KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Politechnika Wrocławska

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

ROZWÓJ POTENCJAŁU I OFERTY DYDAKTYCZNEJ POLITECHNIKI WROCŁAWSKIEJ

Wrocław University of Technology

Medicinal Chemistry

Roman Gancarz

MATHEMATICAL METHODS IN DRUG DESIGN

Wrocław 2011

Projekt współfinansowany ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego

Wrocław University of Technology

Medicinal Chemistry

Roman Gancarz

MATHEMATICAL METHODS IN DRUG DESIGN

Wrocław 2011

Reviewer: Wacław Sokalski

ISBN 978-83-62098-43-9

Published by PRINTPAP Łódź, www.printpap.pl

Table of contests:

Preface

The rational drug design takes advantage upon the knowledge of several disciplines particularly organic and inorganic chemistry, physical chemistry, biochemistry, pharmacology from one site and many mathematical methods. The last are involved in the definition of the relationship between physical properties of the drug and its potency.

The idea behind such an approach is to extract the most important features and relations from a complex set of the available data, physicochemical and medical, in order to understand the process and formulate the next step in the synthesis of a new drug as well.

The presented manuscript is a short presentation of various mathematical approaches in the drug design area. The limited amount of space does not allow detailed descriptions of the presented methods. The student is advised to enhance the knowledge by studying the literature, especially the one suggested at the end of the manuscript.

The algorithms and the way of their application in program R are given in the supplemental material.

Author, November 2010

1. Model

A Model is a simplified representation of a real system. It might be verbal, represented by a picture, it could be model in scale as well as mathematical. The last one may be in a form of a statistical model, differential equation and others. The model describes the essence of the system, it omits less important features in order to be clear. Models schematize and have always less information than the object it represent. It simplifies the object in order to give attention to only selected features. A useful model should be characterized by the following features: m*emorable, simple, self-consistent, not contradictory, powerful*, f*lexible*, *stable to small errors* .

The general rule is that simplification should be as simple as possible but nothing more. The model may be:

- **Iconic when r**esembles the object but not function
- **Analogic when r**esembles function but not shape
- **working** model when it operates on the same physical principles as its object.
- **Pictorial**
- **Verbal**
- **Mathematical**
- **computer model** attempts to simulate an abstract model of a system by a computer algorithm.

Mathematical modelling has become applied in many systems like physics, chemistry and biology, as well as in economics, psychology. The goal is to get a tool which is then useful for the prediction of the behaviour of the real system.

The same is true for the prediction of biological properties of chemical molecules. The modelling plays a very important role in the drug design. From practice we know that one commercial drug is found after 15000 of synthesis of a new compound when it its done not rationally. To obtain such a number of compounds and performing chemical and biological analysis is time consuming and requires (spending) a lot of money.

The Rational drug design tries to find some rules which help to think in a more rational way to select new candidates for synthesis.

1.1 Examples of the most popular models in chemistry

-the scheme of a pilot plant is iconic
-the laboratory scale synthesis is an analogic model of a commercial process.
-The Periodic Table - some trends can be deduced which are hardly to be expected without such a model
-The structural formula of the second major model in chemistry is partly iconic and partly analogic. It helps to understand the properties as well as interaction.
- thermodynamics – a mathematical logical model which helps to predict quantitative consequences.
- Molecular mechanics is a model which helps to predict the behaviour and the property of molecules, mostly the shape.
- quantum mechanics – a more advanced model for the prediction of the property of the molecules
- chemical kinetics describes the behaviour of the system in time.
-QSAR, SAR –a model for prediction of biological properties of the chemical compounds
-others

1.2 QSAR-Quantitative structure-activity relationship

(QSAR) is the model in which chemical structure or its physicochemical properties are correlated with biological activity or chemical reactivity in a quantitative way. In such a model activity and properties are expressed in numbers and mathematical expression is used in a form

$$\text{Activity} = f(\text{physiochemical properties and/or structural properties})$$

The assumptions which are made are that similar molecules have similar properties (which is not absolutely true) and in most cases the assumptions are that fragmental contribution of the molecular properties is additive and depends on the structure – properties are linear. Such characteristics provides some sceptics the arguments to state that such an approach is useless . Many of them rise the question - please give me an example of a positive application of such modeling. We can however think in a little bit different way. If we know from practice that one commercial drug is found after 15000 of synthesis of new compounds then if such

modelling sugests “DO NOT synthetize those 10000, we are saving a lot of effort (for chemists, biologists) and money.

The modern QSAR or SAR belong to the moders approach – the data mining procedures. As such it benefits from the typical data mining procedures like a feature extraction, dimensionality reduction, decision trees, neural networks, pattern recognition and many others.

The oldest application of QSAR in chemistry are boiling points prediction, Hammett equation and Taft equation.

Advantages of QSAR:

- understanding the effect of the structure on the activity, in the quantitative way
- It is possible to make predictions leading to the synthesis of new analogues. Interpolation is justified, but not extrapolation
- It helps to understand interactions between functional groups

Disadvantages of QSAR:

- False correlations may arise
- data collected may not reflect the complete property space.
- physicochemical parameters used to model in most cases are cross-correlated.

2. Analysis of the datasets

In medicinal chemistry there are many cases when it is necessary to compare two or more datasets. It could be a comparison of the biochemical parameters of the patients as well as evaluation of drug design potency.

Below only the idea of the comparison methods are given. It is presented with the help of a very basic dataset and an assumption that data have normal distribution. More detailed information about the methods of calculation procedures are given in additional materials.

2.1 Normal distribution of dataset

Performing several measurements the data collection tends to cluster around the certain value. If the distribution of all data is a bell shaped described by the function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where parameters μ and σ^2 are the *mean* and the *variance* the distribution is called normal. We can normalize the distribution. Then the distribution is called **standard normal**. Then the notation is N(0,1). Such standardization allows us to use the tables for the normal distribution. There are other types of distribution like t-student, Poisson, chi-square, Bernoulli and many others. In most cases we do not know the theoretical parameters (like μ and σ) of the distribution and we have to estimate them by performing the measurements. Such estimated values are estimators. Thus mean value x_{av} is the estimator of μ and standard deviation s is the estimator of σ.

Statistics is the science of the collection, organization, and interpretation of data but also quantities (such as mean and median, standard deviation, skewness) calculated from a set of data.

For each distribution many statistics (meaning quantities) are calculated. They allow to analyze the dataset, but most of them are specific for the particular distribution and are not to be transferred to sets with other distribution.

Before calculation of the statistics one should perform a separate test in order to find out the type of data distribution.

The following description and examples are provided with the assumption that data are normally distributed.

2.2 Average, mean value

To evaluate the unknown one dimensional property like pKa, pH, we perform many measurement and estimate that value in most cases by calculating its estimator - arithmetic mean.

Arithmetic mean for the population is defined by the equation

$$AM = \frac{1}{n}\sum_{i=1}^{n} a_i.$$

for a set of data

2,3,4,3,5,3,4

the $X_{av=}24/7$

There are some other methods to calculate the central tendency of data. For example, the most frequently occurring number on a list is called the mode, (in the example above it is 3), the median is the middle number, (in the example above it is also 3. since (2,3,3,**3**,4,4,5))
Some others, are given in the table below

Table 1. Central tendency measures

Name	Equation or description
Arithmetic mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$
Median	The middle value that separates the higher half from the lower half of the data set
Geometric median	A rotation invariant extension of the median for points in R^n
Mode	The most frequent value in the data set
Geometric mean	$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$
Harmonic mean	$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$
Quadratic mean (or RMS)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}$
Generalized mean	$\sqrt[p]{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^p}$
Weighted mean	$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$
Truncated	The arithmetic mean of data values after a certain number or proportion of the

mean	highest and lowest data values have been discarded
Interquartile mean	A special case of the truncated mean, using the interquartile range
Midrange	$\frac{\max x + \min x}{2}$
Winsorized mean	Similar to the truncated mean, but, rather than deleting the extreme values, they are set equal to the largest and smallest values that remain
Annualization	$\left[\prod (1+R_i)^{t_i}\right]^{1/\sum t_i} - 1$

2.3 Standard deviation

If the data are normally distributed the standard deviation is a widely used measure of the dispersion of the data in the dataset.

It is calculated as :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

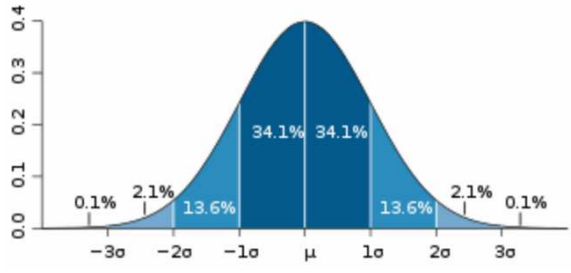
The above value is the theoretical standard deviation which in most cases is not known. We can estimate it by performing a certain number of masurements and then by calculation *standard deviation of the sample.*

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

2.4 Confidence interval

As stated above the standard deviation is a measure of data dispersion. In normal distribution the percentage of the data in the corresponding intervals are presented in the figure below

Figure 1. Plot of normal distribution



The figure above allows to explain a very important term – the confidence interval. For one dimensional data we perform measurements to get the estimation of an unknown value μ. We do it by calculation x_{av}. We have also to evaluate value σ by calculating its estimator s. The Next step is to get the idea about the error of estimation or to define an interval in which the unknown value should be at a certain level of probability (confidence). So the confidence interval is the interval in which theoretical values are expected at a certain confidence level (in most cases the confidence levels are P=99% or P=95%). From the figure above we can find that 68.2% of the data are within the interval +/ 1σ and 95.4% within the interval +/ 2σ. Exactly 99% are within +/- 2.57σ and 95 % are within +/-1.96σ.

For the data with normal distribution if we do not know σ we can replace it with its estimator s so the 99% confidence level is then defined as x_{av}+/- 2.57s and 95 % are within x_{av}+/-1.96s. When the measurements are done for less than 25 data point we can not use normal distribution tables but we have to use t-Student distribution tables and then the 99% confidence level is defined as x_{av}+/- 2.57 t(α,n) and 95 % are within x_{av}+/- 1.96 t(α,n) where t(α,n) values are taken from t-Student distribution tables for the defined confidence level α is ? and degrees of freedom n (for the mean the n=number of measurement -1).

The example is given below. (For the simplicity we use s value instead of t(α,n)).

3,3,4,5,5,4,4

x_{av} =4, s=0.82, so the expected value at the confidence level 95% is within the range

x=4+/-1.96*0.82=**4+/-1.60**

Such a result means that any value in the interval 4+/-1.60 (i.e 2.40-5.60) should be taken as correct.

2.5 Hypotesis testing

Having in mind the example given above and based on the calculated mean and standard deviation we can ask the question if the value 3.5 belongs to the above dataset at a certain confidence interval. For the procedure see additional materials however such question can be answered after looking at the above results. The Answer is yes since 3.5 is in the interval 2.40-5.60.

The above example is very simple. Most of the real problems are much more complicated. The intention of the above discussion was to help to all, even with a low mathematical background students get an idea about very fundamental aspects of measurements and evaluation of its confidence. The available statistical programs allow most of the users to calculate quite complicated problems if the person who is doing the calculation understands some basic relations.

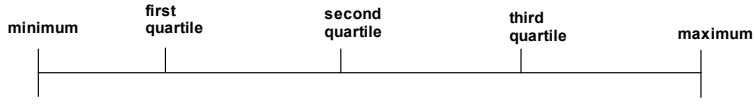
3. Comparing the two datasets

Such an analysis is performed if we want to state "that two patients do not differ from each other" or " two compounds differ in their biological activity". It can be done in two ways - parametric (when we assume that data come from a known type of the probability distribution) and otherwise nonparametric.

3.1 Box and whisker

One very convenient way of presenting several datasets is the box and whisker visualization. The whole dataset is ordered from the lowest to the biggest and divided into four equal parts-quartiles- which are numbered from 1 to 4. The F**irst quartile** (designated Q_1) cuts off the lowest 25% of data, the **second quartile** (designated Q_2) cuts off 50% of data the **third quartile** (designated Q_3) 75%

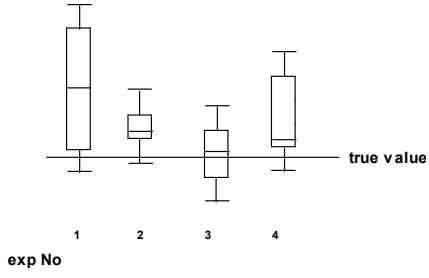
Median divides quartile 1 and 2 from 3 and 4. Then data are presented as shown below.



Boxplots are particulary useful for comparing distributions between several groups or sets of data (see Figure below for an example). Boxplots display differences between populations butno assumptions have been done about the distribution so the analysis is nonparametric.

3.2 Comparing Processes

Figure 2. Box and whisker plot



You can use a Box and Whisker plot to compare the variation and medians in multiple processes. For example, the data shown above displays a biological activity of a new drug measured for five persons. It is easy to see that person one presents more variations than the others. Person one reacts the most, whereas person four and five expresse the lowest influence on the drug action. You can see that the Box and Whisker charts are a great tool for a quick look at how several processes compare.

3.2 Correlation

In statistics, **correlation and dependence** are any of a broad class of statistical relationships between two or more random variables or observed data values. It has to be said that correlation does not imply the causation.

When strong correlation is found between the number of cancer patients and the number of doctors in any population it does not mean that the doctors are responsible for the cancer development in the patients. There are many possibilities to explain such correlation, the simplest is that people are diagnosed by some medical program in highly civilized population

characterized with a high number of doctors. So if two variables correlate no one is allowed to state that one variable is independent and another is dependent.
The most frequently used measure of correlation between two quantities is the Pearson product-moment correlation coefficient, or Pearson's correlation desribed by the formula

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

$\rho_{X,Y}$ the correlation coefficient
X and Y are variables
μ_X and μ_Y are expected values
σ_X and σ_Y are standard deviations

In general we do not know the theoretical values of $\rho_{X,Y}$, μ_X and μ_Y *as well as* σ_X and σ_Y. We estimate them by r_{xy}, x_{av}, y_{av}, s_x and s_y
If we have a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, ..., n$, then we can estimate the Pearson correlation by calculation *sample correlation coefficient*, r.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

where
x and y are the sample means
s_x and s_y are the sample standard deviations of X and Y.
It can also be written as:

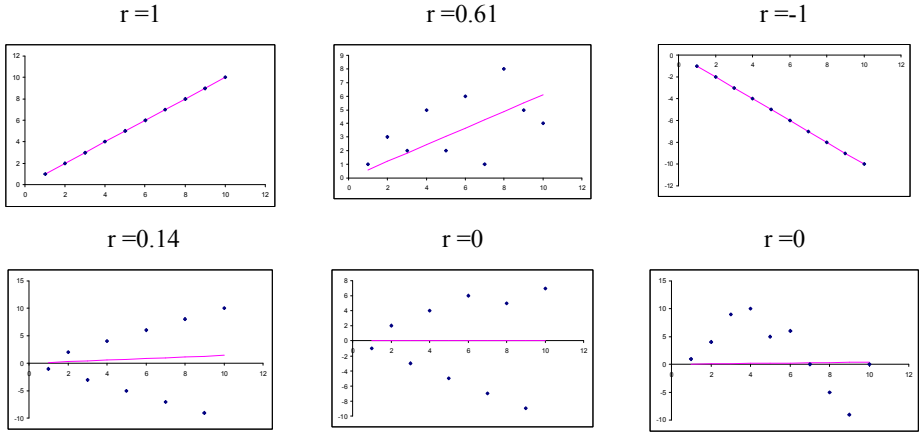
$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

r is the correlation coefficient from a sample and ***it measures how good the correlation is***. A perfect linear relation has r = +1 (positively correlated) or -1 (negatively correlated); no correlation is characterized by r = 0.
Such a test might be used for example in the following case:

Two sets of measurement were done for an analysis of a biological activity of a new compound. The growth of the plant was estimated by measuring its weight gain after a week of cultivation. One hundred plants were treated with the new compound prepared as a potential new herbicide and one hundred plants were grown in the absence of a herbicide. Both datasets have normal distribution so the parametric test can be used. First the average weights of these two sets have to be measured as well as their standard deviation. Then the correlation coefficient will be a measure of the effect of the herbicide. Note that in this case strong correlation will mean NO EFFECT.

The extreme cases of data are when correlation is 1 (strong correlation) and 0 (no correlation). Some other cases with correlation between 0<r<1 are graphically represented in figure below.

Figure 3. Graphical representation of some cases with selected correlations



In medicinal chemistry some researchers suggest that:

Correlation coefficient	The power of correlation
0.9-1.0	full
0.7-0.9	very strong
0.5-0.7	strong

0.3-0.5	moderate
0.1-0.3	weak
00-0.1	very weak
0.0	no correlation

Advantages of the correlation analysis are as follow;

1. Interpretation of an experimental data may be simplified
2. New aspects of information can come to light after the correlation analysis
3. prediction of additional information can be possible

4. Not normal distribution of dataset

If data are not represented by a linear relationship in order to measure the extent to which one variable increases, when the other variable tends to increase, the following rank correlation coefficients are used

-Spearman's rank correlation coefficient

-Kendall tau rank correlation coefficient

-Gamma test (statistics)

Such a test might be used for example in the following case:

Two physicians classified the same group patients in term of values between 1 and 10. The question is if their opinions are similar or not. Such a test is based on the order of ranks given by each physician to each patient.

correlation coeficient (relative quality of fit)

$$r^2 = 1 - \Sigma\Delta^2 / S_{yy}$$

standard deviation (absolute quality of fit)

$$s^2 = \Sigma\Delta^2 /(n - k - 1)$$

F-test (Fisher value, level of statistical significance)

$$F = r^2(n-k-1)/k(1-r^2))$$

Confidence interval for x

$$x = x_{avg} \pm sP(\alpha)$$

where $P(\alpha)$ stand for parmeter taken from statistical tables for normal distribution (if dataset has normal distribution) at significance level α. If dataset has t-Student distribution rhen instead of $P(\alpha)$ the parameter $t(\alpha, n)$, from t-Student statistical table has to be taken (n-means degree of freddom,)

5 Comparing many datasets

For the comparison of many datasets the most popular approach is an analysis of variance. The method is described in additional materials.

