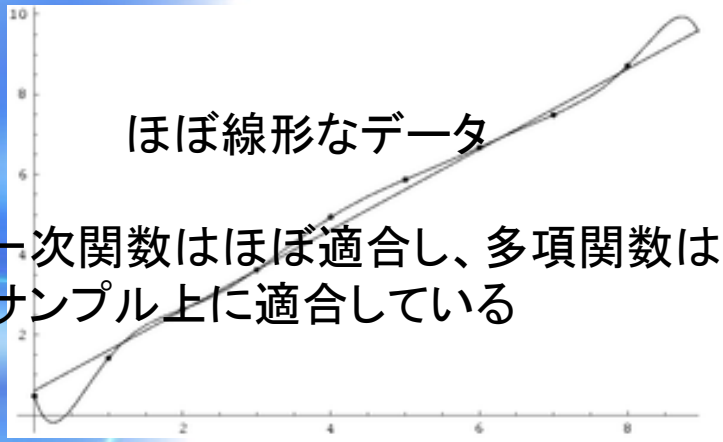
中間層の段数が大きくなると非線形性が強くなる
中間層の層数は急激な場合の数の拡大を伴う

◆線形および非線形フィッテイング



◇過剰適合/過学習

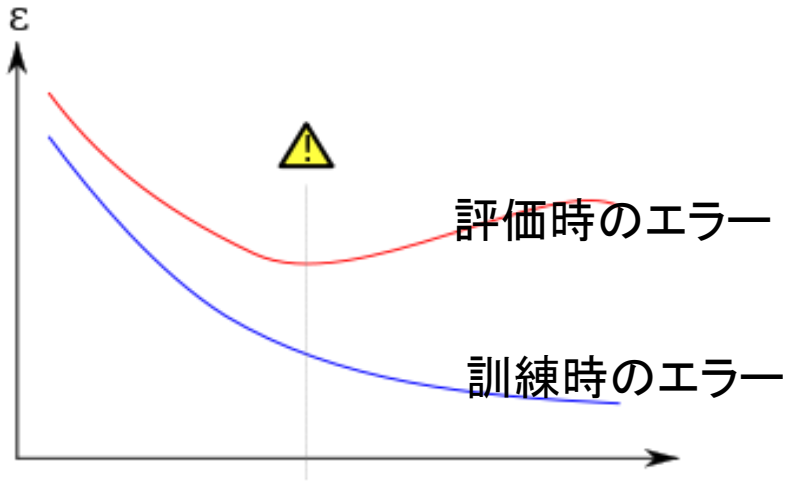
過剰適合（かじょうてきごう、英: Overfitting）とは、統計学や機械学習において、
訓練データに対して学習されているが、未知データ（テストデータ）に対しては
適合できていない、汎化できていない状態を指す。汎化能力の不足に起因する。



