CHEMOMETRICS

Statistics and Multivariate Data Analysis in Chemistry

Kurt VARMUZA

Vienna University of Technology
Institute of Chemical Engineering
Laboratory for ChemoMetrics
Vienna, Austria

WWW.LCM.TUWIEN.AC.AT

kvarmuza@email.tuwien.ac.at

Poster Presentation
2003 Hawaii International Conference on Statistics and Related Fields
5 - 9 June 2003, Honolulu, Hawaii, USA

Introductory Remarks for Non Chemists (1)

1. **Common-live STUFF,**
   such as air, sea-water, food, plastics, fuel, ... ,
   is a very complicated mixture of many, many (chemical) COMPOUNDS (substances).

Typical:	5 - 20	main compounds
	100 - 1,000	minor compounds
	10,000 - ...	trace compounds

2. **Chemical analytical INSTRUMENTS* can**
   - **separate/extract main compounds,**
   - **separate/extract a few, especially interesting trace compounds (present in very low concentrations).**

   * For instance chromatographs.

3. **Separated (pure) compounds are usually characterized/identified by measuring SPECTRA [§ #] that are often characteristic for a compound.**

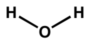
   § Spectroscopy:
   Energy is applied to the molecules, and the effect (absorption, resonance, chemical reaction products, ...) are measured.

   # Spectra types most often used in chemistry:
   IR (infrared spectra), MS (mass spectra), NMR (nuclear magnetic resonance spectra), UV (ultra-violet spectra).

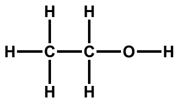
Introductory Remarks for Non Chemists (2)

4. A pure chemical COMPOUND consists of MOLECULES that define the compound.

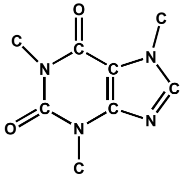
Examples of **molecular structures** (simplified, as colored graphs):



water



ethanol ("alcohol")



caffeine (H-atoms not shown)

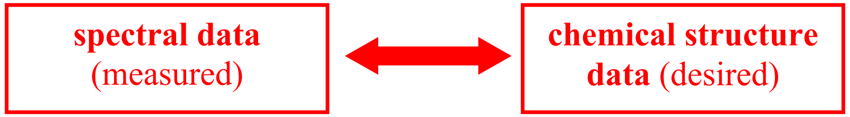
C, H, N, O are carbon-, hydrogen-, nitrogen-, oxygen-atoms, respectively.

5. IDENTIFICATION of a chemical compound means recognition/determination of its molecular structure.

6. Chemical structures cannot be measured/observed directly, but can only be inferred from

- spectral data,
- other chemical/physical/biological properties.

7. Available theory and experiences do not allow to establish generally applicable - and useful - relationships between



spectral data (measured)

chemical structure data (desired)

Introductory Remarks for Non Chemists (3)

8. DATABASES (spectra and chemical structures) with up to ca 200,000 entries (compounds) are used.

From (very) similar spectra is concluded, that the corresponding chemical structures are similar (or even identical).

9. For "well selected" subsets of chemical compounds mathematical MODELS can be developed, such as

chemical structure information = **f** (*spectral data*)
[identification of compounds]

property of compounds = **f** (*chemical structure data*)
[drug design, property prediction]

typically by applying multivariate data analysis or neural networks ("chemometrics").