6. Regresion

Contrary to correlation the regression analysis includes techniques for modeling and analyzing the relation between one (simple linear regression) or many (multiple linear regression) independent variables x_i and dependent variable y in a form of a function. Specifically it helps to understand how the changes of independent variables influences the behaviour of dependent variable.

$$y=f(x_i)$$

simple linear regression (x and y are linearly depended)

$$y=ax+b$$

where a and b are parameters to be calculated

multiple linear regression (x and y are linearly depended)

$$y=a_nx_n+\ldots\ldots\ldots+a_6x_6+a_5x_5+a_4x_4+\ldots\ldots+a_1x_1+b$$

nonlinear regression (x and y are not linearly depended), for example

$$y=a \log(x) +x$$

7. Parameters

The Correlation analysis, regression analysis as well as other methods presented in this study guide the explorer through the dependence between the various parameters. This chapter is devoted to the description of some of the most important features by which the chemical compounds are described. Such parameters are used in many chemical analyses including rational drug design.

7.1 Electronic parameters, Hammet equation

Electronic properties were initially developed from a consideration of substituent effects in aromatic compounds. For example, the **dissociation constants of substituted benzoic acids** ($Kx10^5$ at 25^oC) were used by Hammett.

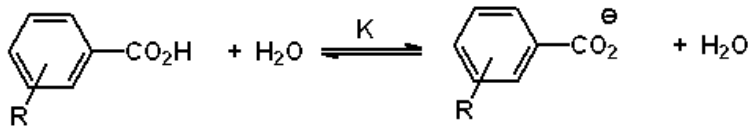


Table 2. Electronic parameters of some selected substituents in aromatic ring

R	H	CH_3	OCH_3	F	Cl	NO_2
ortho	6.27	12.3	8.06	54.1	11.4	671
meta	6.27	5.35	8.17	13.6	14.8	32.1
para	6.27	4.24	3.38	7.22	10.5	37.0

The question is: can we rationalize the effect of R based on the acidity of the acids?

From the table we can conclude that if R is the electron donating then the acid form is stabilized and equilibrium is shifted to the left in respect to the nonsubstituted derivative. Otherwise when R is the electron-withdrawing then the anion form is stabilized which means that equilibrium is shifted to the right.

Electronic properties can be then quantified by equilibrium constant K_a, as follows:

$$\log K_{a\,(\text{substituted acid})} - \log K_{a\,(\text{unsubstituted acid})} = \log K_{a\,(RX)} - \log K_{a\,(RH)}$$
$$= \log \frac{K_{a\,(RX)}}{K_{a\,(RH)}}$$
$$= \sigma$$

where **σ is the substituent constant** for a given group, R, and K_a are acid dissociation equilibrium constants. In more detailed considerations one can distinguish the effects for ortho, para and meta substituents.

The most typical values of the substituent constants, σ_m and σ_p are shown in the table.

Table 3. The most typical values of the substituent constants, σ_m and σ_p

Substituent	σ		Substituent	σ	
	Meta	*Para*		*Meta*	*Para*
O	-0.708	-1.00	F	+0.337	+0.062
OH	+0.121	-0.37	Cl	+0.373	+0.227
OCH_3	+0.115	-0.268	CO_2H	+0.355	+0.406
NH_2	-0.161	-0.660	$COCH_3$	+0.376	+0.502
CH_3	-0.069	-0.170	CF_3	+0.43	+0.54
$(CH_3)_3Si$	-0.121	-0.072	SO_2Ph	+0.61	+0.70
C_6H_5	+0.06	-0.01	NO_2	+0.710	+0.778
H	0.000	0.000	$^{+}N(CH_3)_3$	+0.88	+0.82
SH	+0.25	+0.15	N_2^{+}	+1.76	+1.91
SCH_3	+0.15	0.00	$^{+}S(CH_3)_2$	+1.00	+0.90

When σ is assumed to be transferable to many reactions involving benzene and other aromatic species, it leads to a generalized form of the equation known as the **Hammett equation**:

$$\rho\sigma = \log K_{a\,(R)} - \log K_{a\,(H)}$$

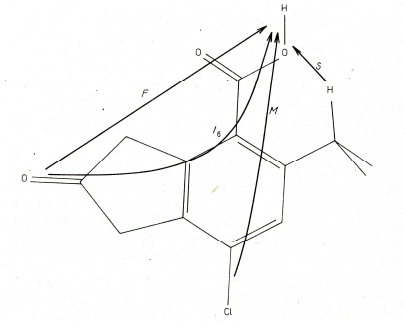
ρ is the reaction constant, σ is the substituent constant, and $\mathbf{K_a}$ is the equilibrium constant (or rate constant, $\mathbf{k_a}$) for the reaction of interest.

There are several other ways of quantifying electronic effects. For example, electronic effects can be represented as a linear combination of a field (**inductive**) effect, **F**, and a **resonance** effect, **R**:

$$\sigma = aF + bR$$

where a and b are coefficients determined from data fitting. The use of σ as well as other parameters described in this chapter has been extended to many types of effects in chemistry as well as biological activity studies.

Figure 4. The definition of the electronic and other effects which are quantified in Hammet analysis.



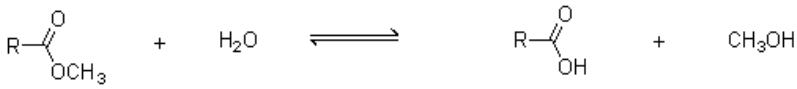
The list of Hammet parameters for most typical substituents is given below.

Steric substituent constants

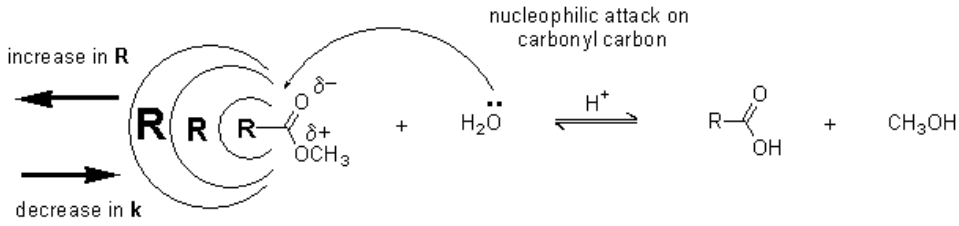
Others can be found in many books for example Otto Exner, Correlation analysis of chemical data, Plenum press, 1988, and www websites, http://www.wiredchemist.com/chemistry/data.)

7.2 Steric parameters

In similar manner like Hammet quantified the electronics effect, Taft quantified the **steric** (spatial) effects using the hydrolysis of esters:



Here, the size of R affects the rate of reaction by *blocking nucleophilic attack by water*.



In this case, the steric effects were quantified by the Taft parameter $\mathbf{E_s}$**:**

$$\mathbf{E_S} = \log \mathbf{k}_{RCO_2CH_3} - \log \mathbf{k}_{CH_3CO_2CH_3}$$
$$= \log \frac{\mathbf{k}_{RCO_2CH_3}}{\mathbf{k}_{CH_3CO_2CH_3}}$$

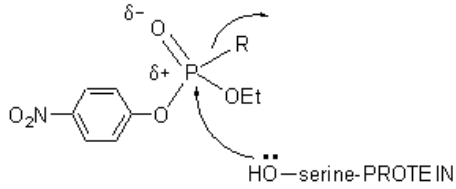
where k is the rate constant for the ester hydrolysis. This expression is analogous to the Hammett equation.

Table 4. E_s Values for Various Substituents

H	Me	Pr	*t*-Bu	F	Cl	Br	OH	SH	NO_2	C_6H_5	CN	NH_2
0.0	-1.24	-1.60	-2.78	-0.46	-0.97	-1.16	-0.55	-1.07	-2.52	-3.82	-0.51	-0.61

Note: H is usually used as the reference substituent (E_s^0), but sometimes methyl (Me) is used as the reference, the value of parameters differ then by 1.24.
As was the case for σ , E_s may be used in other chemical reactions and to explain biological activities,

The example is the hydrolysis of inhibitors of the acetylcholine esterase.



which must be hydrolysed in order to be active.

The observed biological activity in this case nicely correlates with Taft steric parameter $\mathbf{E_S}$ for the substituent R by the equation:

$$\log(1/C) = 2.58\,\mathbf{E_s} + 7.94$$

Below there are given the Taft steric parameters for the most known substituents.

Table 5. Selected Taft parameters

substituent	ν	Es
H	0	1.24
CH3	0.52	0
F	0.27	0.78
C3H7	0.68	-0.37
C6H5	0.57	-2.55

The Taft parameters describe the substituent by a single number. This approach is not adequate for the sterically irregular groups. There are other steric parameters, STERIMOL for example, which describe the size and shape.

STERIMOL size parameters (L, B_1, B_2, B_3 and B_4) were proposed by Verloop and are, defined as:

L = length along the axis of the bond joining R to the parent molecule

B_i = the four width parameters, at right angles to the axis, L, viewed in cross-section,

and $B_1 < B_2 < B_3 < B_4$. (see picture)

Figure 5. Sterimol parameters definition

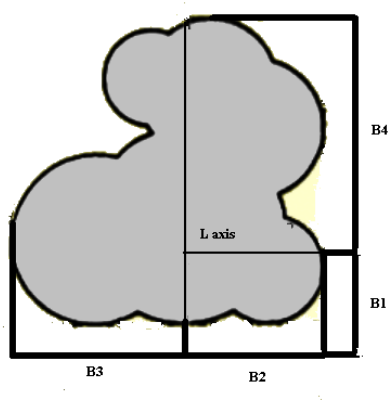


Table 6. Values of STERIMOL parameters for selected substituents

R	H	Me	nPr	tBu	F	Cl	Br	OH	SH	NO2	C6H5	CN	NH2
L	2.05	3.00	5.05	4.11	2.65	3.52	3.83	2.74	3.47	3.44	6.28	4.23	2.93
B1	1.00	1.52	1.52	2.59	1.35	1.80	1.95	1.35	1.70	1.70	1.70	1.60	1.50
B2	1.00	2.04	3.49	2.97	1.35	1.80	1.95	1.93	1.70	1.70	1.70	1.60	1.50
B3	1.00	1.90	1.90	2.86	1.35	1.80	1.93	1.35	2.44	2.44	3.11	1.60	1.50

7.3 Molar refractivity

Molar Refractivity, $\mathbf{MR = \frac{n^2 - 1}{n^2 + 1} \cdot \frac{MW}{d}}$ where *n* is the refractive index, MV= MW/d is the molar volume , MW = molecular weight and d = density. MR is a combination of volume (MV) and polarizability (a property of *n*) in a molecule, has been successfully used in many QSAR studies.

7.4 Topological descriptors

Topological indices are 2D descriptors based on the graph theory concepts (Kier and Hall 1976, 1986; Katritzky and Gordeeva 1993). These indices help to differentiate the molecules according to their size, degree of branching, flexibility and overall the shape.

Wiener index (W)

The Wiener index is the sum of the chemical bonds existing between all pairs of heavy atoms in the molecule.

$$W = \frac{1}{2}\sum_{i}\sum_{j} a_{ij}^{D}$$

Zagreb index (Zagreb)

The Zagreb index is defined as the sum of the squares of vertex valencies (Bonchev 1983):

$$Zagreb = \sum_{i} \delta_i^2$$

Randić index, known also as the **connectivity index**, is the sum of $1 / (d_i d_j)^{1/2}$ where d_i and d_j are the degrees of the vertices making the bond i~j.

Hosoya index (Z)

The Hosoya index, (Z index), is the total number of matchings in a graph. Matching or an independent edge set in a graph is a set of edges without common vertices.
Other descriptors which are given below are more complicated and are given without a detailed descrcription. They can be found in the provided reference literature.

Table 7. Selected topological parameters

RelNo	Symbol	Name	References
1	ZM1	first Zagreb index M1	Gutman, I., Ruscic, B., Trinajstic, N. SWilcox Jr, C.F. J. Chem. Phys., (1975), 62, 3399-3405.
2	ZM1V	first Zagreb index by valence vertex degrees	
3	ZM2	second Zagreb index M2	
4	ZM2V	second Zagreb index by valence vertex degrees	
5	Qindex	Quadratic index	Balaban, A.T. Theor. Chim. Acta, (1979), 53, 355-375.
6	SNar	Narumi simple topological index (log)	Narumi, H. MATCH (Comm. Math. Comp. Chem.), (1987), 22,195-207.
7	HNar	Narumi harmonic topological index	
8	GNar	Narumi geometric topological index	
9	Xt	Total structure connectivity index	Needham, D.E.Wei, I.C. & Seybold, P.O. J. Am. Chem. Soc., (1988), 110, 4186-4194.
10	Dz	Pogliani index	Pogliani, L. J. Phys. Chem., (1996), 100, 18065-18077.
11	Ram	ramification index	Araujo, O. & De La Pena, J.A. J. Chem. Inf. Comput. Sci., (1993), 33, 327-331.
12	Pol	polarity number	Platt.J.R. J. Chem. Phys., (1947), 15, 419-420.
13	LPRS	log of product of row sums (PRS)	Schultz, H.P, Schultz, E.B. & Schuttz, T.P. J. Chem. Inf. Comput. Sci., (1992), 32, 39-72.
14	VDA	average vertex distance degree	E.V.Kostantinova, J. Chem. Inf. Comp. Sci., (1997), 38, 54-57. Skorobogatov, V.A. and Dobrynin, A.A. MATCH (Comm. Math. Comp. Chem.), (1988), 23,105-151.
15	MSD	mean square distance index (Balaban)	Balaban,A.T. Pure & Appl. Chem., (1983), 55,199-203.
16	SMTI	Schultz Molecular Topological Index (MTI)	Schuttz,H.P. J. Chem. Inf. Comput. Sci., (1989), 29, 227-223.
17	SMTIV	Schultz MTI by valence vertex degrees	
18	GMTI	Gutman Molecular Topological Index	Gutman,I. J. Chem. Inf. Comput. Sci., (1994), 34, 1037-1039.
19	GMTIV	Gutman MTI by valence vertex degrees	
20	Xu	Xu index	Ren, B. J.Chem. Inf. Comput. Sci., (1999), 39,139-143.
21	SPI	superpendentic index	Gupta, S., Singh. M. & Madan, A.K. J. Chem. Inf. Comput. Sci., (1999), 39, 272-277.

22	W	Wiener W index	Wiener, H. J. Am. Chem. Soc., (1947), 69, 17-20.
23	WA	mean Wiener index	
24	Har	Harary H index	Ivanciuc, O., Balaban, T.-S. & Balaban, A.T. J. Math. Chem, (1993), 12, 309-318.
25	Har2	square reciprocal distance sum index	
26	QW	quasi-Wiener index (Kirchhoff number)	Mohar, B., Babic, D.& Trinajstic, N. J. Chem. Inf. Comput. Sci., (1993), 33, 153-154.
27	TI1	first Mohar index TI1	Mohar, B. MATH/CHEM/COMP 1988 (Graovac, A, ed.), Elsevier, Amsterdam (The Netherlands)
28	TI2	second Mohar index TI2	
29	STN	spanning tree number (log)	
30	HyDp	hyper-distance-path index	Diudea, M.V. J. Chem. Inf. Comput. Sci., (1996), 36, 535-540.
31	RHyDp	reciprocal hyper-distance-path index	Diudea, M.V. J. Chem. Inf. Comput. Sci., (1997), 37, 292-299.
32	w	detour index	Amic, D. & Trinajstic, N. Croat. Chem. Acta, (1995), 68, 53-62.
33	ww	hyper-detour index	Diudea, M.V. J. Chem. Inf. Comput. Sci., (1996), 36, 535-540.
34	Rww	reciprocal hyper-detour index	Diudea, M.V. J. Chem. Inf. Comput. Sci., (1997), 37, 292-299.
35	D/D	distance/detour index	Randic, M. J. Chem. Inf. Comput. Sci., (1997), 37, 1063-1071.
36	Wap	all-path Wiener index	Lukovits, I. J. Chem. Inf. Comput. Sci., (1998), 38, 125-129.
37	WhetZ	Wiener-type index from Z weighted distance matrix (Barysz matrix)	Barysz, M., Jashari, G., Lall, R.S., Srivastava, A.K. & Trinajstic, N. Chemical Applications of Topology and Graph Theory (King, R.B., ed.), Elsevier, Amsterdam (The Netherlands), (1983), pp. 222-230.
38	Whetm	Wiener-type index from mass weighted distance matrix	
39	Whetv	Wiener-type index from van der Waals weighted distance matrix	
40	Whete	Wiener-type index from electronegativity weighted distance matrix	
41	Whetp	Wiener-type index from polarizability weighted distance matrix	
42	J	Balaban distance connectivity index	Balaban A.T. Chem. Phys. Lett., (1982), 89, 399-404.
43	JhetZ	Balaban-type index from Z weighted distance matrix (Barysz matrix)	
44	Jhetm	Balaban-type index from mass weighted distance matrix	
45	Jhetv	Balaban-type index from van der Waals weighted distance matrix	
46	Jhete	Balaban-type index from electronegativity weighted distance matrix	
47	Jhetp	Balaban-type index from polarizability	

		weighted distance matrix	
48	MAXDN	maximal electrotopological negative variation	Gramatica, P., Corradi, M., Consonni, V. Chemosphere, (2000), 41,783-777.
49	MAXDP	maximal electrotopological positive variation	
50	DELS	molecular electrotopological variation	
51	TIE	E-state topological parameter	Voelkel, A. Computers Chem., (1994), 18, 1-4.
52	S0K	Kier symmetry index	Kier, L.B. Quant. Struct. -Act. Relat., (1987), 6, 8-12.
53	S1K	1-path Kier alpha-modified shape index	Kier, L.B. Quant. Struct. -Act. Relat., (1985), 4, 109-116.
54	S2K	2-path Kier alpha-modified shape index	
55	S3K	3-path Kier alpha-modified shape index	
56	PHI	Kier flexibility index	Kier, L.B. Quant. Struct. -Act. Relat., (1989), 8, 221-224.
57	BLI	Kier benzene-likeliness index	Kier, L.B . & Hall, L.H. Molecular Connectivity in Structure-Activity Analysis. RSP-Wlley, Chichester (UK), (1986).
58	PW2	path/walk 2 - Randic shape index	Randic,M. J. Chem. Inf. Comput. Sci., (2001), 41, 607-613.
59	PW3	path/walk 3 - Randic shape index	
60	PW4	path/walk 4 - Randic shape index	
61	PW5	path/walk 5 - Randic shape index	
62	PJI2	2D Petitjean shape index	Petitjean, M. J. Chem. Inf. Comput. Sci., (1992), 32, 331-337.
63	CSI	eccentric connectivity index	Sharma, V., Goswami, R. & Madan, A.K. J. Chem. Inf. Comput. Sci., (1997), 37,273-282.
64	ECC	eccentricity	E.V.Kostantinova, J. Chem. Inf. Comp. Sci., (1997), 38, 54-57. Skorobogatov, V.A. and Dobrynin, A.A. MATCH (Comm. Math. Comp. Chem.), (1988), 23,105-151.
65	AECC	average eccentricity	
66	DECC	eccentric	
67	MDDD	mean distance degree deviation	
68	UNIP	unipolarity	
69	CENT	centralization	
70	VAR	variation	Entiger, R.C., Jackson, D.E. and Snyder, D.A. Czech. Math. J., (1978), 26, 283-296.
71	BAC	Balaban centric index	Balaban, A.T. Theor. Chim. Acta, (1979), 53, 355-375.
72	Lop	Lopping centric index	
73	ICR	radial centric information index	Bonchev, D. & Rouvray, D.H. Eds. Chemical Graph Theory. Gordon & Breach,
74	D/Dr03	distance/detour ring index of order 3	
75	D/Dr04	distance/detour ring index of order 4	