ほぼ線形なデータ

一次関数はほぼ適合し、多項関数は
サンプル上に適合している

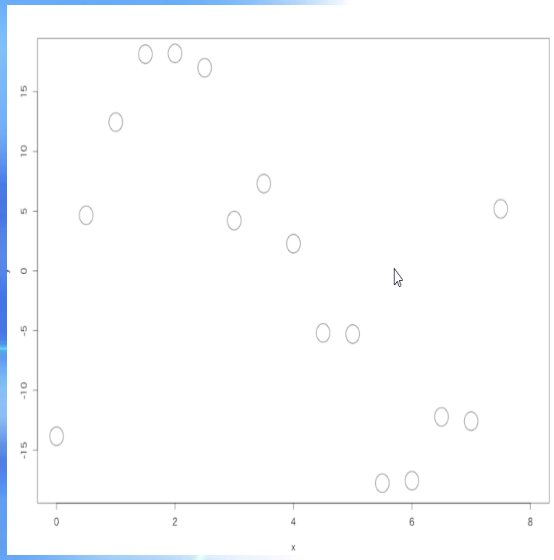
一次関数は両端で値が安定するが
多項関数は両端で値が大きく変動



ニューラルネットワークでの
過剰適合の状況

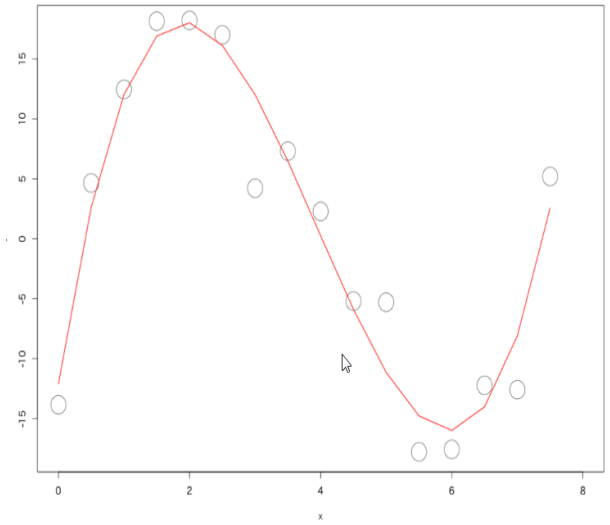
https://ja.wikipedia.org/wiki/過剰適合

◇過剰適合/過学習



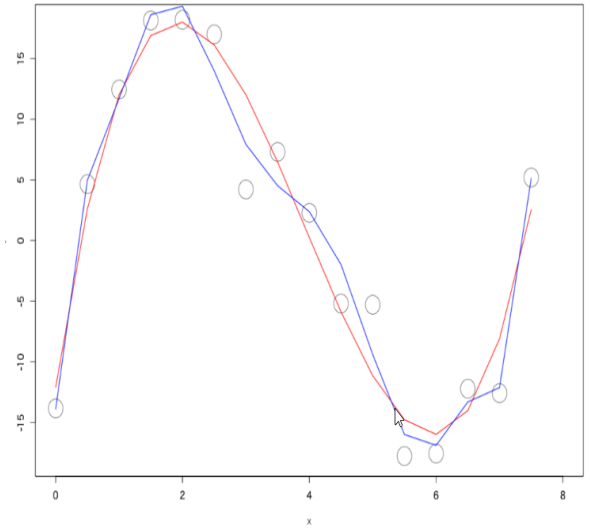
全サンプル数21
学習用16サンプル

3パラメーターを用いた重回帰
解析信頼性を保ったパラメーター数



相関係数：０．９６８
汎化能大

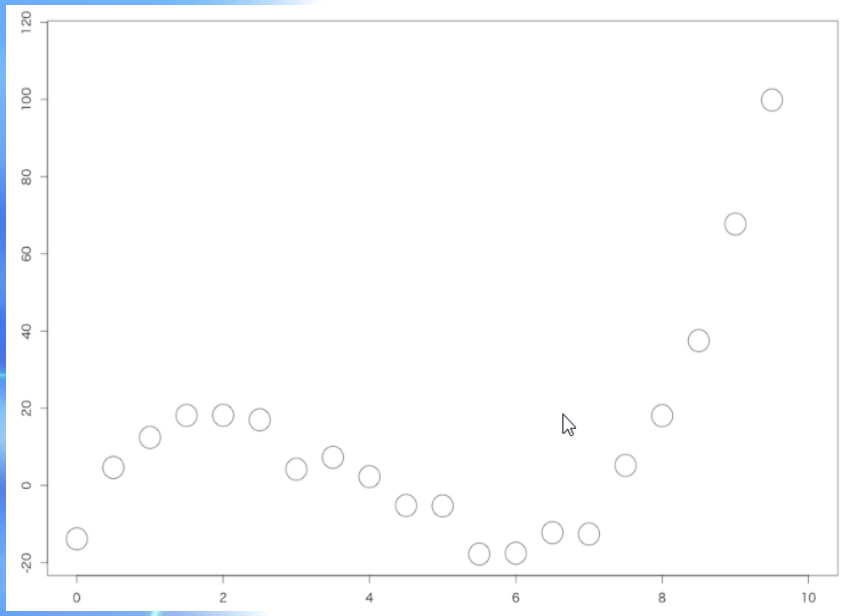
9パラメーターを用いた重回帰
解析信頼性を伴わない