10. CHEMOMETRICS

is an interfacial discipline between

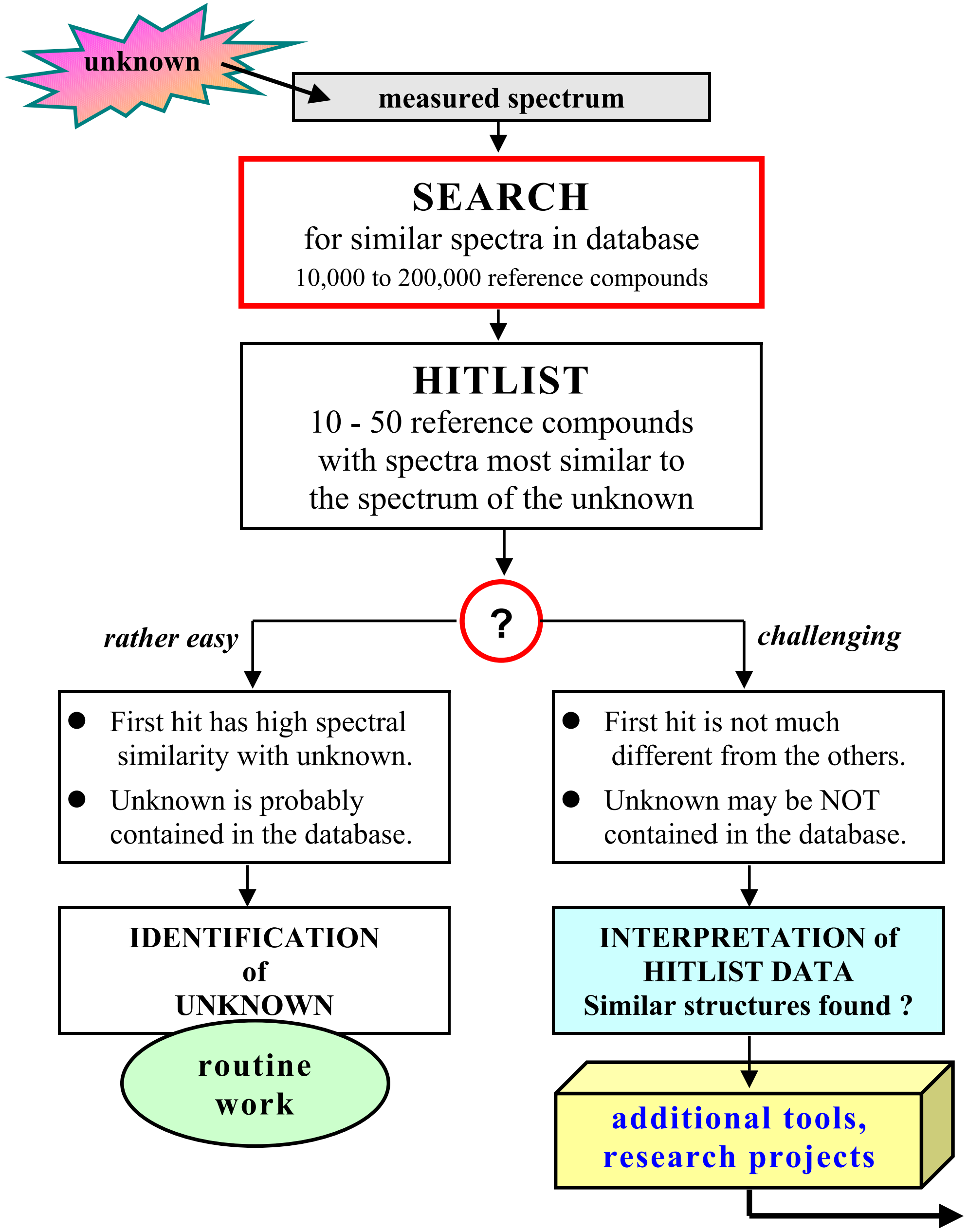
- **instrumental, measurement-oriented chemistry, and chemical technology,**

  and

- **applied statistics, and computer science.**

Extraction of information from chemistry-relevant data is essential.

Identification of a chemical compound



unknown

measured spectrum

SEARCH

for similar spectra in database

10,000 to 200,000 reference compounds

HITLIST

10 - 50 reference compounds with spectra most similar to the spectrum of the unknown

?

rather easy

- First hit has high spectral similarity with unknown.
- Unknown is probably contained in the database.

IDENTIFICATION of UNKNOWN

routine work

challenging

- First hit is not much different from the others.
- Unknown may be NOT contained in the database.

INTERPRETATION of HITLIST DATA

Similar structures found ?

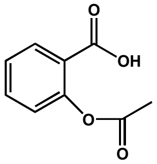
additional tools, research projects

Spectra as Vectors / Similarity of Spectra

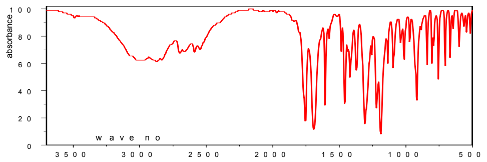
Salicylic acid acetate (aspirin)

$C_9H_8O_4$

molecular weight 180



IR spectrum



Spectra can be easily represented by vectors. Sometimes mathematical transformations (for instance autocorrelation) are applied, as well as transformations guided by spectroscopic experiences.