76	D/Dr05	distance/detour ring index of order 5	New York (NY), (1991). Trinajstic,N., Chemical Graph Theory. CRC Press, Boca Raton (FL), (1992). Devillers, J. & Balaban, A.T. Eds. Topological Indices and Related Descriptors in QSAR and Drug Design. Gordon&Breach, Amsterdam (The Netherlands), (2000).
77	D/Dr06	distance/detour ring index of order 6	
78	D/Dr07	distance/detour ring index of order 7	
79	D/Dr08	distance/detour ring index of order 8	
80	D/Dr09	distance/detour ring index of order 9	
81	D/Dr10	distance/detour ring index of order 10	
82	D/Dr11	distance/detour ring index of order 11	
83	D/Dr12	distance/detour ring index of order 12	
84	T(N..N)	sum of topological distances between N..N	
85	T(N..O)	sum of topological distances between N..O	
86	T(N..S)	sum of topological distances between N..S	
87	T(N..P)	sum of topological distances between N..P	
88	T(N..F)	sum of topological distances between N..F	
89	T(N..Cl)	sum of topological distances between N..Cl	
90	T(N..Br)	sum of topological distances between N..Br	
91	T(N..I)	sum of topological distances between N..I	
92	T(O..O)	sum of topological distances between O..O	
93	T(O..S)	sum of topological distances between O..S	
94	T(O..P)	sum of topological distances between O..P	
95	T(O..F)	sum of topological distances between O..F	
96	T(O..Cl)	sum of topological distances between O..Cl	
97	T(O..Br)	sum of topological distances between O..Br	
98	T(O..I)	sum of topological distances between O..I	
99	T(S..S)	sum of topological distances between S..S	
100	T(S..P)	sum of topological distances between S..P	
101	T(S..F)	sum of topological distances between S..F	
102	T(S..Cl)	sum of topological distances between S..Cl	

103	T(S..Br)	sum of topological distances between S..Br	
104	T(S..I)	sum of topological distances between S..I	
105	T(P..P)	sum of topological distances between P..P	
106	T(P..F)	sum of topological distances between P..F	
107	T(P..Cl)	sum of topological distances between P..Cl	
108	T(P..Br)	sum of topological distances between P..Br	
109	T(P..I)	sum of topological distances between P..I	
110	T(F..F)	sum of topological distances between F..F	
111	T(F..Cl)	sum of topological distances between F..Cl	
112	T(F..Br)	sum of topological distances between F..Br	
113	T(F..I)	sum of topological distances between F..I	
114	T(Cl..Cl)	sum of topological distances between Cl..Cl	
115	T(Cl..Br)	sum of topological distances between Cl..Br	
116	T(Cl..I)	sum of topological distances between Cl..I	
117	T(Br..Br)	sum of topological distances between Br..Br	
118	T(Br..I)	sum of topological distances between Br..I	
119	T(I..I)	sum of topological distances between I..I	

Table 8. Most frequently used other parameters

Parametr	***Symbol***
Hydrophobic parameters	
Partition coefficient	log *P*
Substituent constant	π

Hydrophobic fragmental constant	f, f'
Distribution coefficient	log D
Apparent partition coefficient (fixed pH)	log P', log P_{app}
Capacity factor in HPLC	log k, log k_w
Solubility parametr	log S
Electronic descriptors	
Hammett constants	σ, σ^-, σ^+
Taft's inductive (polar) constants	σ^*, σ_I
Swain and Lupton field parameter	F
Swain and Lupton resonance parameter	R
Ionization constant	pK_a, ΔpK_a
Chemical shifts (^{13}C and ^{1}H)	δ
Theoretical parameters	
Atomic net charge	q^σ, q^π
Superdelocalizability	S^N, S^E, S^R
Energy of highest occupied molecular orbital	E_{HOMO}
Energy of lowest unoccupied molecular orbital	E_{LUMO}
Electrostatic potential	$V(r)$
Steric descriptors	
Taft's steric parameter	E_s
Molar volume	MV
Molecular weight	MW
Van der Waals radius	r
Van der Waals volume	V_w
Molar refractivity	MR
Parachor	P_r
STERMOL parameters	L, B_i

8. Hansch analysis

Hammet equation has been defined above in the section describing the electronic parameters. The modification of this approach in the explanation of the relation between the biological activity and structural parameters has been introduced by Hansch and is known as Hansch equation or QSAR (quantitative structure activity relationship).

A QSAR generally takes the form of a linear equation

$$\log (1/C) = k_1(\log P)^2 + k_2 \log P + k_3 s + k_4 \rho + k_5$$

for: C = minimum effective dose

P = octanol - water partition coefficient

s = Hammett substituent constant

k_i= constants derived from regression analysis

Lipophilicity parameters

Corvin Hansch in the 1960s pointed out that in the analysis of a drug action it is necessary to consider the additional parameter lipophilicity.

It was a very important step in medicinal chemistry as pointed out by S. L. Carney (DDT 9, 158-160 (2004)): "Has there been a single development that, in your opinion, has moved the field of medicinal chemistry ahead more than any other?" and Robert Ganellin: "I would go back to the 1960s to the work of Corwin Hansch on the importance of lipophilicity. ... I think that changed the way of thinking in medicinal chemistry. .... I think that the application of physical organic chemical approaches to structure–activity analysis have been very important."

In chemistry and the pharmaceutical sciences, a **partition-** (P) is the ratio of concentrations of a compound in the two phases water and octanol. In a form of a equation it is written:

$$log\ P_{oct/wat} = log\left(\frac{[solute]_{octanol}}{[solute]_{water}^{un-ionized}}\right)$$

Fragmental substituent constant

We can also define the fragmental lipophilic parameter π as

$$\pi(x)=\log P_X-\log P_H$$

and assume that it is additive.

Table 9. Fragmental constants for a chosen functional groups.

substituent	π
H	0
CH3	0.56
CN	-0.57
NO2	-0.28

Hansch equation

Typical Hansch equation with other important parameters describing the relation of a biological activity of a group of compounds in a form of the linear regression model is given below.

$$\log (1/C) = 1.20(\pm0.2)\ \pi + 1.46\ (\pm0.1)\ \sigma + 0.6\ (\pm0.02)$$

9. Free Wilson metod

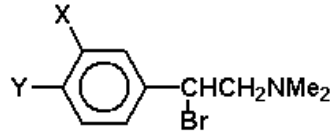
Free and Wilson assume that the biological activity for a set of analogues could be described by the contributions of that substituents or structural elements.
Instead of using π, σ_m, σ_p, F, R, E_S, and other parameters, Free Wilson equation (below) describes the biological activity in a form:

$$\log(1/C) = \sum a_j X_{ij} + \mu$$

where a_j is the group X contribution (takes value 1 if the substituent is present in position j in molecule i and 0 in the absence of it) and μ is the reference value for a parent compound.
It is possible to add mixed variables as a combination of Free-Wilson analysis and Hansch analysis.
In some works other indicators like for example substructures, chiral centers and special substituents might be used.

A typical tabulation for a set of compounds in Free Wilson approach is shown in the following table:

Table 10. Free Wilson analysis example



meta	*para*	*meta-*					*para-*					log 1/C	log 1/C
(X)	(Y)	F	Cl	Br	I	Me	F	Cl	Br	I	Me	obsd.	calc.a)
H	H											7.46	7.82
H	F						1					8.16	8.16
H	Cl							1				8.68	8.59
H	Br								1			8.89	8.84
H	I									1		9.25	9.25
H	Me										1	9.30	9.08
F	H	1										7.52	7.52
Cl	H		1									8.16	8.03
Br	H			1								8.30	8.26
I	H				1							8.40	8.40
Me	H					1						8.46	8.28
Cl	F		1				1					8.19	8.37
Br	F			1			1					8.57	8.60
Me	F					1	1					8.82	8.62
Cl	Cl		1					1				8.89	8.80
Br	Cl			1				1				8.92	9.02
Me	Cl					1		1				8.96	9.04

Cl	Br		1						1			9.00	9.05
Br	Br			1					1			9.35	9.28
Me	Br					1			1			9.22	9.30
Me	Me					1					1	9.30	9.53
Br	Me			1							1	9.52	9.51

After performing the regression analysis the equation is

$$\log(1/ED_{50}) = -0.301[m\text{–F}] + 0.27[m\text{–Cl}] + 0.434[m\text{–Br}] + 0.579[m\text{–I}] + 0.454[m\text{–Me}] + 0.340[p\text{–F}] + 0.768[p\text{–Cl}] + 1.020[p\text{–Br}] + 1.429[p\text{–I}] + 1.256[p\text{–Me}] + 7.821$$

$$n = 22,\ r^2 = 0.94,\ s = 0.194,\ F = 17.0$$

The independent variables *indicate the status* of these groups. A negative coefficient indicates that the presence of that group is *unfavourable* to the activity; a positive coefficient indicates that the presence of that group is favourable to the activity. The indexes in brackets correspond to a symbol; of a group.

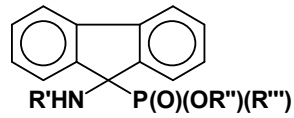
10. Misleadings in regression analysis.

In general in a typical regression model the data used are correlated, which means that the same information is introduced into the model many times. Introduction of correlated data brings no new information to the model but increases the number of independent variables and as a consequence the increase of the degrees of freedom for the errors. The goodness of a fit performed by the R-squared will raise with addition of a newly correlated variable. It results in the increase of R-squared but not in the predictive power of the regression model. To avoid such a situation three methods most frequently applied in regression model analysis will be presented : the stepwise regression analysis, the leave one out and the principal component analysis.

Methods will be presented on a dataset of herbicidal aminophosphonates. The dataset was elaborated by Gancarz and Kosior.

Example.

The 50 compounds of a structure presented below were made.



Ten plants were chosen for the biological herbicidal activity,

1 - ryegrass (*Arrhenatherum elatius*)

2 - oats (*Avena sativa*)

3 - maize (*Zea mays*)

4 - mustard (*Sinapis arvensis*)

5 - peas (*Pisum sativum*)

6 - bean (*Phaseolus vulgaris*)

7 - cucumber (*Cucumis sativus*)

8 - flox (*Linum usitatissimum*)

9 - red beet (*Beta esculenta*)

10 - buckwheat (*Fagopyrum sagittatum*)

Then some physicochemical parameters were calculated. A part of the dataset is presented below.

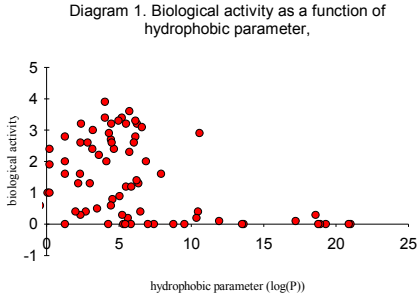
Table 11. The selected data for leave one out analysis

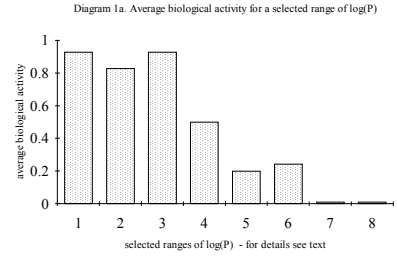
No.	R'	R", R'"	Activity for particular plant in 1-4 scale	BA (average)	v	S	log(P)
17	nBu	nPr	3 4 3 4 3 3 4 3 3 4	3.4	1.8	-3.573	5.194
18	nBu	iPr	3334334343	3.3	2.18	-3.819	4.962
19	nBu	nBu	2334333344	3.2	1.84	-3.657	6.248
20	nBu	CH	0001011100	0.4	1.84	-3.599	10.464
21	nBu	CH	0000000000	0	1.84	-3.805	13.626
22	nBu	CH	0000000000	0	1.84	-3.957	18.896
23	nBu	CH	0000000000	0	1.84	-3.957	21.004
24	nBu	Ph	0001222203	1.2	2.08	-2.181	5.474
25	iBu	Et	4434444444	3.9	1.94	-3.499	4.024
26	iBu	nBu	2104334344	2.8	2.14	-3.735	6.132
27	iBu	CH	0001000100	0.2	2.14	-3.677	10.348

28	iBu	CH	0000000000	0	2.14	-3.883	13.51
29	iBu	CH	0000000000	0	2.14	-4.035	18.78
30	secBu	Et	3334344343	3.4	1.98	-3.571	4.024
31	secBu	nBu	3324343344	3.3	2.18	-3.807	6.132
32	secBu	CH	0000000000	0	2.24	-4.107	20.888
33	nCH	iPr	2324244443	3.2	2.18	-3.895	5.489
34	nCH	Et	0004333343	2.3	1.69	-3.312	5.721
35	Ph	Me	0000020002	0.4	2.38	-2.411	2.699
36	Ph	nBu	0002032203	1.2	2.82	-2.919	5.861

The correlation of the biological activity with every physicochemical parameter is given in figures below.

Fig. Biological activity as a function of hydrophobic parameters



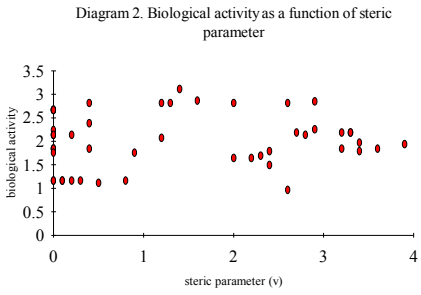


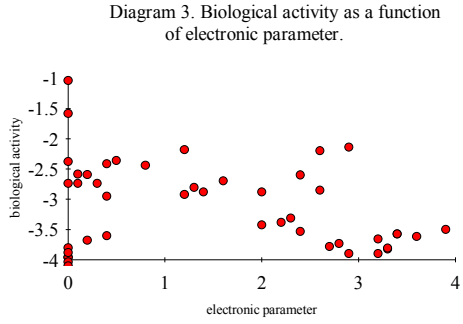
Biological activity vs. LogP

a) all compounds presented individually

b) all compounds classified into 8 groups (the height of the bar means an average activity of a class).

Figure 6. Biological activity as a function of steric and electronic parameters





The Analysis of the attached diagrams leads to the conclusion that the biological activity of the analyzed group of compounds is only a function of log(P) and is independent from neither the electronic either steric parameters. So we present it I in a form

$$BA=a_1\log(P)+const$$

With such a knowledge we can analyze how adequate are the regression models.

11. Multiple regression

Performing the standard procedure for the calculation of the regression coefficients for the problem defined above for the herbicidal aminophosphonates, the following equation can be obtained.

$$\mathbf{Ba = 0.134\ logP\ -68.1\ log^2P\ +2.461\ log^3P - 3.203\ \sigma + 0.292\ s\ -2.271}$$

It is evident that this is not the relation deduced according to the correlation analysis of dependent variables and independent one.

Stepwise regression

The stepwise regression as the name suggests is the process of developing the model in several steps. It starts from the simple linear regression (only one independent variable with the highest predictive power) and every new independent variable is introduced after the calculation only if it covers new information in explaining the behaviour of dependent variable at a significant level. The significance of the variables in most cases is done after performing the F-test, but also t-test and others are applied.

The main approaches are:

Forward selection,

The procedure starts with the regression model with no variables,

$Y=const$

Then the calculation of the significance of all independent variables is performed. The Next step includes the next variable, statisticaly the most significant for the regression model. Let it be variable x_4. Then the model is.

$Y=a_4x_4+const$

The next step is the calculation of the statistical significance of all independent variables, not yet included into the model, then selection of the most significant variable and introducing it into the model only if its significance is bigger than the defined by the user (in most cases the value is set to be greater than 4 at F-test). Let it be x_2, then the model is

$Y= a_2x_2+a_4x_4+const$ if F-test for variable x_2 is > assumed by a user, usually 4

The procedure is stopped at the moment when none of the independent variables not included into the model yet exceeds the minimal significance level.

It is important to state that significance means that it correlates strongly with the dependent variable (y) and it brings new, not yet brought to the model, information in explaining y. It means that if there are two strongly intercorrelated variables and both are also strongly correlated with the y value, then only one will be present in a model. When the first one will

be introduced into the model the F-test will manifest that the other is no longer important and will be rejected in the process of the model development.

Backward elimination,

The Procedure starts with the regression model with all variables, testing them one by one for the statistical significance, and deleting the most not significant at every step until none has less statistical significance than assumed.

Combination methods are also applied when at each stage variables are included or excluded.

Study case

For the example described above where the biological activity was modeled by a physicochemical parameters of a compound the several regression models were calculated. All of them are presented in the table below.
Analyzing the data of multiple regression models and in terms of multiple regression correlation coefficient (column R) one should conclude that the best model is the one with all the physicochemical parameters included in the model (R=0.943). The Analysis done in the previous chapter indicates that this is not true. Moreover the addition of a random variable RND makes the model even "better" (R=0.944).
According to such an analysis the regression model should be in the form:

$$\mathbf{Ba = a1\ logP + a2\ log^2P + a3\ log^3P + a4\ s + a6\ RND\ const}$$

Table 12. Results of multiple regression analysis

const	Log^3P*10^3	Log^2P*10^3	$LogP*10^3$	σ	v	RND	R	s	F	D
3.4	-	-	-0.187	-	-	-	0.748	0.780	$F_{1,24}$	0.837
4.1	-	7318	-0.360	-	-	-	0.757	0.725	$F_{2,23}$	0.815
-0.182	4.350	-147	1.191	-	-	-	0.826	0.665	$F_{3,22}$	0.808
-1.626	2.587	-73.11	0.198	-2.867	-	-	0.942	0.399	$F_{4,21}$	0.488
-2.271	2.461	-68.10	0.134	-3.023	0.292	-	0.943	0.399	$F_{5,20}$	0.512
-1.495	2.423	-67.20	-0.134	2.780	-	0.366	0.944	0.402		0.517

Leave one out

When performing the regression analysis the whole data set is used to built the regression model and then the same model is used for the analysis. The goodness of the fit in general is performed by the R-squared, analyses of of residuals, hypothesis testing, and statistical significance by an F-test of the overall fit, and by t-tests of individual parameters. The Regression model can lead however to big mistakes. The Regression coefficient which is assumed to measure the quality of the fit has the tendency to increase (never decrease) with the increase of independent variables. It means that adding some new data even completely not correlated with the depended variable (biological activity vs random variable for example) will result in the increase of the correlation coefficient. This is because every new independent variable means addition of the additional degree of freedom to the errors. To avoid such a misinterpretation **Cross-validation**, sometimes called **rotation estimation**, technique is used. It involves partitioning data into two complementary subsets, one called the *training set* for performing the analysis and the second called the *validation set* or *testing set* for validation of the predictive power of a model.