パラメーター数



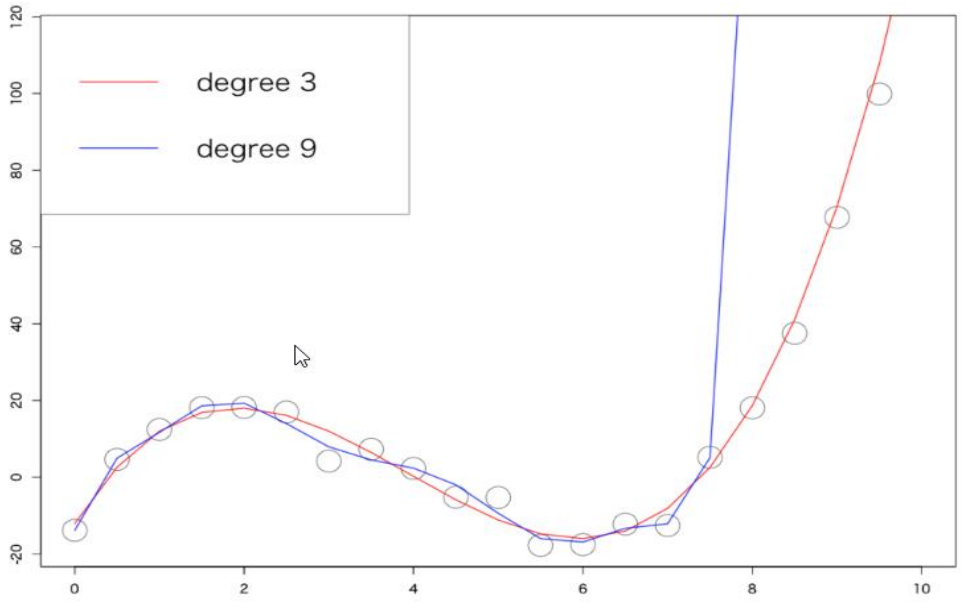
相関係数：０．９８６
過学習

https://tjo.hatenablog.com/entry/2016/04/14/190000

◇過剰適合/過学習



全サンプル21
学習用16＋テストデータ5



パラメーター3の場合と9の場合の回帰図

https://tjo.hatenablog.com/entry/2016/04/14/190000

□創薬での展開例

◇創薬時における化合物と種々特性の関係

創薬への**インテグレーテッド概念**の提案

◇創薬時における「並列創薬」の提案

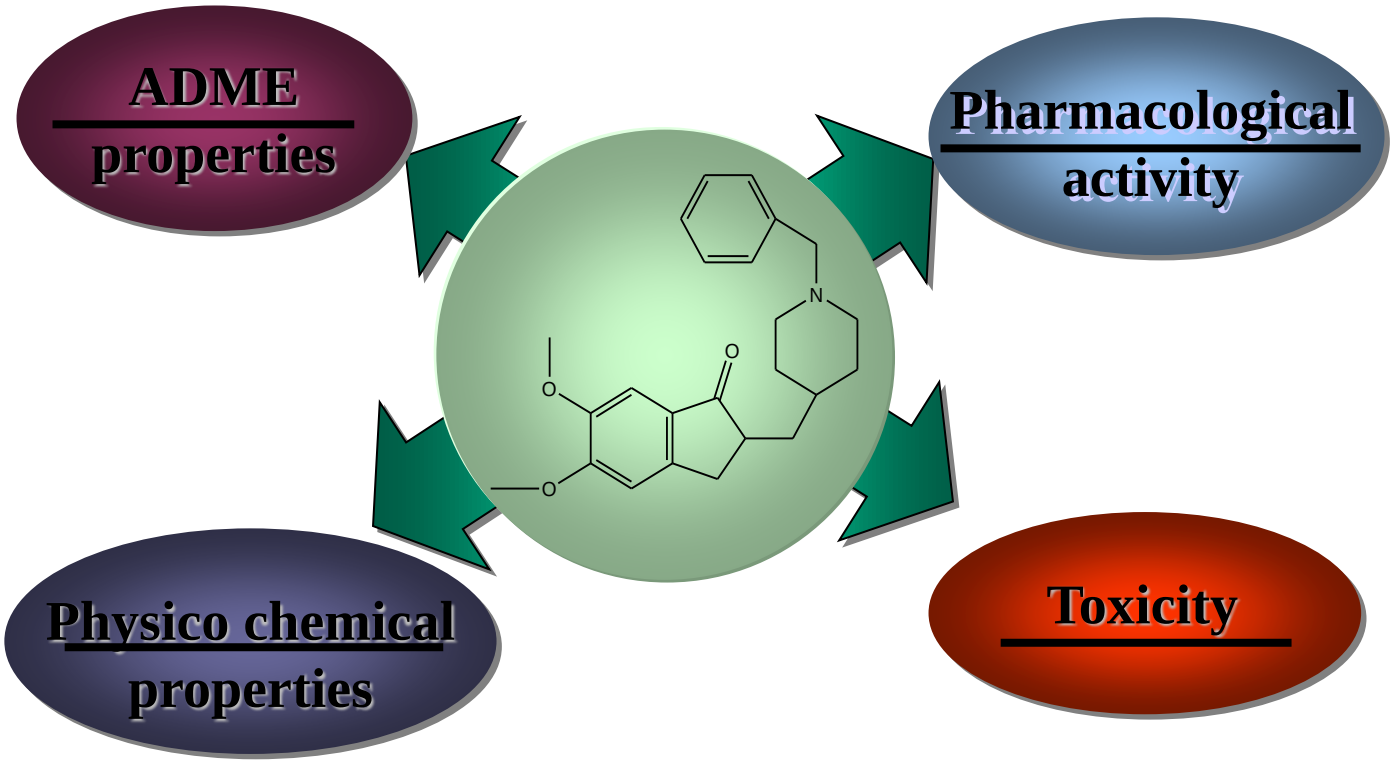
上記**インテグレーテッド概念**を実施するインシリコ手法として「**並列創薬**」を提案し、システム上で実施