Number of vector elements: 200 - 1,000 (depending on resolution)

Similarity/diversity of spectra

Most often based on

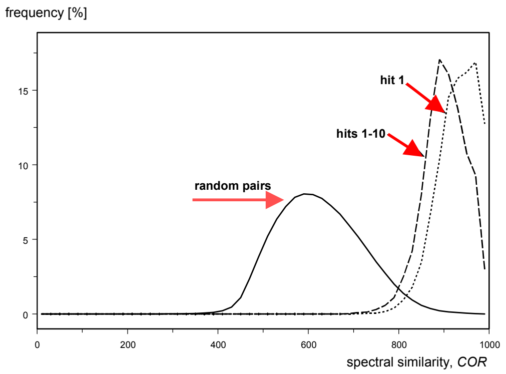
x_A, x_B are vectors representing spectrum A and B, respectively

Correlation coefficient $= (x_A^T . x_B) / (\| x_A \| * \| x_B \|)$

or

Euclidean distance $= \| (x_A - x_B) \|$

sometimes extended by spectroscopic ideas.



Frequency distribution of spectral similarity for 10^6 pairs of randomly selected different structures from a database with 13484 compounds.

hit 1, 1-10, frequency distributions of most similar and 10 most similar spectra (each of the 13484 spectra was used as query spectrum).

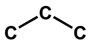
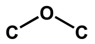
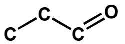
Frequency is given in % for 50 intervals, each 20 units wide.

Structures as Vectors / Similarity of Structures

Several methods have been developed for the representation of chemical structures by vectors. Only one approach is mentioned here:

Representation of a chemical structure by a binary vector, with each binary vector element being a molecular descriptor, that indicates presence/absence of a predefined substructure.

Demo example (subgraph isomorphism)

encoded structure	predefined substructures		
			$C=O$
	1	0	1

Actual number of vector elements (substructures): 200 - 2,000.

y

Similarity/diversity of chemical structures

Widely used is the

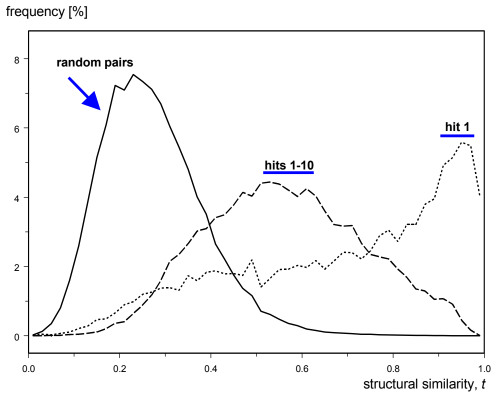
y_A, y_B are binary vectors representing structure A and B, respectively

Tanimoto index (Jaccard similarity)

$$t = (y_A^T . y_B) / (y_A^T .1 + y_B^T .1 - y_A^T . y_B)$$

$$= \Sigma \text{ AND}[y_A(j), y_B(j)] / \Sigma \text{ OR}[y_A(j), y_B(j)]$$

as a similarity measure of two chemical structures (range of *t* is 0 ... 1).