Below one of such cross validation is presented

Leave-one-out cross-validation

The method is multistep. As the name suggests, leave-one-out involves omitting in every one step one observation from the whole original dataset used, which is then the validation data, and the rest as the training data. Such a developed regression model is then used for the calculation of the omitted observation. The difference (d_i) between the calculated and real values of the dependent variable is a measure of the quality of the fit. The procedure is repeated many times and each observation in the dataset is used once as the validation data set. Then the squared sum of differences ($D=\sum d_i$) is a measure of the prediction power of the model.

Using other validation, root mean squared error or median absolute deviation can also be used. The best model is the one for which the D value is the smallest. According to the data in the table above the model suggests the following equation

$BA=a_1 log(P)+const$

It is exactly what was concluded at the beginning by analysing the influence of particular features on the biological effect.

12. Principal Component analysis

Another approach is to convert a set of correlated features, a1……an, into a set of new not correlated "features", let's say b1….bn, which are a linear combination of the original one. The Procedure involves calculation of eigenvectors (vi) and eigenvalues (wi) . The first one defines how to transform the original dataset to a new coordinate system, whereas the second yields the information about the significance of each of a new coordinate axis in description and the explanation of the depended variable.

$$bi = \sum (a1 * v1).$$

So the old model was

$$Y= f(ai)$$

Now the model is

$$Y=f(bi)$$

Where

$$bi=f(ai)$$

Such conversion of a coordinate system can be done by PCA method. More about the method is given in chapter dealing with PCA

The new b_i features, which are now not correlated are used in the regression analysis.

13. Pattern recognition methods

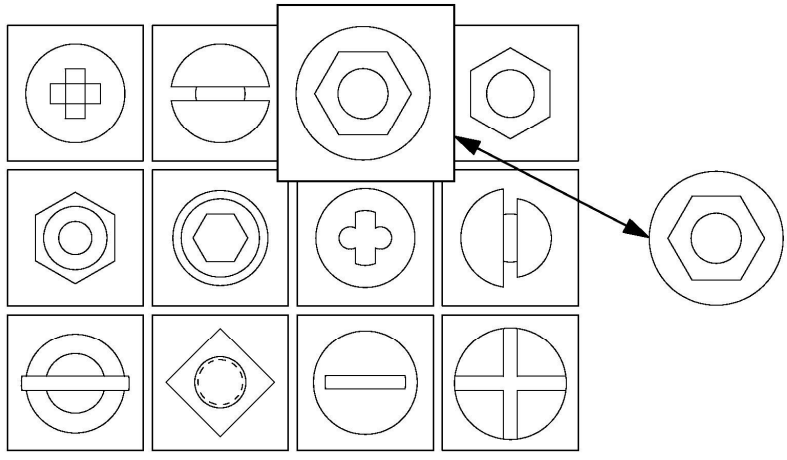
Introduction

Pattern recognition methods are very useful for the classification of objects (for example faces, chemicals, spectra and so on). They also try to find the relationship: a pattern vs. physical, biological or any other properties. It is important that such an analysis is objective as it is done without using the chemical knowledge or some prejudices. It tries to discover the internal similarities between the objects in the data set.

The examples of questions which can be solved are:

1. is he or it in the database (for example police searching in the database of faces, fingerprints or in the database of spectra)
2. is this his voice (in the identification of a person)
3. is it the letter a or b (in programs recognizing the text)
4. many other similar problems

Figure 7. Example of object search



As stated before the limitation of the space in this study guide does not allow to describe all pattern recognition methods. Only the most important will be presented in order to get an idea what the pattern recognition is and what kind of information it can provide especially for the chemist. Some examples are not strictly chemical or medicinal. They are presented as illustration of variation of applications of the pattern recognition methods and additionally some of them better illustrate the methods.

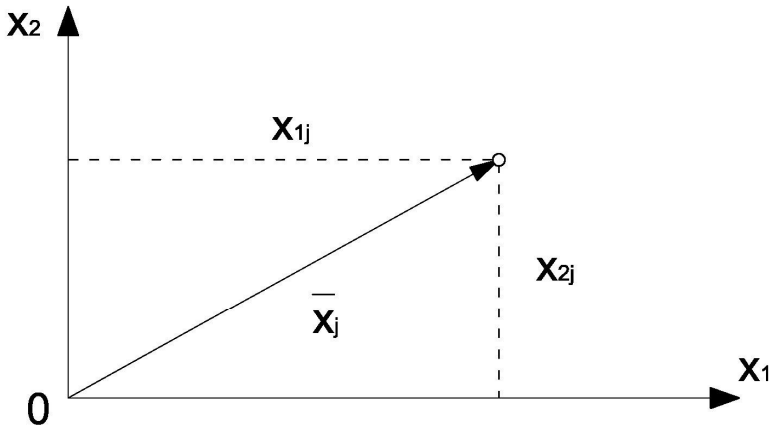
In some pattern recognition methods advanced mathematics is used however in this presentation we restricted the explanation of the mathematical background to the necessary minimum even if the method itself is mathematically complex. Only such mathematical formulas are presented which are necessary to understand the idea of the methods without the previous knowledge of the pattern recognition.

Pattern space

Definition

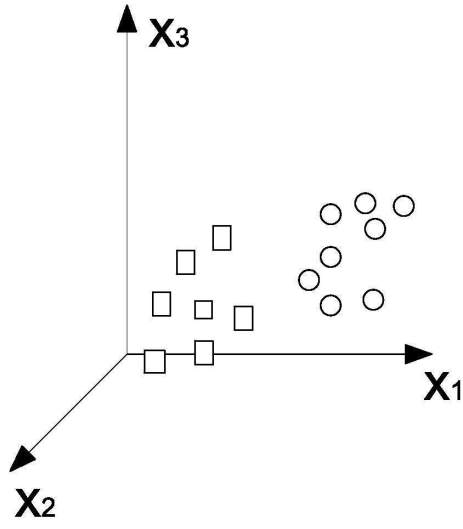
The fundamental concept in the pattern recognition methods is pattern space. Let's imagine we have two dimensional objects, named here patterns, with two descriptors x1 and x2 (for example a drug with the known steric parameter-x1 and known lipohilicity-x2). We can represent such an object in the two dimensional space, called here the pattern space, like it is shown on figure below.

Figure 8. Pattern space definition



In a similar manner we can represent a set of three dimensional patterns (drugs) which are defined by three parameters, x1, x2, x3. They can be represented in the three dimensional space by the set of points (patterns), see below. We can measure the biological activity of these compounds and label them: active – triangles, not active-squares.

Figure 9. Two kind of objects in pattern space



The goal in the pattern recognition methods is to identify similar objects. The figure above shows that squared boxes are similar to each other and distinct from the circles. If a new object appears in the area of circles it will be classified as a circle (active). This is the basic idea of judgment in the pattern recognition method – the similarity criterion.

Let's define a more complicated object, for example MS spectra. In the simplest way every spectrum can be described by the set of two parameter points, mass and corresponding intensity. For clarity of the further discussion let's measure the mass spectrum at mass ranges

1-200 and with the precision to one unit mass. Then every compound is described by a list of 200 points, whereas each point is characterized by the corresponding intensity.

We can represent every compound in 201 data space (200 coordinated for every mass point plus one coordinated for intensity) . In such space every mass spectrum of a specific compound is represented by a point. We can measure many MS spectra for the set of the known compound and place them in such space. Each compound is then represented in this space by a single point.

Identification of a new compound is based on measuring its MS spectrum and placing it as a point in the defined pattern space and identifying the closest neighbours.

In the summary we can say the following. An object is defined by a set of n features (x1 to xn). All features of the particular object define a pattern. Then the object can be represented as a point in n dimensional coordinate system called the pattern space. Identification means the analysis of neighbours. In all pattern recognition methods the assumption is taken that points similar in their properties are close together, in the sense of distance in the pattern space.

Some typical applications of the pattern recognition methods are:

- recognition of printed or handwritten characters
- analysis of spectra
- speech recognition
- fingerprint identification
- interpretation of clinical data
- medical diagnosis
- drug design
- interpretation of chemical data
- recognition of faces
- quality control
- recognition of shapes
- analysis of photographs

Classification in pattern space

The objects are placed in pattern space and they are classified into several groups (clusters, classes). Once such a structure (clusters, classes) is formed then the prediction of properties is possible. If object belongs to a certain class it may be assumed that its properties are similar to the properties of other members of a such class.
Most of the pattern recognition methods are nonparametric which means that the underlying statistics is not known.

Binary classification

Binary classification is the the classification process of the members of a given set of objects into two groups on the basis of whether they have some property or not.

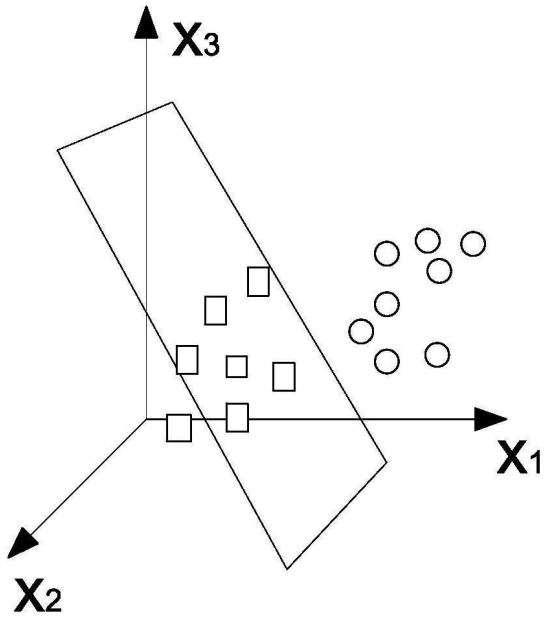
Some typical binary classification problems are:

- to determine if a patient has a certain disease or not
- to determine if the proposed for the synthesis compound will have the expected properties or not

The points in the picture below show two mutually exclusive classes (class 1 circles, class 2 squares). In such a situation it is possible to draw a plane (line in two dimensions or

a hyperplane in pattern space with more than three dimensions) that separates them completely like it is shown in the figure below (two classes form well separated groups which we can call clusters). Such a plane will be termed the decision plane.

Figure 10. Decision plane in pattern space



The Decision plane is calculated in the optimization process called training. This process is performed for the set of points with a known score (for example active or non active). Once the decision plane is formed then the plane is used for classification of the unknown point. Does it lay on the site of the plane where all the active compounds were clustered or on the other site.

The decision plane might have the finite thickness. The optimum of the thickness can be achieved by the training algorithm. Without going into the details in the two pictures below two examples are given with the plane having positive and negative thickness.

Figure 11.Plane with positive thickness

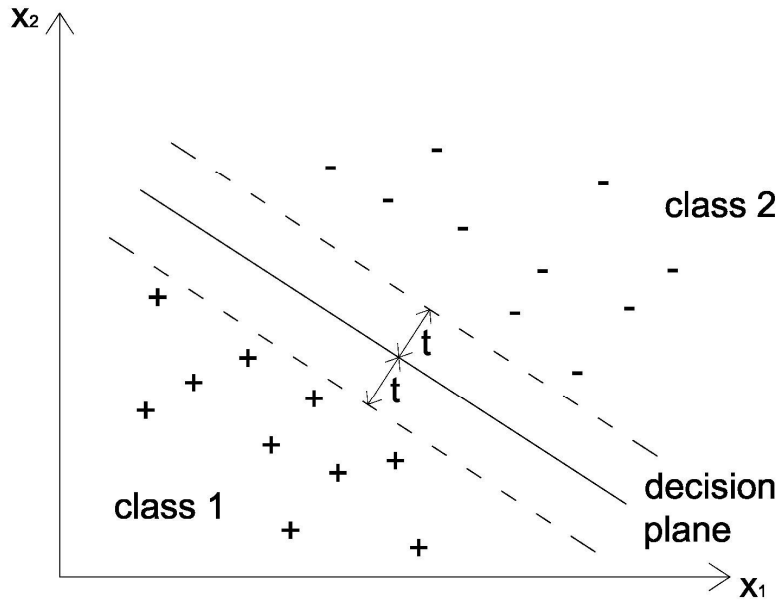
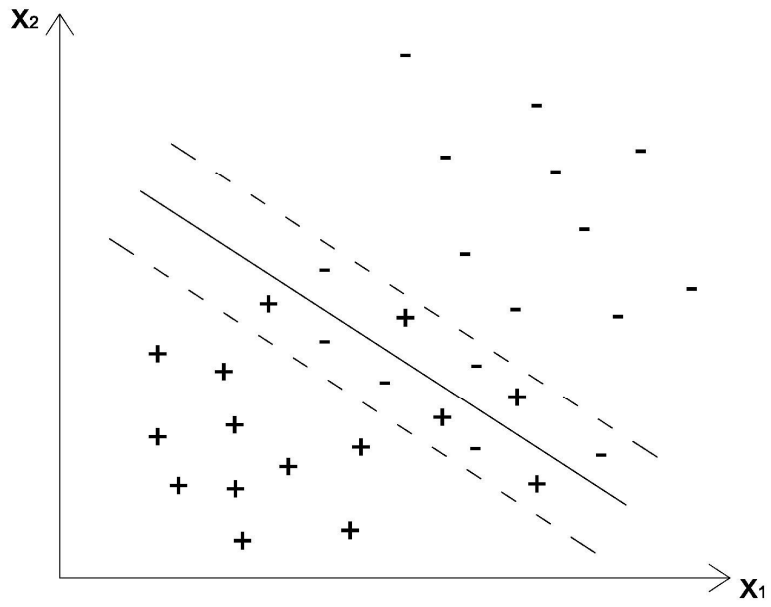


Figure 12. Plane with negative thickness



a plane with negative thickness

Selection of the parameters

Nobody knows a priori which set of parameters will give the chance of clusters separation. The selection of them could be than by the trial and error method. In the training step after a set of calculations using the appropriate algorithm the process will result in the output with the information which set of parameters separates the best the points with the known scores and will also indicate how many (percentage probability) of points were classified correctly. After that we can evaluate the position of a new point and get the information to which class it belongs (with the same probability).

14. Projection

Projections are the display methods which aim is to visualize the structure of multidimensional pattern space by the two or three dimensional representation. The Human eye is the best pattern recognizer but only in the two and three dimensional space. The Display methods try to reproduce the distances in the original multidimensional pattern space as far as it is possible into two or three dimensions. Of course some deformation is necessary because the exact reproduction can not be done when there is reduction of dimensionality. There are two basic approaches: linear (called projection) and nonlinear (called rather mapping).

Linear methods

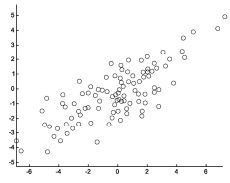
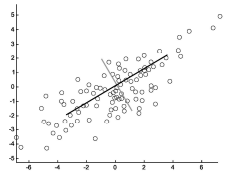
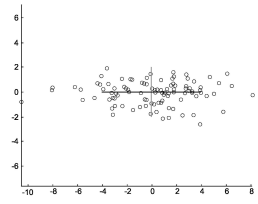
The simplest method of the linear projection is the variable by variable plot. This transformation is rather not fruitful when the starting dimensionality is large. However in many cases very valuable information is obtained when the projection is done on the two most important features. The case from chapter on QSAR provides a good example of the linear projection. Three different two dimensional projections log(P) vs BA, electronic parameters vs. BA and steric parameters vs. BA were presented above.
Only the first one gives valuable information that there is optimal log(P) and only the compounds for which log(P) is within the range 3.5-4.5 have the potential herbicidal activity. The other projections are useless.

Principal component projection

A very useful and the most optimum projection is Karhunen Loeve transformation. In this method new variables are formed which are linear combinations of the original one and in addition are orthogonal to each other. The calculation of new principal components were described in the previous chapter where the new variables were used in the regression development. The new coordinate system calculated in such a way can serve as a new coordinate system. It results in rotation of the coordinate system in such a way that the first coordinate system contains the most variance (information) about the dataset and the rest are ordered according to the decreasing amount of variance (information).

The mathematical procedure is not described here but the geometrical effect is presented below.

Figure 13. Example of principal coordinate system definition

Patterns in original coordinate system	*Calculation of principal components*	*Patterns in new coordinate system*
		

Data presented in the original coordinate system and in the two dimensional coordinate system are defined by the two the most important principal axis PC1 and PC2 (containing the most of the variance-information).

Figure 14. Example of principal coordinate system definition, coordinate system reduction.

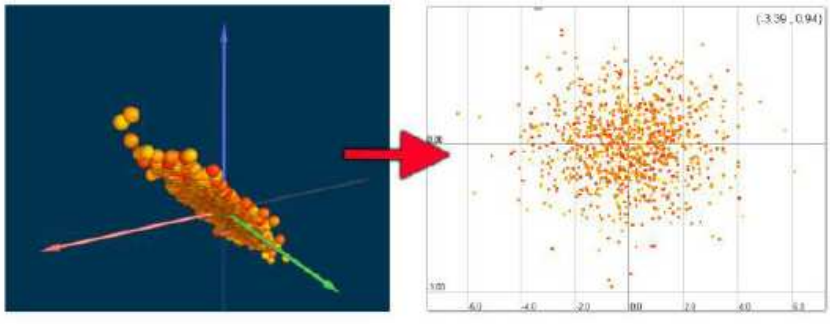
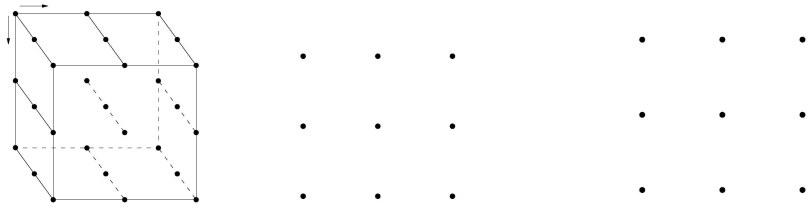


Figure 15.Variable by variable projection of a cube .



cube X-Z and X-Y projection

Fig Eigenvector Projection of a cube

cube PC1- PC2 projection

In general the procedure is done in such a way that the new coordinate system is orthogonal which results that the variables being not correlated. The new axis is called Principal components axis.