◇化合物と種々特性の関係：インテグレーテッド概念

■化合物関連の様々な研究や技術は何らかの基本概念を基に開発／展開される
■新たに基本概念を設定
■新たな基本概念に従った化合物関連の様々な手法が展開可能となる
■この新たな基本概念として「**インテグレーテッド概念**」を提案した

化合物は
構造式が決まれば
総ての特性が決定する

□インテグレーテッド概念図



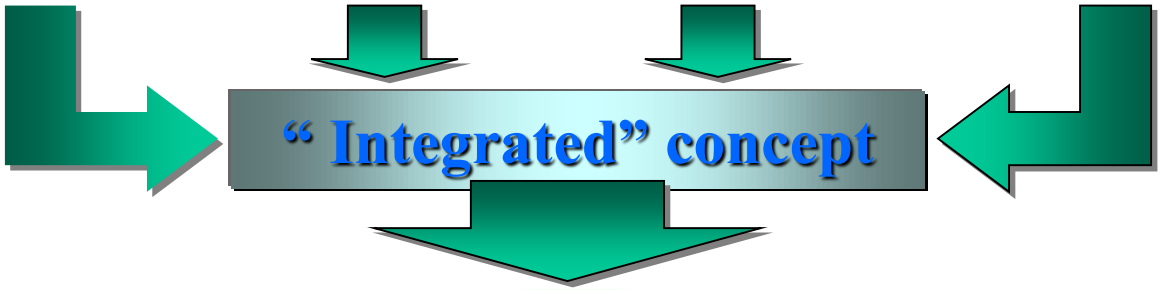
Relations between compound structure and various properties.

□インテグレーテッド概念図

‘Integrated’ concept for drug development

Activity + ADME + Toxicity + Property

All drug properties shall be considered at the same time



“ Integrated” in silico screening & drug design

The basic structure of ‘Integrated’ conception.