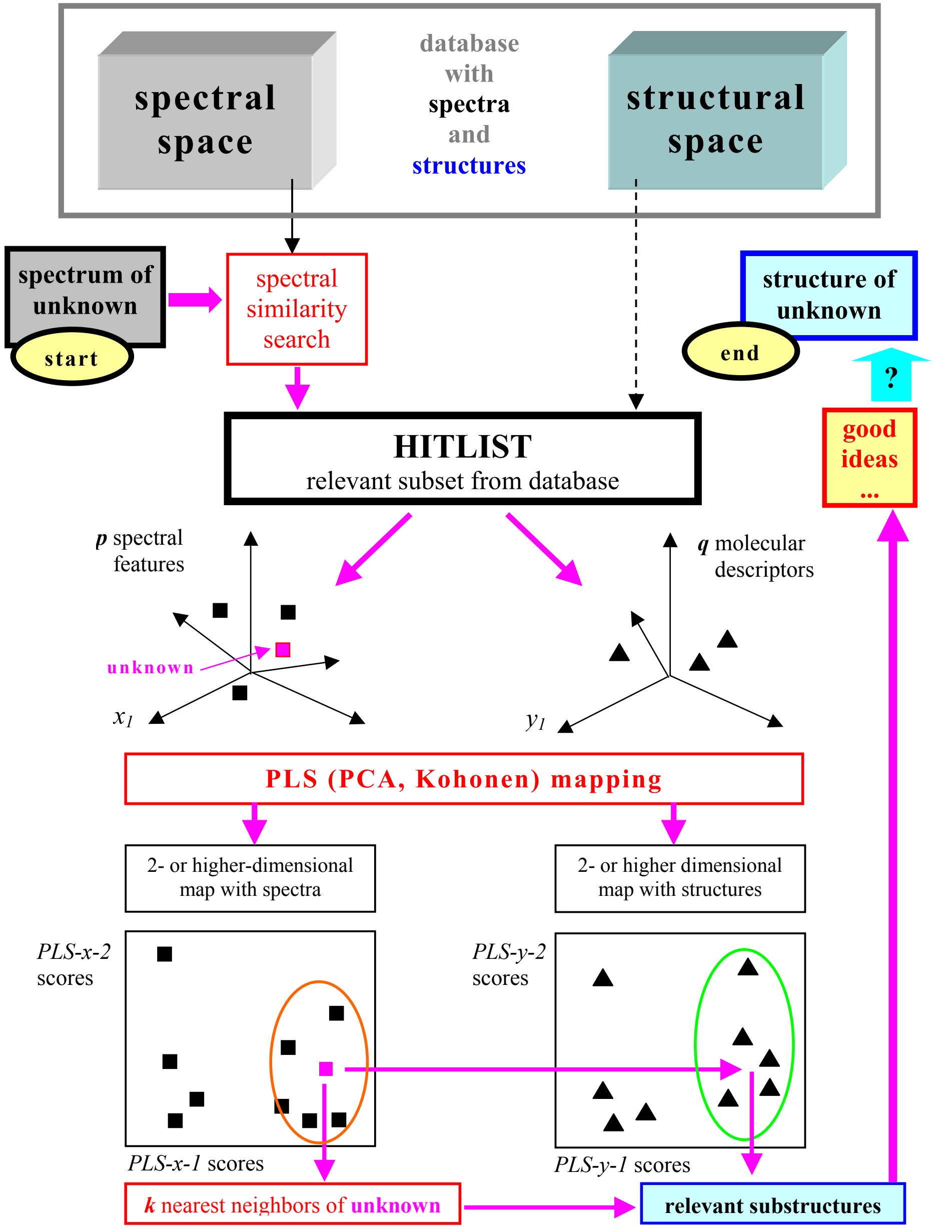


Frequency distribution of structural similarity for 10^6 pairs of randomly selected different structures from a database with 13484 compounds.

hit 1, 1-10, frequency distributions for compounds with most similar and 10 most similar spectra (each of the 13484 spectra was used as query spectrum).

Frequency is given in % for 50 intervals, each 0.02 units wide.

Exploration of hitlist data



spectral space

database with spectra and structures

structural space

spectrum of unknown

start

spectral similarity search

structure of unknown

end

?

good ideas ...

HITLIST
relevant subset from database

PLS (PCA, Kohonen) mapping

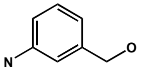
2- or higher-dimensional map with spectra

2- or higher dimensional map with structures

k nearest neighbors of unknown

relevant substructures

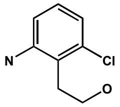
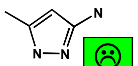
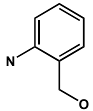
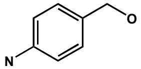
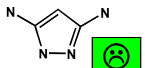
Example: IR spectrum similarity search (1)



Query compound	**3-amino-benzylalcohol**
Database	13484 compounds (IR spectra and structures, SpecInfo)
Spectral similarity	correlation coefficient of absorbance units
Structural similarity	Tanimoto index (*t*) based on 1365 substructures