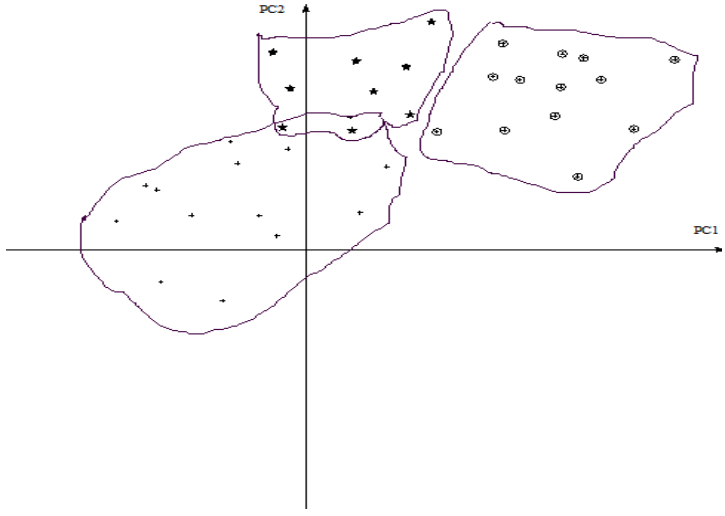
In the Paper “Fucosylation of serum glycoproteins in lung cancer patients” by Barbara Kossowska, Mirosława Ferens-Sieczkowska, Roman Gancarz, Ewa Passowicz-Muszyńska3 and Renata Jankowska the serum of patients with lung cancer were diagnosed and compared

with the serum of healthy people. The level of four proteins of the acute phase and four degrees of their fucosylation were measured and used for the diagnosis. Thus every patient formed a pattern in the eight dimensional pattern space.

Clinical subjects

There were 18 patients (Nos. 1–18) diagnosed with non-small cell lung cancer (NSCLC) at different stages of cancer development, 11 patients (Nos. 19–29) diagnosed with small-cell lung cancer (SCLC) and for the comparison of a serum of the control group of 10 healthy blood donors. (Nos.30–39). The following diagram presents the data in two dimmensional pattern space PC1, PC2. Three slightly overlapping areas correspond to the patterns of NSCLC, SCLC patients, and HEALTHY blood donors.

Figure 16. The result of lung cancer patient projection patient on two first PC components



The diagram above can serve as a tool for the diagnosis of a new patient. If the level of the four above defined proteins and the degree of their fucosylation is known then the patient pattern (the level of four proteins and their fucosylation degree) can be projected onto this above presented pattern space. Then depending on which area the patient pattern falls the appropriate diagnosis can be formulated. The judgment can be done in the sense to which class it belongs as well as from which "cancerogenous" patterns it derives.

Craigs and Toplis approach

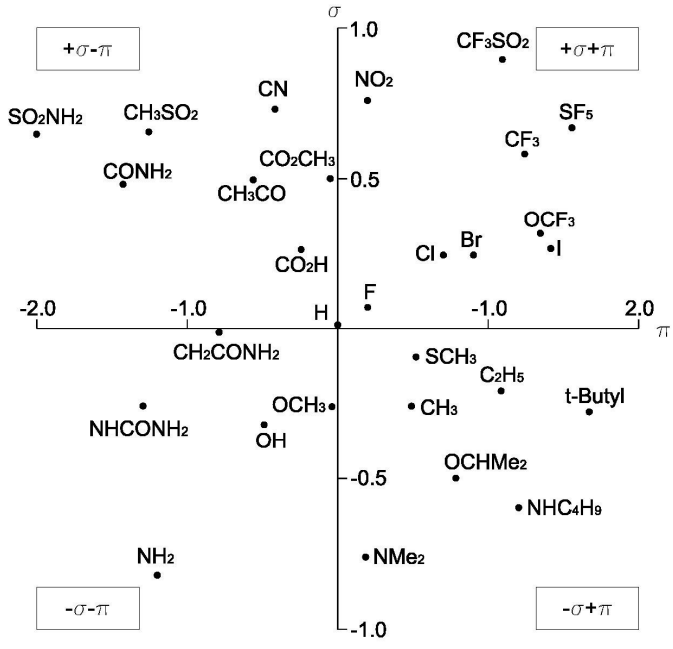
The below described Craigs presentation and the following Topliss procedure could be considered also as analysis of patterns projected on the special coordinate system.

Craigs

In the diagram below the axis of a coordinate system of pattern space is defined by physicochemical properties of chosen functional groups present in the analyzed group of the compounds. For example, if a compound contains nitro group it is placed in the upper right corner of the diagram. If all compounds of the known biological activity are analyzed and the position of the most active synthesized compounds is located then the analysis relies on the identification in which part of the diagram the new potential compound (a drug for example) with certain functional groups will be. Such an analysis is helpful in decision taking, synthesizing or not the drug candidate.

For example, if most of the active compounds were found in the part $+\sigma$, $+\pi$. Then the conclusion should be drawn that the next candidates for the synthesis should have $+\sigma$, $+\pi$ characteristics.

Figure 17. Craigs diagram



Topliss scheme

This is an algorithm which also suggests the next step in the synthesis in order to improve the biological activity. The procedure for modification of the parent compound by the aromatic substitution is shown in the Fig below.

Figure 18. Toplis algorithm

If 4-Cl >H then 3,4 Cl

If 3,4 Cl > 4 Cl then 3, 4 Br, I CF3

If 3,4 Cl > 4 Cl then Br, I =, NO2

If 4-Cl<H then 3,4 Cl then 4 OMe

If 4-OMe > H then 4 NMe2

If 4-NMe2 > 4 OMe then 4 NEt2

If 4 NMe2 < 4 OMe then NH2, OH

If 4OMe < H then 3 Cl

If 4 Cl =H then 4Me

If 4Me > H, 4 cl then Et Pr iPr, Bu, tBu

If 4 Me < H, 4 Cl then 3 Cl

If 3 Cl > $ Me then 3,5 Cl, 3Br, 3 I

If 3 Cl < 4Me then 3 CH3 3 NMe2

4 F, 4 NO2, 4 CN, 4 CONH2

The algorithm is as follows:

First *we start with substituents H and 4-Cl*

Then according to the biological results the next step is taken.

Next

- if 4-Cl substititution increases the activity when compared with the not substituted analogue then make 3,4 dichlorosubstituted compound.

- *if 4-Cl substititution decreases the activity when compared with the not substituted analogue then make methoxy substituted compound.*
- *- if 4-Cl substititution is as potent as not substituted analogue then methyl substituted compound.*

This and the other *step can be found in J.Med Chem 15, 1007 (1972)*

Nonlinear projections – mapping

These methods rely on the displaying of the original data in the new coordinate system which is not the linear combination of the original coordinates. There are many possibilities of such transformation. The most popular method used now was proposed by Samson. The transformation is done in such a way that it tries to preserve interpoint distances as far as possible.

The difference between the linear and nonlinear mapping is pictorially shown in picture below.

Figure 19. The difference between the linear and nonlinear mapping

No single linear projection gives as much information as the nonlinear one. In the example given above the projection provides the most important information about the top, bottom and the body, the same picture is not possible in the linear projection.

The mathematical procedure of the calculation of the new coordinate system is given below

Each pattern defined by the coordinates X_i –X_n in a given multidimensional coordinate system is projected to the new two dimensional coordinate system with coordinates X_1' and X_2'. The projection is done in such a way that the error

$$E(\rho) = \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{(d_{ij}^*)^\rho}$$

is minimal.

Where d_{ij} are the distances between patterns in the original system and d^*_{ij} in a new system.

The physical meaning is that the changes in distances are the minimal possible/ as minimal as possible.

Example

30 antibiotics were tested against 27 bacteria species so the antibiotics are spanned in 27 the dimensional space. Such space is difficult to analyze.

Antibiolics:

1. Penicillin-G
2. Methicillin
3. Oxacillin
4. Ampicillin
5. Cloxacillin
6. Sulbenicillin
7. Carbenicillin
8. Cephalothin
9. Cephaloridine
10. Cephaloglycin
11. Cephoxitin
12. Cephazolin
13. Cephalexin
14. Streptomycin
15. Neomycin
16. Kanamycin
17. Paroxomycin
18. Gentamycin
19. Vistamycin
20. Erythromycin
21. Spiramycin
22. Kitasamycin
23. Oleandomycin
24. Chlortetracycline
25. Oxytetracycline
26. Tetracycline 1
27. Tetracycline 2
28. Pyrrolidinomethyltetracycline
29. Colistin
30. Polymyxin B

Bacteria:

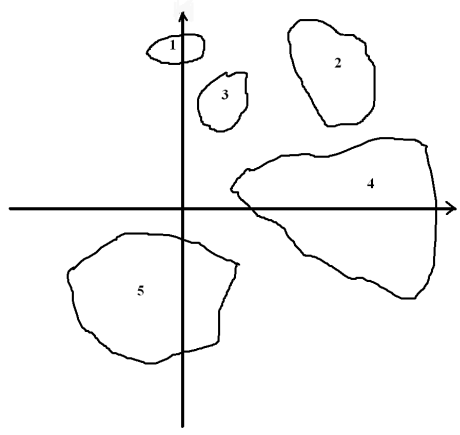
1. Escherichia coli
2. Shigella flexneri
3. Salmonella eneteretidis
4. Salmonella typhimurium
5. Proteus vulgaris
6. Proteus mirabilis
7. Citrobacter freundi
8. Haemophilius aleci
9. Yersina enterocolitica

10. Klebsiella pneumonia
11. Entereobacterium cloacae
12. Entereobacteriumaerogenes
13. Serratia marcescens
14. Pseudomonas aeruginosa
15. Pseudomonas capacia
16. Pseudomonas multophilia
17.Pseudomonas putida
18.Achremobacter xylosoxidans
19.Acinebacter anitratus
20.Agrobacterium feccalis
21.Flavobaterium meningosepticum
22.Staphylococcus aureus
23.Staphylococcus epidermidis
24.Streptococus pyogenes
25.Streptococus pneumoniae
26.Streptococuus fecalis
27.Baciluus subtilis

Nonlinear mapping to two dimensions are shown below. The Data are clustered into six groups which defines six groups of antibiotics.

1. Aminoglycosides(14-19)
2. Cephalosporins(8-13)
3. Makrolides (20-23)
4. Pencillins(1-7)
5. Peptides(29-30)
6. Tetracycline(24-28)

Figure 20. Nonlinear mapping to two dimensions of antibiotics. For details see text



The diagram might be used for the classification of an unknown antibiotic. When the unknown antibiotic is to be classified then it has to be exposed to the bacteria set then transformed into the two dimensional system and analyzed in which cluster it is found.

15. Distance in a pattern space.

The similarity of the patterns in the pattern space can be measured as a distance between patterns. There are many distance definitions. The most frequently used are given below.

Minkovski

$$d(x,y)=L_p(x,y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}$$
$$p>0$$

Manhattan city block

$$d(x,y)=L_{p=1}(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

Euclides

$$d(x,y)=L_{p=2}(x,y)=\|x-y\|=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Mahalanobis

$$d(x,y)=-\sum_{i=1}^{n}z_i x_i y_i$$

$$d(x,y)=\frac{1}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}}\sum_{i=1}^{n}z_i x_i y_i$$

$$z_i=\sqrt{\frac{\lambda_i}{\lambda_i+\alpha^2}} \qquad \alpha=0,25$$

$$z_i=\sqrt{\frac{\lambda_i}{\lambda_i+\alpha^2}} \approx \sqrt{\frac{1}{\lambda_i}}$$

Angle

$$d(x,y)=-\cos(x,y)$$

$$\cos(x,y)=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{m}x_i^2\sum_{i=1}^{m}y_i^2}}$$

Euklides quadratic with Euklides average squared error.

$$d(x,y)=L^{2}_{p=2}(x,y)=SSE=\|x-y\|^2=\sum_{i=1}^{n}(x_i-y_i)^2$$

$$d(x,y)=\frac{1}{n}L^{2}_{p=2}(x,y)=MSE=\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2$$

16. Classification

Centers of gravity.

If all pattern points are grouped into several, more or less separated groups then each group can be represented by a center of gravity of that group. The center of gravity in each dimension is calculated as the mean of all patterns belonging to that group. C_j (mean for dimension j) for i points in n dimensional space (with coordinates x1….xj) is calculated as follows:

$C_j = 1/n \Sigma x_{ij}$ for i=1 to the number of points (patterns)

Such a calculation has to be repeated for every dimension. As a result the point which represents the whole group in the n dimensional space is calculated. The Figure below shows the center of gravity in the two dimensional space.

Let's have the set of points: 2,3 2,4 3,5 5,3 3,5. Then the center of gravity is characterized by the point with coordinates 3,4 since 1/5(2+2+3+5+3)=3 and 1/5(3+4+5+3+5)=4.

Classification by measure to the center of gravity.

When all data points have been assigned to the defined groups/classes, then the new point is classified by measuring the distance to the centers of gravity for all
the defined classes in the pattern space. The point belongs to this class to which the calculated distance was the shortest.

The example below comes from the work done by Ilona Dudka, Barbara Kossowska, Roman Gancarz and co. The whole dataset of the people exposed the heavy metal environment was divided into ten classes depending on the distance from the center of gravity of the data obtained from volunteers not exposed to the risky environment.

Problem definition.

For 50 volunteers and 400 workers in the cooper mine in Lower Silesia 38 physiological parameters were measured. Thus every worker is represented by a point in 38 dimensional space. The clinical parameters measured are presented below:

Table 13. Clinical parameters of workers in the cooper mine in Lower Silesia

Nr	***Parameter***	***Unit***	***Reference range***
1.	magnesium in serum	µg/ml	18 - 31
2.	calcium in serum	µg/ml	80 - 105
3.	cooper in serum	µg%	80 - 150
4.	zinc in serum	µg%	70 - 160
5.	lead in whole blood	µg/l	to 500
6.	cadmium in whole blood	µg/l	to 5,0
7.	cadmium in urine	µg/g CREA	to 5,0
8.	arsenic in urine	µg/g CREA	to 35
9.	FEP in erythrocytes	µg/100 ml eryt.	to 100
10.	WBC	K/µL	4.00 – 10.0
11.	NEU	K/µL	1.50 – 7.00
12.	LYM	K/µL	1.00 – 3.70
13.	MONO	K/µL	0.00 – 1.00

14.	EOS	K/µL	0.00 – 4.00
15.	BASO	K/µL	0.00 – 0.100
16.	RBC	M/uL	3.50 – 5.00
17.	HGB	g/dL	12.0 – 17.2
18.	HCT	%	33.0 – 49.0
19.	MCV	fL	81.0 – 98.0

Table 14. Reference range of all measured parameters - continuation.

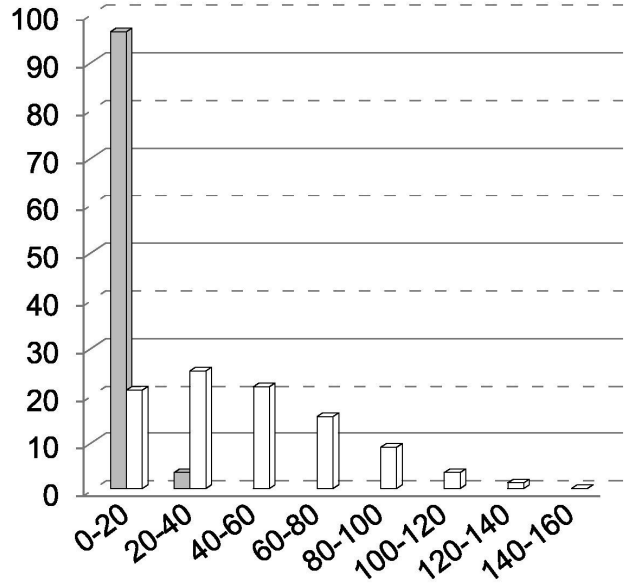
Nr	***Parameter***	***Unit***	***Reference range***
20.	MCH	pg	26.0 – 43.0
21.	MCHC	g/dL	31.0 – 37.0
22.	RDW	%	11.6 – 15.8
23.	PLT	K/µL	130.0 – 400.0
24.	MPV	fL	0.00 – 99.9
25.	Glucose in serum	magnesium %	65 – 105
26.	Creatinine in serum	magnesium %	0.60 – 1.35
27.	Molar creatinine	µmol/l	53 - 120
28.	Completed iron in serum	µmol/l	9.0 – 28.6
29.	TIBC in serum	µmol/l	44.8 – 73.4
30.	UIBC in serum	µmol/l	26.9 – 53.1
31.	Saturation in serum	%	20.0 – 45.0

32.	Trasferrin in serum	g/l	2.00 – 4.00
34.	PSA total	ng/ml	under 4.00
35.	PSA free	ng/ml	-
36.	PSA free/PSA total	-	-
37.	HbA1c	%	4.40 – 6.40
38.	Ferratine	ng/ml	-

It was necessary to divide the group of workers into 10 different groups with progressive deviation from the standard physiological stage and to choose the representative worker from each group for further biochemical studies. The aim of the work was devoted to the search of biomarkers indicating changes in the prolonged exposition to heavy metals.

The problem was solved in such a way that the center of gravity for the volunteer group was defined as the standard physiological stage. Then the distance of all workers to that center of gravity was calculated and the value for the largest distance was divided by ten. In this way the workers could be classified to ten classes. The First class contained all workers distant from the center of gravity of volunteers by 0-1/8 of the largest distance, second 1/8 -2/8, third 2/8- 3/8 and so on. The population of the workers in each class is shown in the figure below.

Figure 21. The ten classes of workers in the cooper mine in Lower Silesia



The Green bar represents the volunteers not exposed to heavy metals, yellow-workers exposed to heavy metals. Most of the volunteers are grouped in the first class, the class closest to the calculated reference center of gravity. However there is a small amount of volunteers which is characterized by some distortion from the standard physiological state. There are also some groups of workers characterized by the standard physiological parameters.
This classification allowed to choose a representative worker from each group and to study the progress of changes. It was not possible to find such representative workers without the above described approach.

classification with potential functions

A powerful method in decision making about the class membership of a new point x in a defined pattern space is the potential function method.
In the learning step for all known patterns with the known score the "electric charge" is the set corresponding to the score of that pattern.