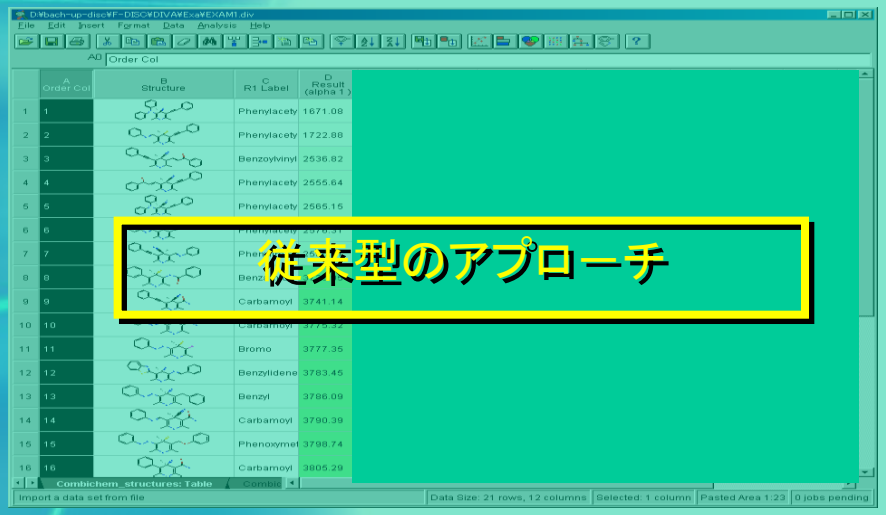
□インテグレーテッド実施例

◇複数の特性を同時に評価しながら化合物をスクリーニングする
　予め複数の予測モデルを用意し、これらの複数の要因について同時に予測／評価する

“インテグレーテッド”インシリコスクリーニング

1次元スクリーニング

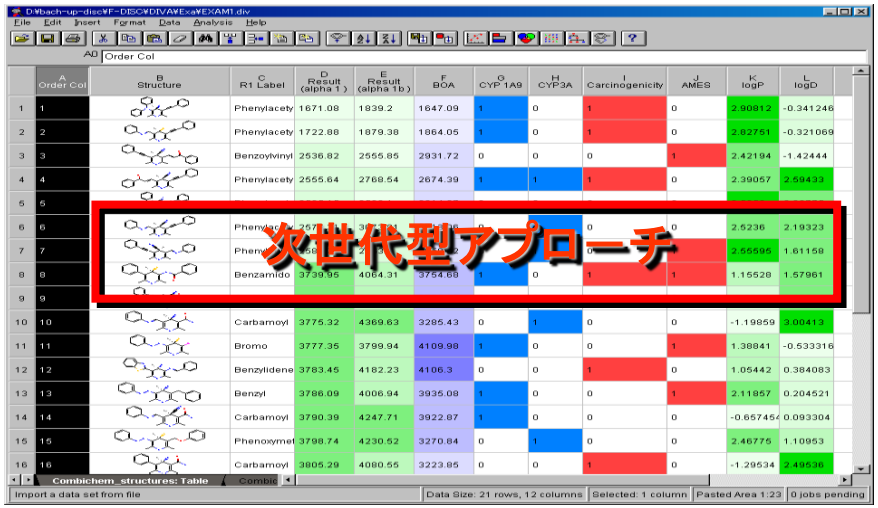
1薬理活性／化合物



2次元スクリーニング

複数の薬理活性
複数のADME特性
複数の毒性
複数の物性

／ 化合物

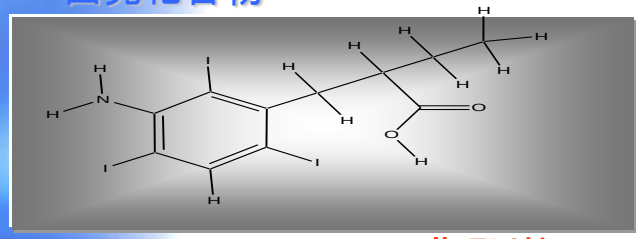


□インテグレーテッド実施例

“インタラクティブ（リアルタイム）”ドラグデザイン

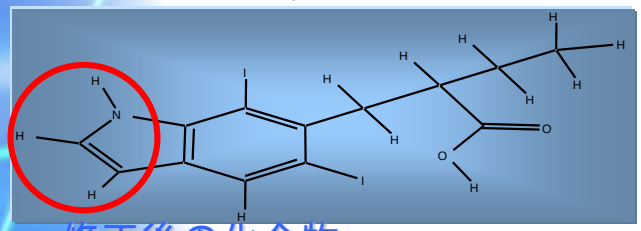
＊構造修正による薬理活性／ADME／毒性／物性変化の
即時チェックによるドラグデザインの実施

出発化合物



構造修正

薬理活性
ADME
毒性
物性



修正後の化合物

antibacterial	Antiinfra mmatory	anticancer	・・・・・	pesticide
carcinogenicity	Ames test	LD50	・・・・・	others

ADME and properties

Caco－2	BBB	CYP
LogP	pKa	$LogD_{7.4}$

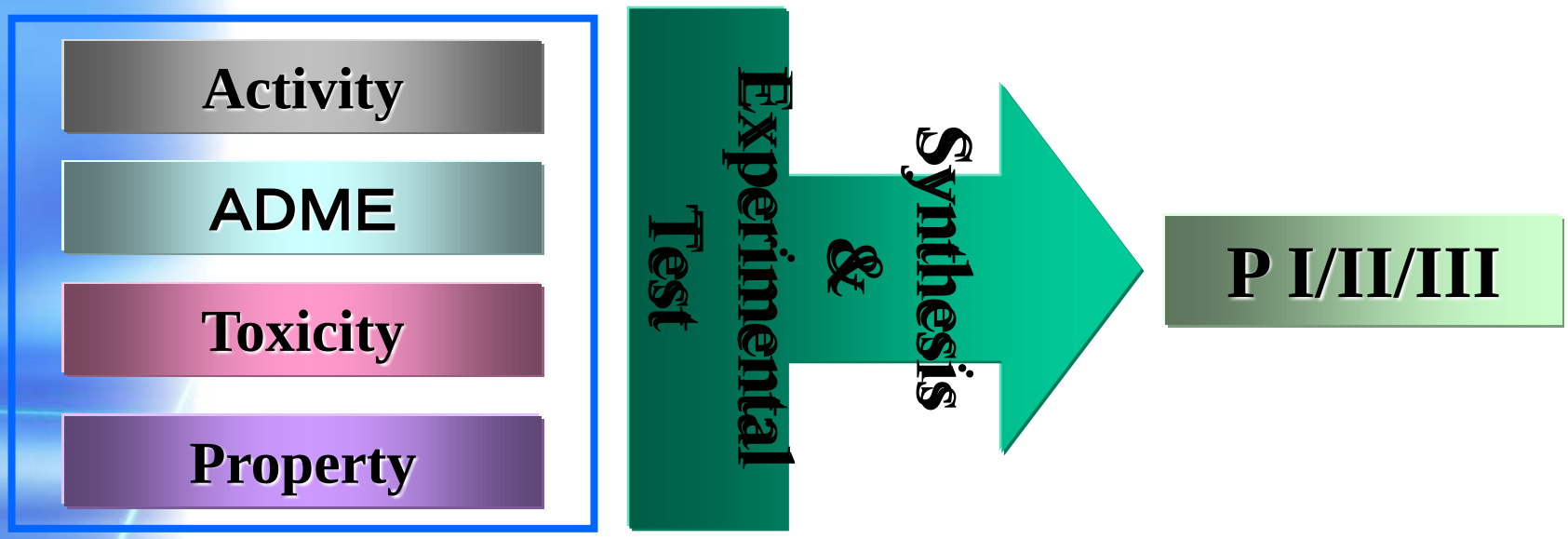
antibacterial	Antiinfra mmatory	anticancer	・・・・・	pesticide
carcinogenicity	Ames test	LD50	・・・・・	others

場所や時間を問わず、思いついた時に
構造式を書くだけで必要な情報が得られる

複数の予測モデルを用意し、構造式を変更するとリアルタイムで予測値が変わる。
この変化を見ながら、化合物構造式を対話型でデザインする

□並列創薬概念図

‘Parallel Drug Design’