Each “charged” pattern influences the environment in such a case that the charge at every point at a distance d is given by the function:

$Q(d)=1/(1+qd^2)$

or

$Q(d)= exp\ (-d^2/q)$

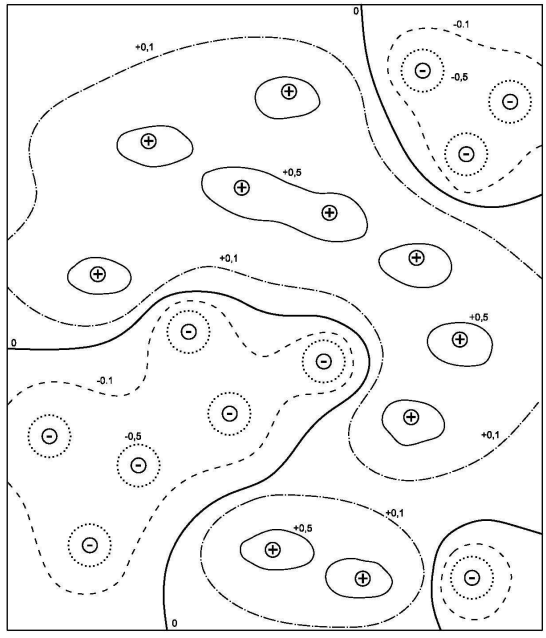
Then the potential for every point in multidimensional space is calculated as the sum of the effects from each pattern.

The Explanation below is given for the two parameters problem i.e. when the score, for example, as a a biological response is given as function Y=f(p1, p2)

The figure below is an example of applying a potential method for a set of patterns belonging to two different groups. Every pattern in the first group was charged with positive values +1 and in the second with the negative charge -1. The result of the superposition of charges is given as a d+1 dimensional surface in a form of a potential contour lines. The solid lines represent the borderline between classes. The values of the contour lines indicate the centers of each class.

Then the potential of a new pattern to be classified is determined upon its position and the value of the potential at that point.

Figure 22. Example of classification by potential method



The simplex method

The center of gravity calculations was to find the point which represents the particular class of patterns. Another approach to look for the best representation is the simplex algorithm.

Let's have a set of compounds described by the two physicochemical parameters (it is much simpler to explain the problem in the small dimensional pattern space, but there is no problem to expand this method to the higher dimensional pattern space).

Let each compound be characterized also by a biological response (let say MIC – minimal inhibitory concentration) which is a measure of its potency against pathogenic bacteria. The problem we want to solve now is to find the point in the two dimensional physicochemical space with the highest response. Such a point could serve as a reference point.

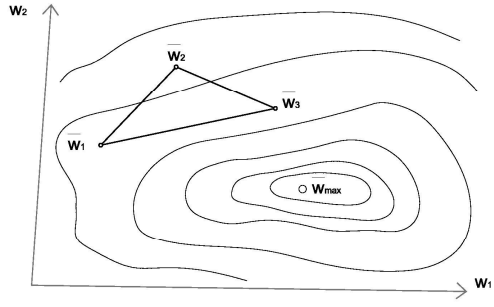
Graphically the problem can be defined as climbing uphill in order to find the top of the hill.

Lines on the graph represent the points of the same activity in a similar manner like for example isobars.

Step 1 define three points w1,w2,w3

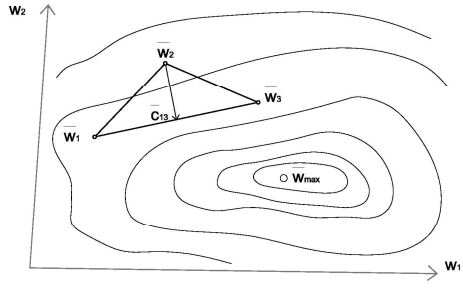
Step 2 draw a triangle w1, w2, w3

Figure 23. Simplex metod step 2



Step3 find the vertice with the lowest score (in our example it is w2)

Figure 24. Simplex metod step 3



Step 4 draw a line from the vertice with the lowest score and passing through the middle of the edge opposite to the vertice with the lowest score -c13.5 Find the point w4 such that w4c13 =w2 c13

Step 6 Draw the new triangle w1, w3, w4

Step 7 Go to step 3

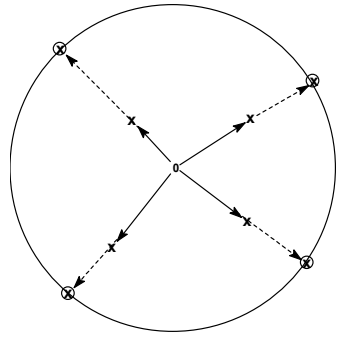
The algorithm is stopped when in two subsequent steps the point with a higher score is not found.

Once the reference point is found then the distance to that point is a measure of similarity. If the surface contains many maxima then all of them can be found and the tops of them defined as the centers of the classes.

Modelling by hypershere

Classification by the distance to the center of gravity can be realized by assigning the critical distance. Such a procedure will define the circle, sphere or hypersphere depending on the dimensionality of the pattern space. The classification then is simple is the object inside or outside the declared geometrical figure. See the diagram:

Figure 25. Modelling by hypershere

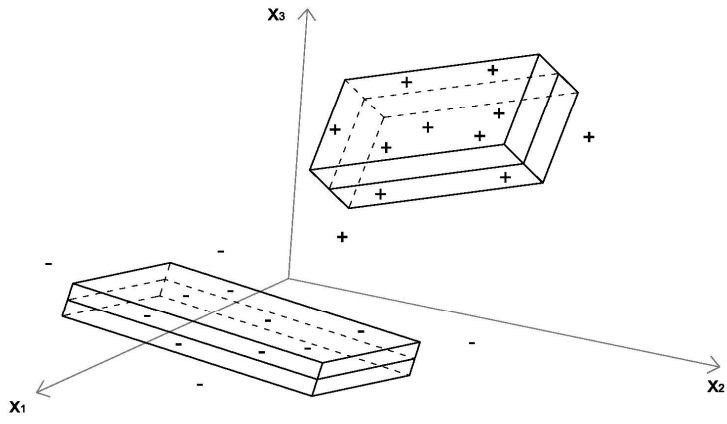


Simca

Simca means the **s**tatistical **i**solinear **m**ulti**c**ategory **a**nalysis. In this method many dimensional hyperplanes are defined for each class as shown in the figure below.

The Mathematical description is complicated. More information can be found in Materials

Figure 26. Simca classification



kNN classification

In the ***k*-nearest neighbours algorithm** (*k*-NN) the classification of the objects is based on the analysis of the closest neighbours in the pattern space. Once the correct classification is made in the training step then a new object is classified by the examination of a certain number of the closest neighbours (1 in 1kNN, 3 in 3kNN,5 in 5kNN).

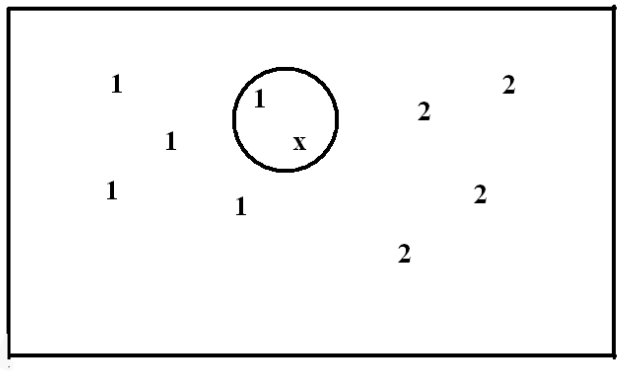
For example

Let's have two classes – "1" and "2".

1kNN method example

In the picture below the new object is classified as the "1" class since in 1kNN the closest neighbour was the "1".

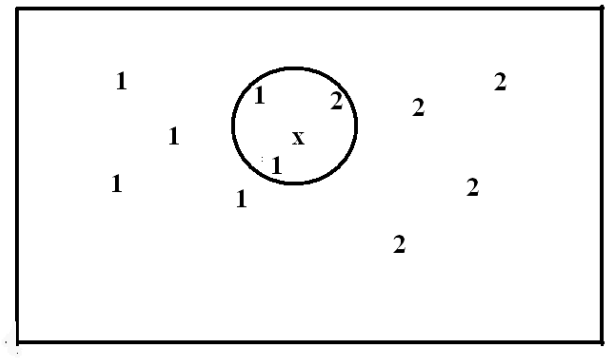
Figure 27. 1NN classification



3NN method example

In the picture below the new object is classified to the triangle class since in 3kNN the closest neighbours were two “1” and one “2”.

Figure 28. 3NN clasification



17 MST – minimal spanning tree

Before describing the minimal spanning tree method a few definitions from the graph theory are necessary.

graph

In mathematics and computer science, the **graph theory** means the study of *graphs*: the Graph is a mathematical structure which describes pairwise relations between objects.

The graph consists of vertices and edges. The last connected pairs of vertices.
A graph may be *undirected*, when there is no distinction between the two vertices associated with each edge, or *directed* when the edges indicate the direction from one vertex to another.

The graph may be represented as a set of pairwise relations or as a graph.
Two possible representations of a graph relations.

Figure 29. Graph example. Representation 1

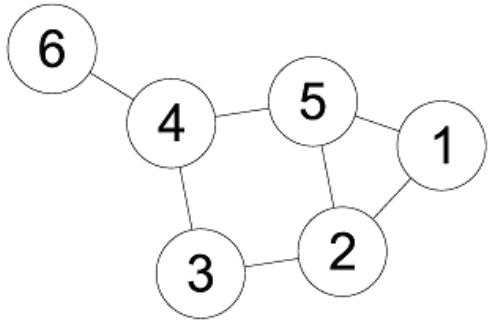


Figure 30. Graph example. Representation 2

1,2

1,5

2,5

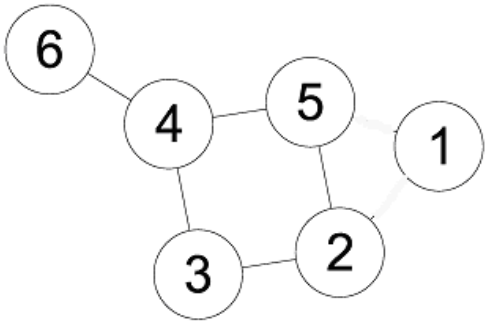
2,3

3,4

4,5

4,6

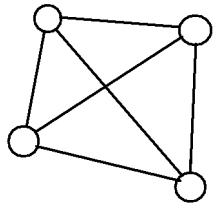
Figure 31.Graph with isolated vertices



A Full graph is the graph in which every vertex is connected to another vertex.

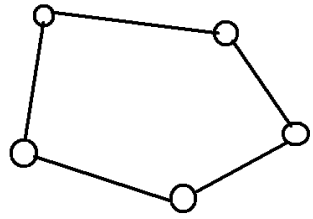
Example is K4 graph with 4 vertices

Figure 32. Full graph



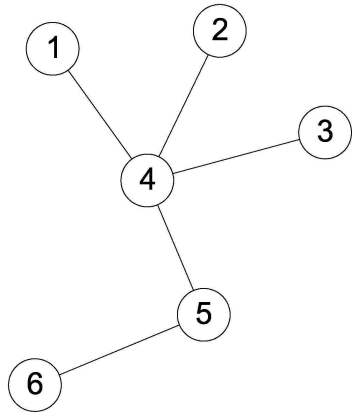
A *cycle* in a graph is a closed sequence of vertices. *Closed* means, that it begins and ends in the same vertex. The Cycle may contain from 1 to n vertices. Above the graph with 2 cycles is presented:

Figure 33. Cycle



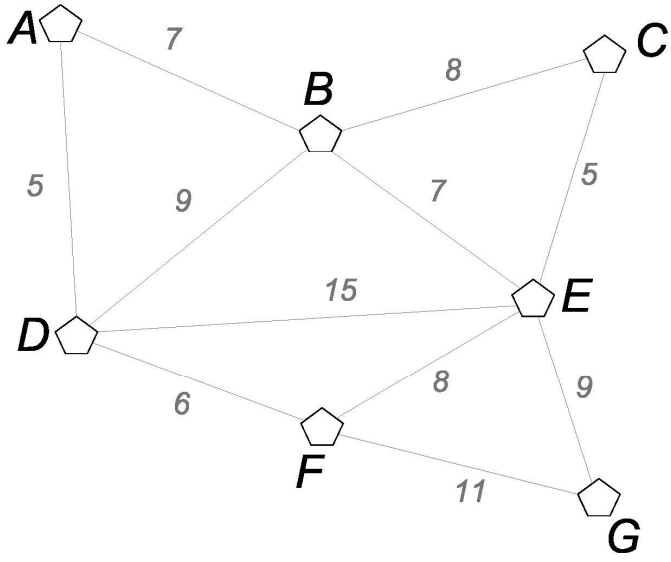
A **tree** is a connected graph with no cycles. A vertex of a degree 1 is called a **leaf**,

Figure 34. Tree



A Graph or tree is weighted if the edges are weighted according to some criteria

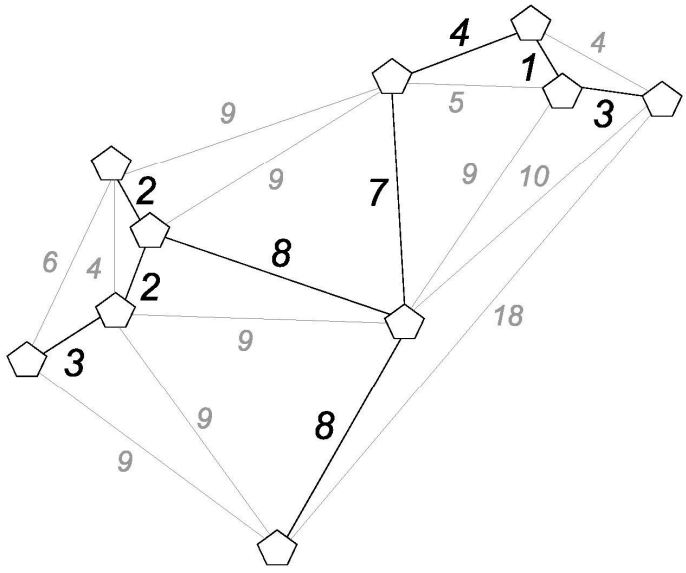
Figure 35. weighted graph



Spanning tree

The Spanning tree is constructed from a given graph by removing some edges to get the tree. There are many possibilities to construct the spanning tree from a given graph. One of the example is given below.

Figure 36. Minimal spanning tree



The Minimal spanning tree is the tree which is defined for a set of patterns in such a case that the sum of edges is minimal. The above tree is minimal spanning tree (MST)
There are two algorithms for making a minimal spanning trees - Prim's and Kluskal's. In both cases the start is from a full graph

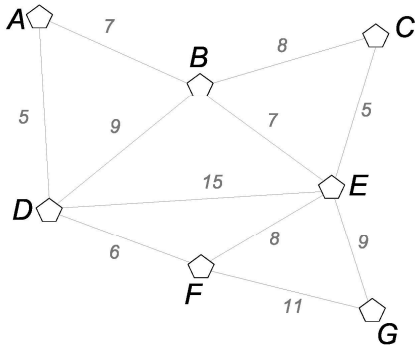
Prism algorithm

1.The first step is to find the edge with the lowest weight
2. Join the corresponding vertices and make a branch of the tree
3. Find the edge of the lowest density but such that one of the vertices is a part of the tree
4.Check if joining the vertices will not result in formation of the cycle
5.Draw the next branch
6.If not all points are a part of the tree go to 3

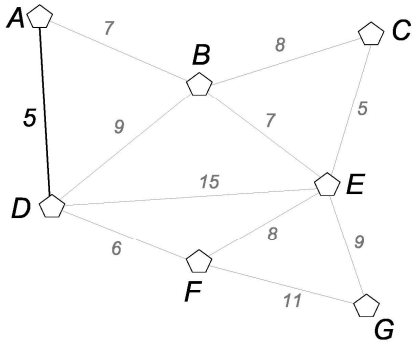
Example

The following pictures show the steps in developing the minimal spanning tree from a chosen graph. The starting graph should be in general the full graph, however for the clarity not all the edges of a full graph are written. The red line appearing in pictures means that such edges can not be considered in the next step of the minimal spanning tree construction as they lead to cycles.

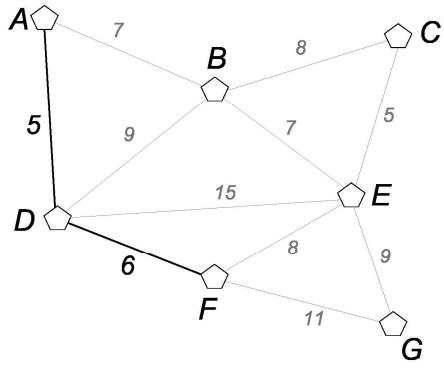
Figure 37. Ilustration of Prism algorithm



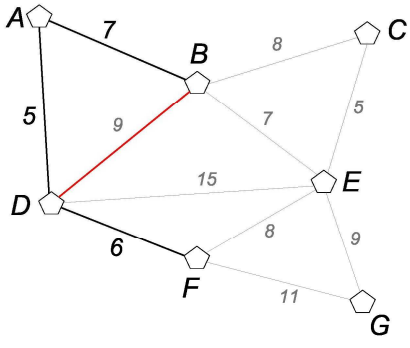
1.



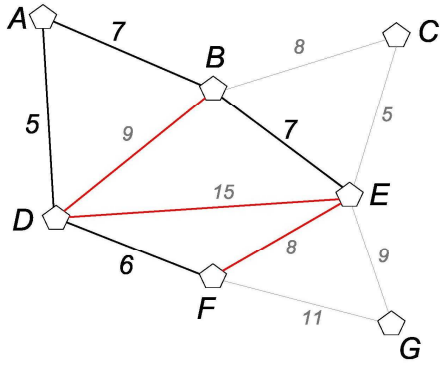
2.


A
B
C
D
E
F
G
7
8
5
9
7
5
15
8
9
6
11

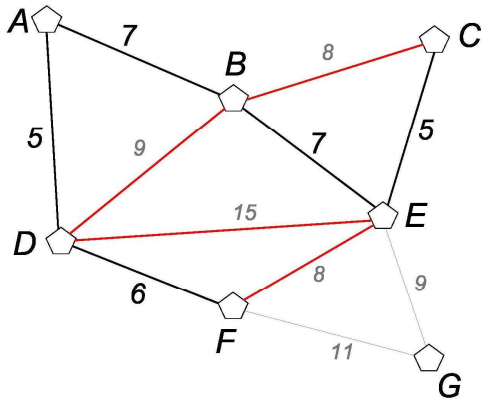
3.


A
B
C
D
E
F
G
7
8
5
9
7
5
15
8
9
6
11

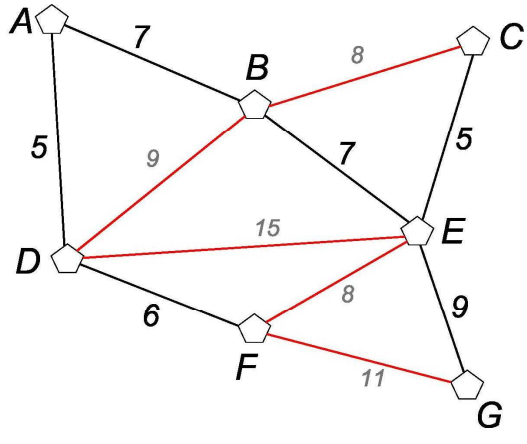
4.


A
B
C
D
E
F
G
7
8
5
9
7
5
15
8
9
6
11

5.



6.



7.

F D A B E G

Kruskal algorithm

a. *Find the edge with the lowest weight*
b. *Join the vertices*

c. *Find the next edge with the lowest weight*
d. *Check if after joining the new vertices no circle is formed*
e. *if the answer in step 4 is YES find the new edge with the next lowest weight*
f. *otherwise join the edges*
g. *repeat until all the vertices are a part of the tree*

diameter of MST

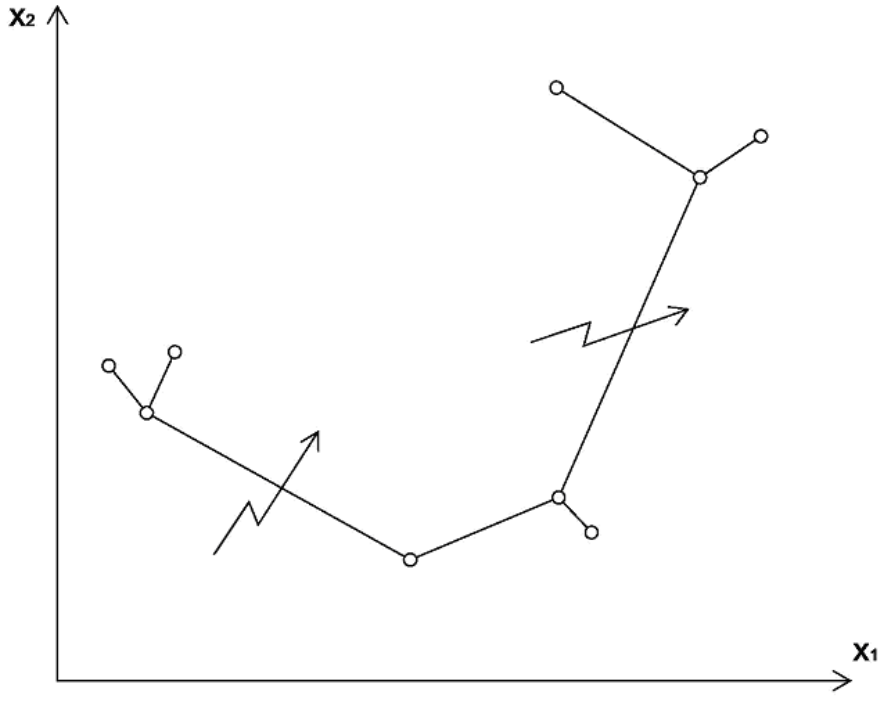
The Diameter of the MST is the longest path which can be obtained from the minimal spanning tree. In the exaple above it is: *F D A B E G*

The patterns on the path defined by the diameter of MST can be treated as the most representative objects along the "path change" of the dataset. For example, analyzing a large set of biologically active compounds one can select only those that lay on the diameter and analyze which properties are changing the most when moving from one end to the other end.

subgraphs

The minimal spanning tree allows to divide the whole dataset to as many subsets as we want. The Picture below shows the procedure which results in the division into three subsets by removing the edges with the two highest weights.

Figure 38. Division of MST to three subgraphs



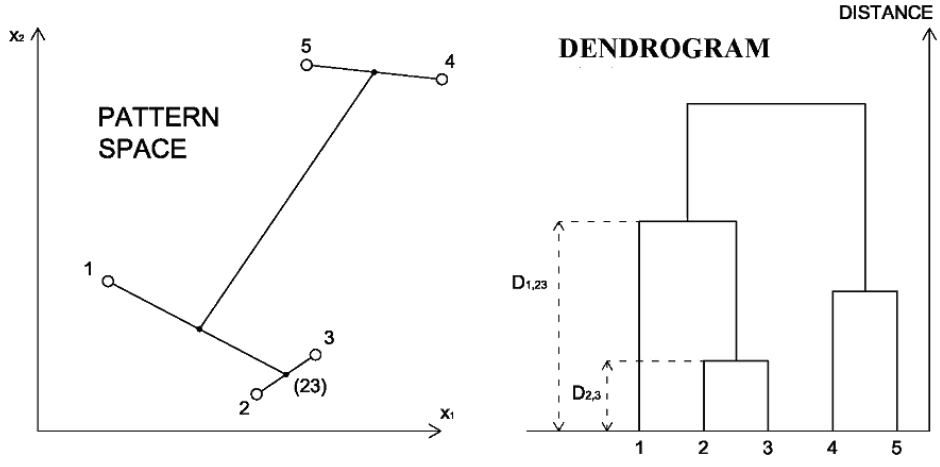
Such a procedure aloows to dividecan into three classes – most active, active, not active.

18. Clustering of the data

In the picture below the pattern space with five patterns are shown. Clustering algorithm starts with the calculation of all possible distances of all patterns. In the picture below the lowest distance found was between points 2 and 3. The points are connected and replaced by one new point at the position depending on the chosen method (see below). The number of points is then reduced by one. In the next step the the new edge is formed between the two closest points. The new point which replaced the two points plays the same role as the other points in the pattern space. The algorithm ends when all points are joined and form a tree.

The tree can be drawn as a dendrogram, see the figure below on the right. The horizontal lines joining the points are drawn at the level corresponding to the distance between points (the vertical scale). The example of clustering of the data is given below.

Figure 39. Two representation of a clustering results.



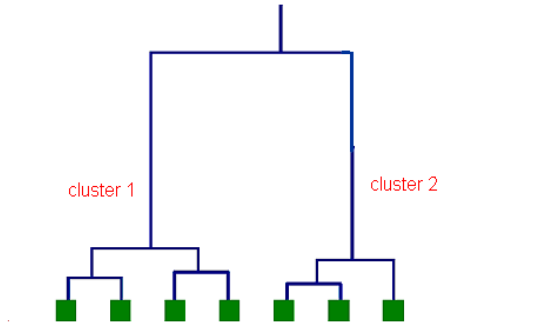
When the dendrogram is calculated some questions can be asked

1. divide the whole dataset to a three subsets. Picture below show the horizontal line drawn which makes three distinct clusters (subclusters)

Figure 40. Example of a cluster

2. How many subclusters can be found in the dataset?. Such division is made after the declaration that when the two points joint with distance, for example, four times bigger than the distance in the previous joint it means that the two separate clusters are joint. See the picture below.

Figure 41. Example of a cluster



There are many possibilities of drawing the dendrogram. They differ from each other in the method of distance measurement and also the method of calculation of the coordinates of the new point obtained after joining points when each of them is representing represents the cluster itself. There is a difference if we join two clusters with almost the same population of patterns or when we join a cluster containing only a few patterns with a highly populated cluster . Intuitively the new point should lay close to the last cluster. For these reasons there are many ways of the calculation of the coordinates of a new point. The detailed description falls outside the scope of this study guide.

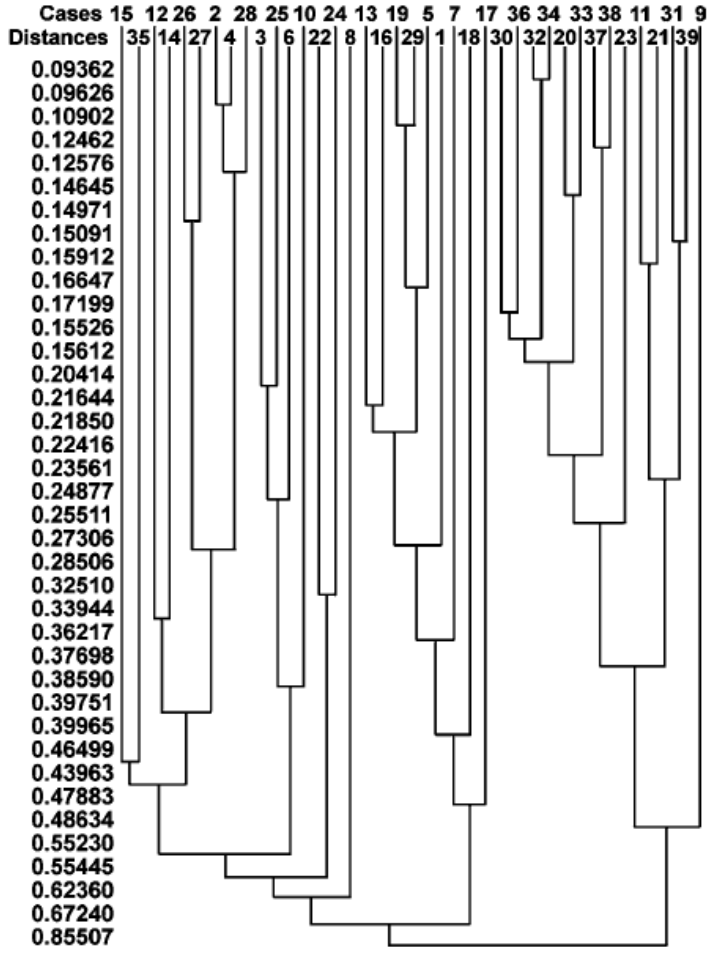
Example of the cluster analysis

In the Paper "Fucosylation of serum glycoproteins in lung cancer patients" by Barbara Kossowska, Mirosława Ferens-Sieczkowska, Roman Gancarz, Ewa Passowicz-Muszyńska3 and Renata Jankowska the serum of patients with lung cancer were diagnosed and compared with the serum of the healthy people. The level of four proteins of the acute phase and four degrees of their fucosylation were measured and used for the diagnosis. Thus every patient formed a pattern in the eight dimensional pattern space. All patterns were subjected to the cluster analysis.

Clinical subjects

There were 18 patients (Nos. 1–18) diagnosed with non-small cell lung cancer (NSCLC) at different stages of cancer development, 11 patients (Nos. 19–29) diagnosed with small-cell lung cancer (SCLC) and for the comparison a serum of the control group of 10 healthy blood donors. (Nos.30–39). The similarities between the points were calculated using different clustering parameters, including different metrics (Euclidean, Manhattan city block, and Canbery), on the original data points, as well as the scaled points (25, and references therein). The mean values for each point in the cluster were calculated as the mean average or the weighted mean average.
The obtained dendrogram is shown below.

Figure 42. Cluster analysis of LGOM patients. For details see text.



The analysis of the dendrogram allowed to classify the patients into five group.

19 Artificial neural network

The whole presented study guide is devoted to methods of data analysis. The fundamental task we want to learn is how to extract the knowledge from a set of observations. The oldest approaches of human studies were statistical methods and pattern recognition methods described in the previous chapters. The Human brain however works in a different way. The mathematical models and used algorithms which mimic the information processing by the human brain are named neural networks. We can say that a neural network is a computational method inspired by studies of the brain and nervous systems in the biological organisms.

The aim of this chapter is to provide basic principles of such method.

At first glance we can imagine a neural network as a black box which yields some outputs (new information) based on several inputs (data). The input can be from a market, medical investigation, spectrometric analysis and many others. For most of users it is not necessary to understand the processes inside the black box however some basic information about such processes will be provided.

History

The concept of neural networks started in the late-1800s
The first step toward the artificial neural networks was done in 1943 by Warren McCulloch & Walter Pitts.
Donald Hebb wrote *Organization of Behavior* in 1949.
First computer simulation of the biological neural network took place in 1950's
McCullogh and Pitts defined the first artificial neuron 1951
The Comparison of artificial intelligence and neural networks 1956. (both ANN and AI can ***"think"***)
John von Neumann the first imitation of simple neuron functions
The work on perceptron by Frank Rosenblatt 1957.

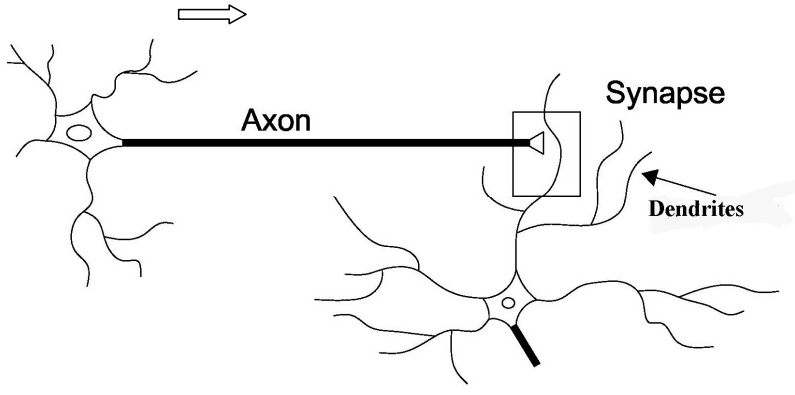
What is neural network?

This is a system that simulates intelligence that occurs in animal brains and which will transform a set of n inputs (observations) into n outputs (conclusions).
The transformation of the data is performed in basic processing units called neurons. The set of neurons with the interconnections forms the neural network.
In some publications the term perceptron is used interchangeably for a single neuron as well as network. To avoid the misunderstanding this term will not be used in this work.

Real and artificial neurons

In biological systems the information is passed from one neuron to another neuron and depending on the system we can state that in many cases the information is analysed, modified when necessary and passed to the next neuron or effector (output).
Dendrites receive the signal and axon passes it to synapse. The degree of change of the signal at synapse is called the synaptic strength.
Schematically the biological neuron net can be represented as in the picture below:

Figure 42. Neuron

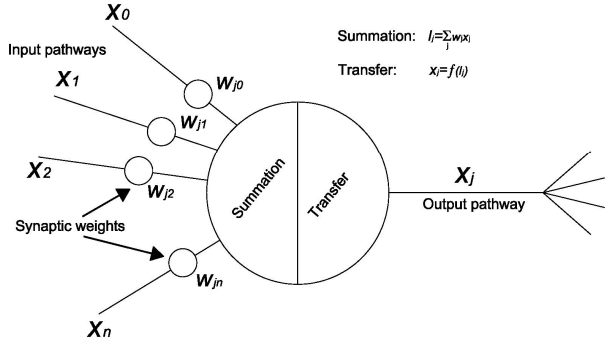


The Biological neuron can receive many signals at one time and then the collective output is formed. The net input into the artificial neuron (NET) having m synapses can be calculated according to the equation below where the synaptic strength is modeled by the weights (w_i).

$Net= s_1w_1+s2w2+s3w3+\ldots\ldots s_nw_n$

Schematically it can be represented as in the picture below.

Figure 43. Artificial neuron



The differences between ANNs and Computers can be specified in the following ways:

ANN	Computers
Contain pre-defined library of data (experiences).	No pre-defined data provided.
Rely on a large number of already existing examples of the behavior they are supposed to exhibit.	Do not rely on “experiences”.
ANNs are useful even when we can’t formulate an algorithm.	Fail to give satisfactory results when lacking the proper algorithm.
Computes arithmetic functions very fast.	Better suited to massive parallelism.
Have a high fault tolerance. Cannot act on very noisy data but powerful in deducing inferences.	Have minimal fault tolerance. Can operate on noisy data as well. However, deduction of inferences is generally very weak.

The areas where neural nets may be useful

Signal processing
Control
Pattern Recognition
- pattern association
- pattern classification
- regularity detection
- image processing

Medicine
Modeling and simulating the functions of the brain and sensory organs
Signal processing, Bioelectric signal filtering and evaluation

Classification and interpretation of physical and instrumental outputs.
Prediction, providing prognostic information based on retrospective parameter analysis.

Other
Speech Recognition
Speech Production
Business
speech analysis
optimization problems
robot steering
processing of inaccurate or incomplete inputs

Neural networks in medicine

Artificial Neural Networks (ANN) becomes a 'hot' research area in medicine. It is believed that they will increase application in these areas in the next few years. Now it is applied mostly to recognizing diseases from cardiograms, ultrasonic scans, etc. Neural networks are learned by examples of many variations of the disease.

Modelling and Diagnosing the Cardiovascular System
Neural Networks are used to model the human cardiovascular system. First the model of the cardiovascular system is built based on the relationship among physiological variables (i.e., heart rate, systolic and diastolic blood pressures, and breathing rate) at different physical activity levels. Then the prediction can be made.

Instant Physician
A More general application is called the application developed in the mid-1980s called the "instant physician". When a number of medical records, including information on symptoms, diagnosis, and treatment is collected then the "best" diagnosis and treatment can be stated.
.
Electronic noses
The electronic nose should identify the subject, individual, compound based on data collection in the training step..