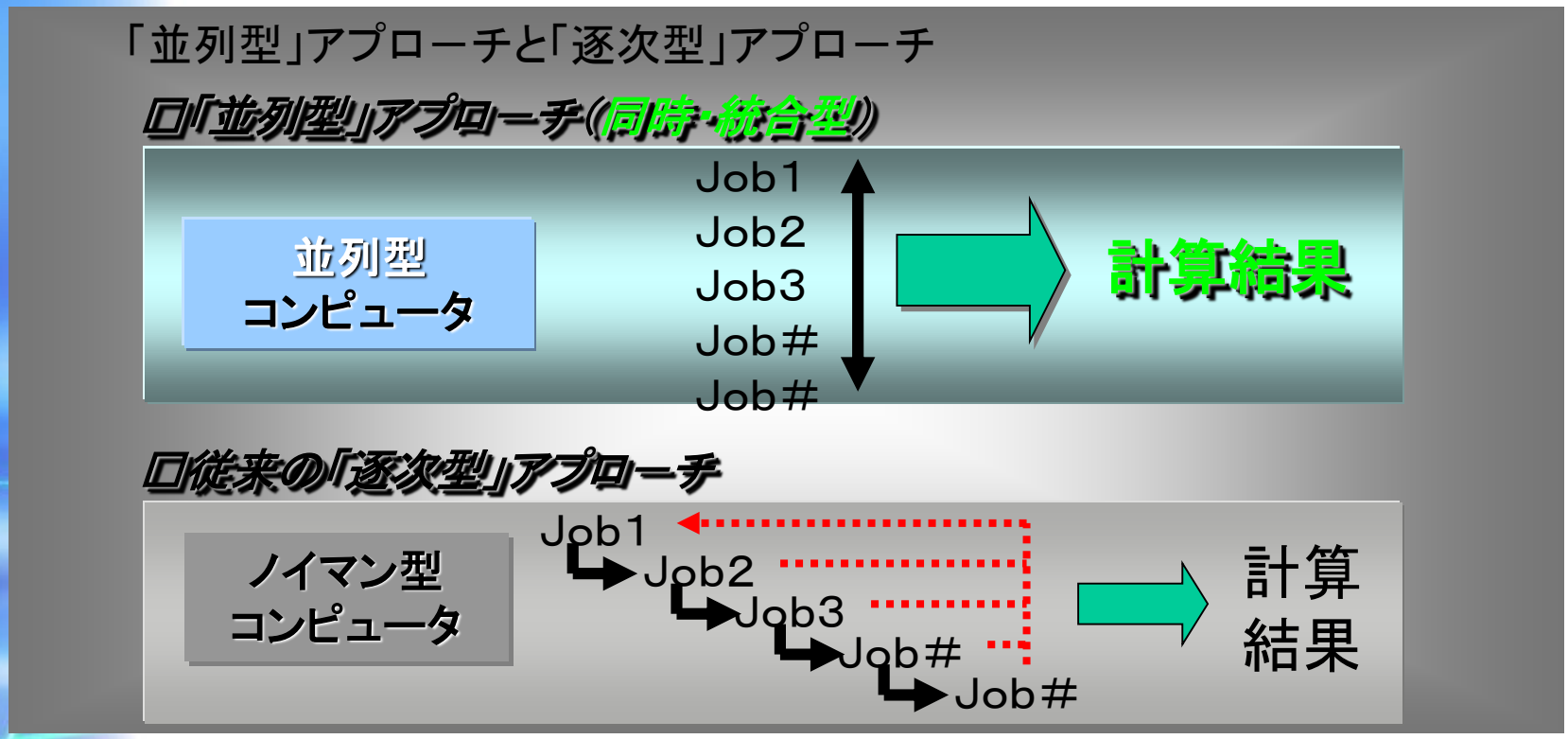


Basic flow of the ‘Parallel Drug Design’.

並列創薬:
薬理活性／ADME／毒性／物性等の諸特性を同時評価しつつ化合物デザイン実施

□並列創薬概念図

コンピュータ技術に例えた比較



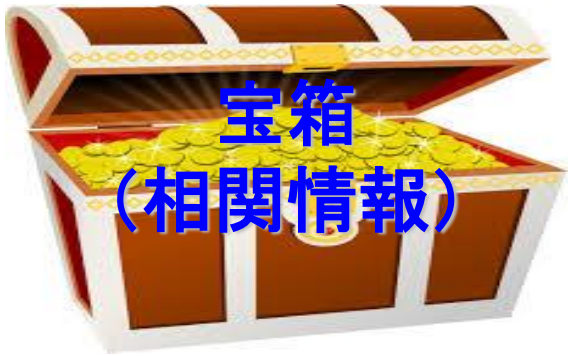
並列型アプローチ:
インテグレーテッド概念で提案する様々な化合物デザイン手法次世代型アプローチ

逐次型アプローチ:
従来型で、特性単位で値を決定し、残る特性を順番に決定してゆく

◇構造－活性/毒性/ADME/物性相関、メタボロミクス、機器スペクトル解析

化合物データ
発現プロフィールデータ
機器スペクトルデータ
画像データ
数値パラメータ

カテゴリー変数
離散変数
連続変数



種々の解析手法

判別分析
クラスタリング
マッピング
フィッティング
その他

薬理活性
毒性
副作用
代謝
分解性
光分解性
その他

◇構造－活性相関への展開

ケモメトリックス解析による構造－活性相関は関連技術の発展や開発により
様々な適用パターンで急速に展開されている

①構造－活性相関（Hansch-Fujita法）
データ解析手法として重回帰、パラメーターはHansch-Fujitaパラメータ

②ケモメトリックスによる構造－活性/毒性/ADME/物性相関
データ解析手法はケモメトリックス、パラメーターは多種利用

③バーチャルスクリーニング
化合物は化合物データベースより、パラメーターは種々

④インシリコHTSスクリーニング
HTS結果の化合物、パラメーターは種々

⑤ドラグリポジショニング
種々薬理活性予測モデル、パラメーターは種々

◇バイオ関連解析への展開

バイオ関連研究分野でも多変量解析／パターン認識の展開による
様々な解析研究が展開されている

①遺伝子（ゲノム）解析

遺伝子配列をターゲットとしてホモロジー（相同性）検索、モチーフ検索

②発現プロフィール解析

マイクロアレイやシーケンシングにより、ある個体の器官、組織、細胞ごとに調べられた遺伝子発現の全体的な様子で、理想的には全遺伝子の発現量として表現される。

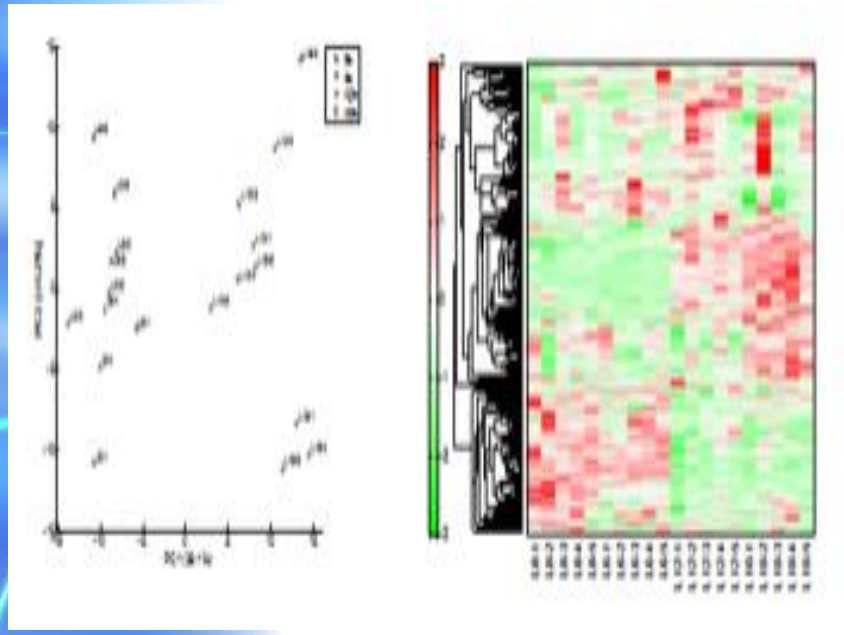
https://www.weblio.jp/content/発現プロファイル

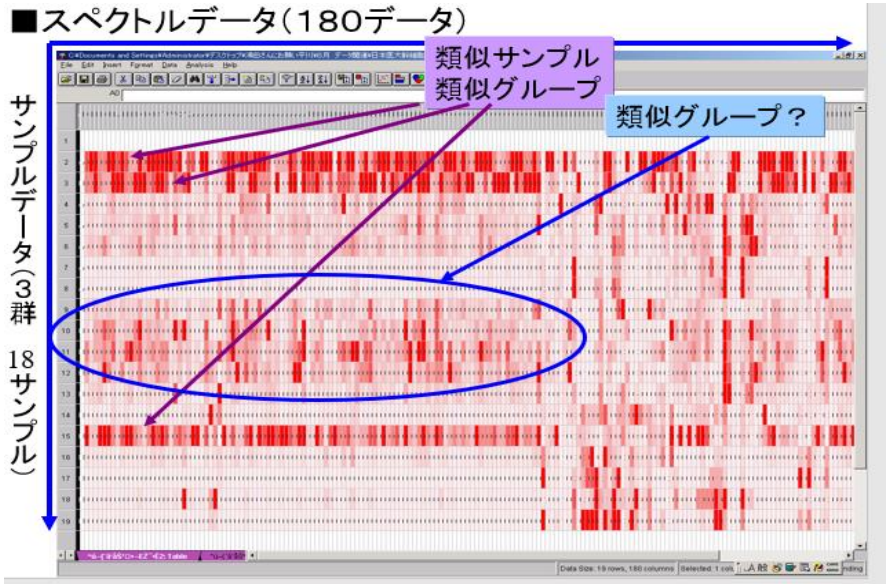
③SNPs解析

SNP(1塩基多型）を遺伝子中より発見するアプローチ。
この解析には統計手法が適用される

◇メタボロミクス解析への展開

メタボロミクス（Metabolomics）あるいはメタボローム解析（Metabolomic analysis）とは、細胞の活動によって生じる特異的な分子を網羅的に解析することである[1]。
メタボロームという語は、ある生物の持つ全ての代謝産物（メタボライト）を表す。伝令RNAの発現データやプロテオームの解析だけでは細胞で何が起こっているのか分からないが、メタボロームのプロファイルは細胞のある瞬間の生理を明らかにすることができる。





幹細胞のNMRデータを用いた解析

http://catalog.takara-bio.co.jp/jutaku/basic_info.php?unitid=U100009204