OTHER APPLICATIONS IN MEDICINE

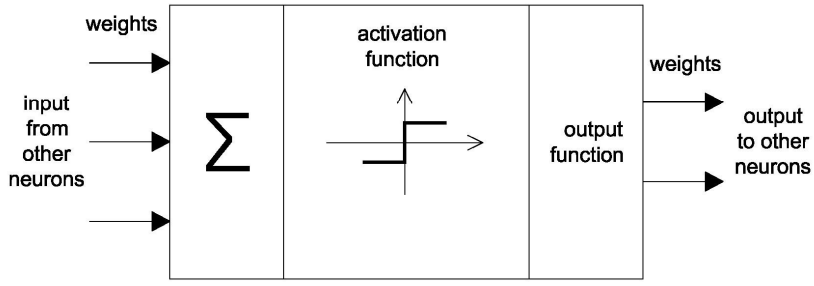
Between 1990 and 1997 applications of neural networks were introduced in nearly 2000 papers

Table 15. Application of neural network in medicine

Discipline	Application field
Cardiology	diagnostics, prognostics
ECG	diagnostics
Intensive care	prediction
Gastroenterology	prediction
Pulmonology	diagnostics
Oncology	diagnostics, prognostics
Paediatrics	diagnostics
Neurology	signal processing, modelling
EEG	diagnostics
Otology-Rhinology-Laryngology	signal processing, modelling
Obstetrics and Gynaecology	prediction
Ophthalmology	signal processing, modelling
Radiology	signal processing (x-ray, US, CT)
Clinical chemistry	signal processing, diagnostics
Pathology	diagnostics, prognostics
Cytology	diagnostics, re-screening
Genetics	diagnostics
Biochemistry	protein sequence, structure

Model of neuron

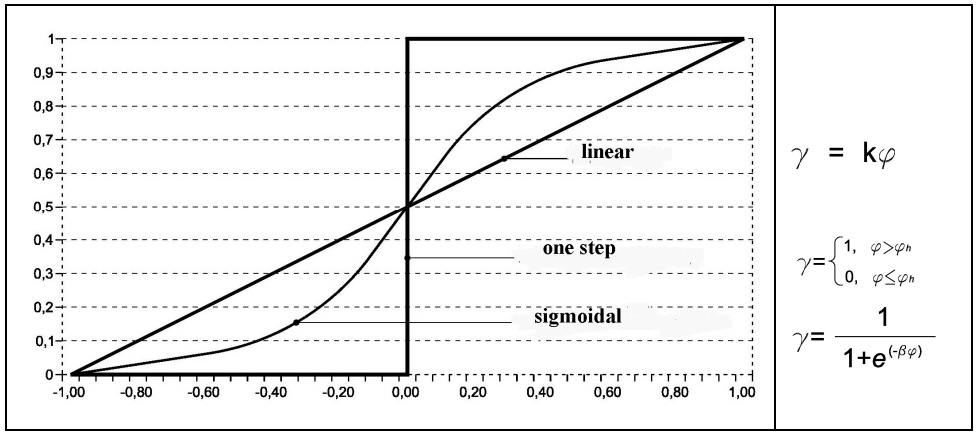
Figure 44. Model of single neuron



The Structure of a neuron in a neural net. The net input to the neuron undergoes the transformation.

The activation functions can be various. Some examples are given below.

Figure 45. Single neuron activation function



Linking neurons- neural network formation. Types of Neural Nets

Neural networks can be distinguished by:

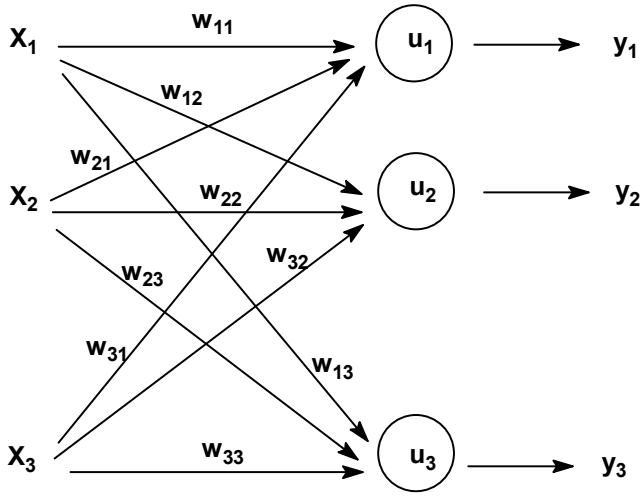
- their type (feed forward or feedback)

- their structure
- the learning algorithm they use

Single layer network

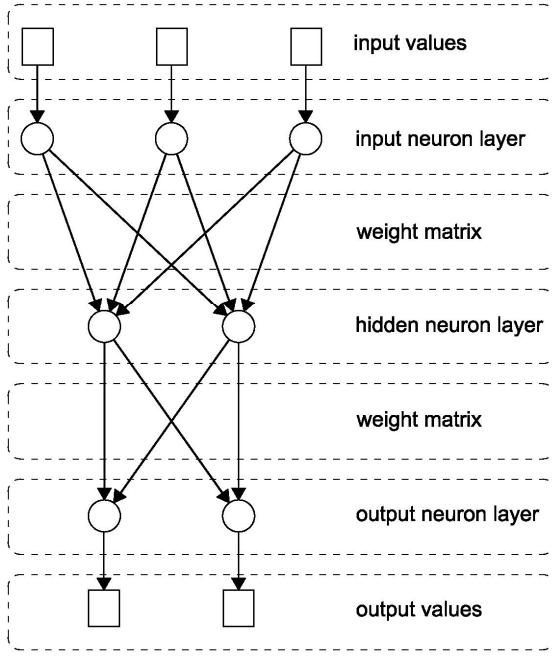
A single-layer neural network consists of a set of units organized in a layer. Each unit *Ui* receives a weighted input *xj* with weight *wji*. The **Figure** shows a single-layer linear model with *m* inputs and *n* outputs.

Figure 46. Single layer network



Three Layers Neural Net

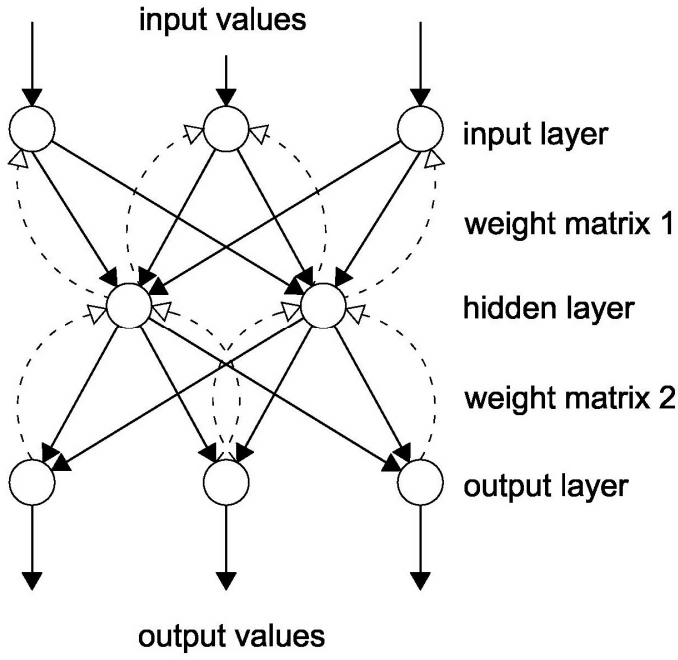
Figure 47. Three Layers Neural Net



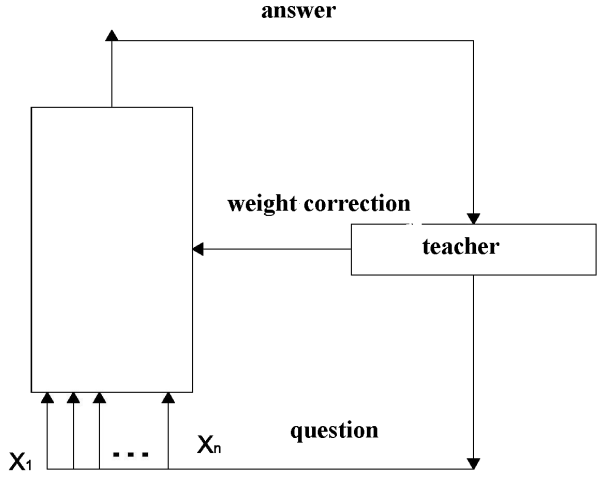
Backpropagation Neural Networks (BPNNs)

BPNNs are one of the most common neural network structures, as they are simple and effective, and have found home in a wide assortment of machine learning applications, such as the character recognition. In general the message is passed from one layer to the next as in the previously described networks but additionally in this case the message is also passed back to the previous neuron and weights are modified.

Figure 48. Backpropagation Neural Networks



How the neuron is thought?



Kohonen Networks

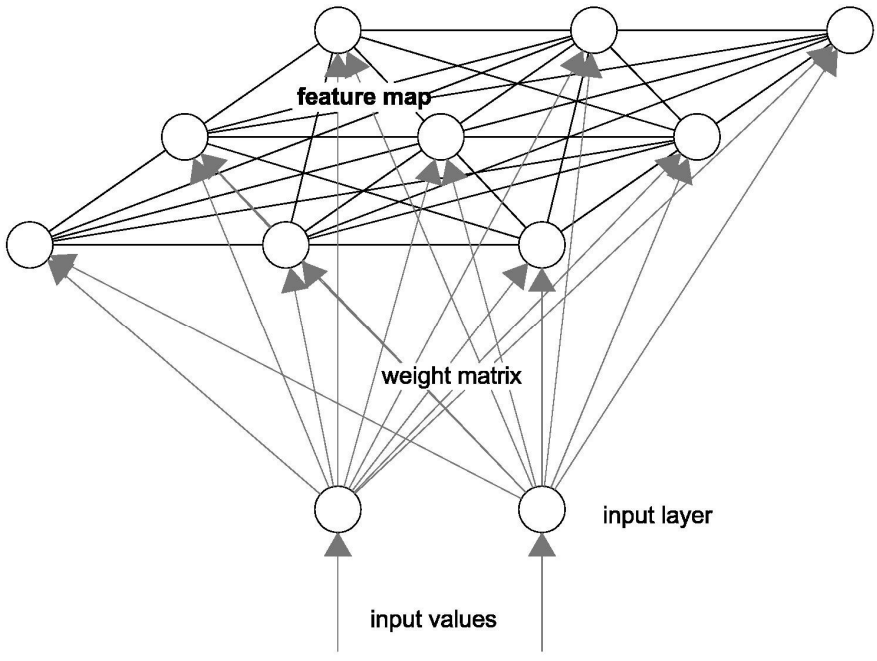
We often think about the data set in a way of looking for the numerical values. There is another way of analyzing the data set, topological one which deals with the possible relation between objects using a neighbourhood function to preserve the topological properties of the input space.

Such information in general is given by the mapping of a multidimensional dataset to a smaller one The compression should be done with the smallest possible loss of

the information. Some of the projection methods were given before. Here another approach is given , a mapping of multidimensional information into the two dimensional plane of neurons, a method introduced by Kohonen and called the self organized maps approach (SOM).

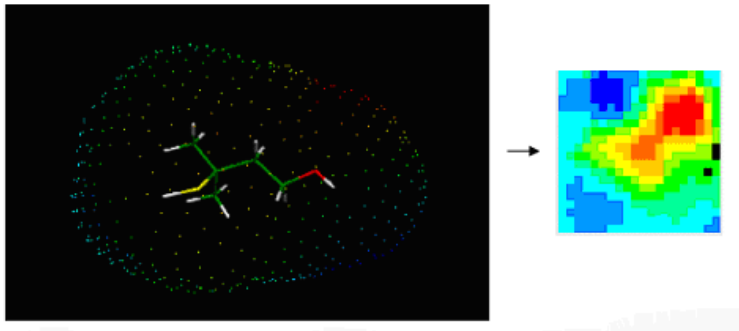
Without going into details a **self-organizing map** (**SOM**) or **self-organizing feature map** (**SOFM**) is a type of an artificial neural network that produce a low-dimensional (typically two-dimensional) representation of the input space called a **map**.

Figure 49. Kehonen Feature Map structure.

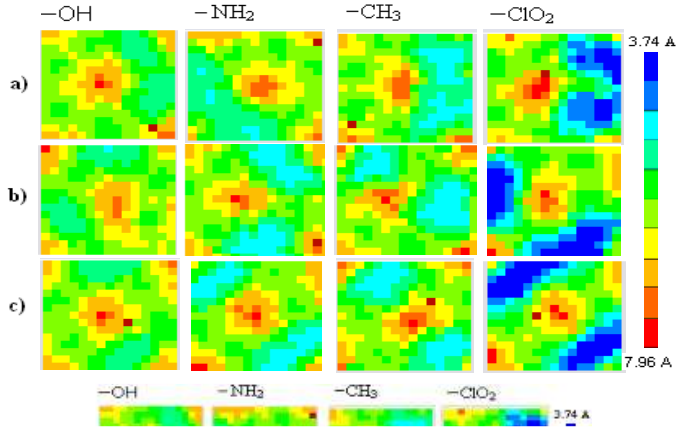


Kohonen maps are representation of the objects. As such they can be used for example for finding molecular similarities between objects represented in a form of Kohonen maps. For example a pair of bioisosteric groups or compounds of similar biological properties have very similar Kohonen maps. The compounds which fit into the same receptor will also be represented by the similar Kohonen maps of their molecular electrostatic potentials.

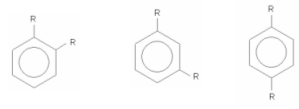
Figure 50. Example of 2D Kohonen transformation of chemical structures



transformation of 3D structure to 2D Kohonen map



2D Kohonen maps of two substituted benzene derivatives a) ortho, b)meta, c) para



Rysunek 9.1. Wzory strukturalne izomerów benzenu z podstawnikami -R w pozycjach orto, meta i para.

Some of the freely available software packages for NN simulation ?

- Rochester Connectionist Simulator
- UCLA-SFINX
- NeurDS
- PlaNet5.7
- GENESIS
- Mactivation
- Cascade Correlation Simulator
- Quickprop
- DartNet
- SNNS
- Aspirin/MIGRAINES
- Adaptive Logic Network kit
- NeuralShell
- PDP
- Xerion
- Neocognitron simulator
- Multi-Module Neural Computing Environment (MUME)
- LVQ_PAK, SOM_PAK
- SESAME
- Nevada Backpropagation (NevProp)
- Fuzzy ARTmap
- PYGMALION
- Basis of AI backprop

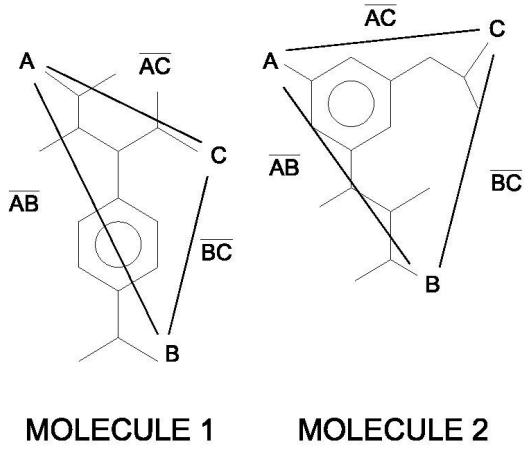
20 Active analogue approach

Molecular modeling covers a huge area of methods of modeling of a potential drug to a known receptor. Such an approach will not be covered in this study guide.
However there is mathematical approach that aims at getting an idea about the shape of the receptor in the case when the structure of the receptor is not known or even the receptor is not known. Such a method is named the active analogue approach.

The procedure of finding the shape of an unknown receptor is given below:

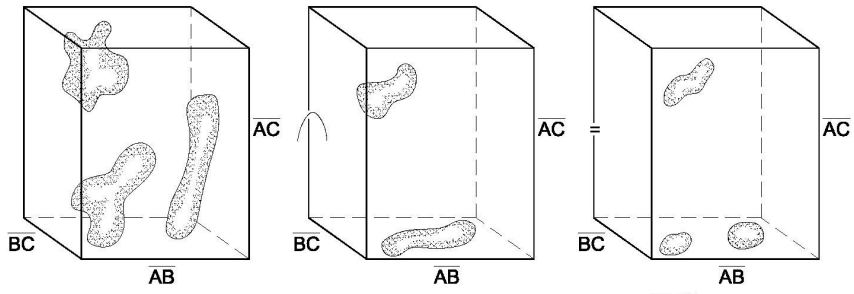
- Pharmacophoric group fragments are represented by points A,B,C
- Then the pharmacophoric fragment is represented by a triangle with the edges AB, BC, AC.

Figure 51. definition of a pharmacophoric pattern



- All possible conformation of a given compound is then represented by the areas in
- the three dimensional space AB, BC, AC as in the fig below – conformational pattern.

- The common volumes for all active compounds is are then represented by the three dimensional pharmacophoric pattern. The Figure below shows the procedure for two compounds.
- *Figure 52. Result of conformational analysis of two compounds*
-

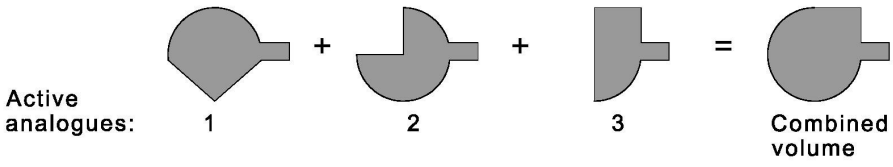


1 2 3

1. conformation pattern for molecule 1
2. conformation pattern for molecule 1
3. conformational pattern common for molecule 1 and 2

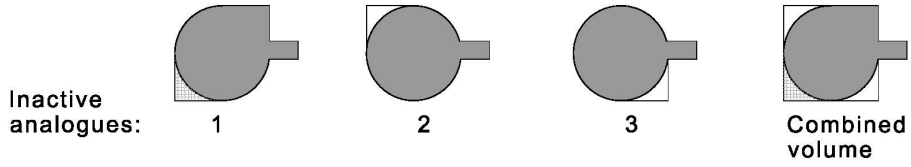
- If the set is empty, pharmacophoric groups were incorrectly assumed or
- the activity of the group of compounds is related to a different mechanism. If the set is not empty, we assume that we have found the pharmacophore.
- Superimposing the shapes/volumes of all active compounds in their active conformations represent the free volume in the active site.

Figure 53. Representation of pharmacophore



- The Combined volume of all inactive compounds in the conformation of the pharmacophoric pattern corresponding to the "active state" indicates the interaction of the compounds with the "body" of the active site.

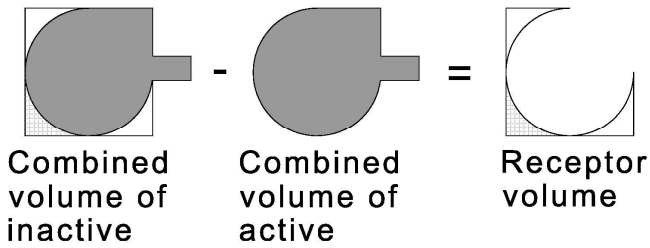
Figure 54. Representation of the enzyme pocket



- the fragment representing the interaction of the compound with the "body" of the receptor

- The analysis of the combined volumes of all the active and all the inactive yields the information about the active site volume as shown in the figure below.

Figure 55. Representation of the receptor volume



21. Literature

1.K.Varmuza, Pattern Recognition in Chemistry, Springer Verlag 1980

2.O. Exner, Correlation Analysis of Chemical Data, Plenum press, 1988

3.J.Zupan, J. Gastgeiger, Neural Networks in Chemistry and Drug design, Wiley-VCH, 1999

4.M.Bland, An Introduction to Medical Statistics, Oxford University Press, 2008

5.T.N.Toyer, M.Rowland, Introduction to Pharmacokinetics and Pharmacodynamics. The Quantitative Basis of Drug Therapy, Lippincot Williams and Wilkins , 2006
6.P.Krogsgaard Larsen, T. Liljefors, U.Madsen, Textbook of Drug Design and Discovery, Taylor and Francis, 2002