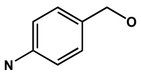
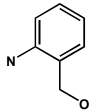
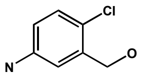
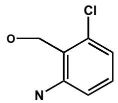
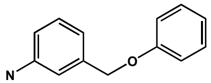
(A) Most similar spectra in database

Tanimoto mean 1 - 5: **0.66**

	 ☹			 ☹
t = 0.74	0.48	0.96	0.96	0.14

(B) Most similar structures in database

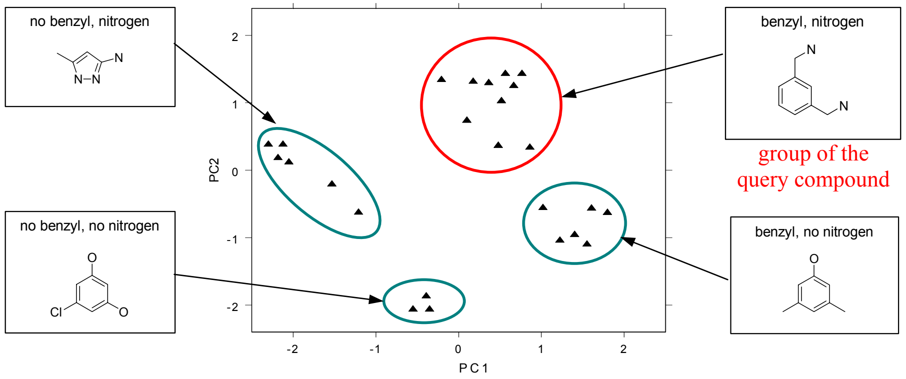
Tanimoto mean 1 - 5: **0.91**

				
t = 0.96	0.96	0.89	0.89	0.85

two best reference structures in yellow

(C) Cluster analysis of hitlist structures by PCA

18 binary substructure descriptors; variance retained in PC1, PC2: 36%, 28%



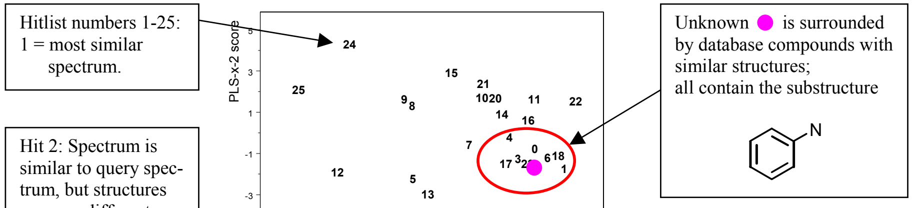
group of the query compound

Four groups of chemical structures found (example structures shown)

Example: IR spectrum similarity search (2)

(D) PLS mapping of spectra (*X*) and structures (*Y*)

X: averaged absorbances (autoscaled) of 50 wavenumber intervals between 500 and 3700 cm^{-1}
Y: 18 binary substructure descriptors (autoscaled)
PLS-*x* components are defined by the first two eigenvectors of $\boldsymbol{X^T Y Y^T X}$



Hitlist numbers 1-25: 1 = most similar spectrum.

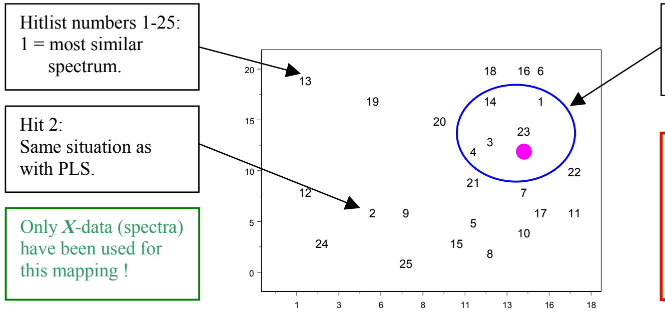
Hit 2: Spectrum is similar to query spectrum, but structures are very different.

Unknown ● is surrounded by database compounds with similar structures; all contain the substructure

Distances in PLS-component space can be used for a new ranking of the hitlist compounds.

(E) Kohonen mapping of spectra (*X*)

X: averaged absorbances of 50 wavenumber intervals between 500 and 3700 cm^{-1}
Software SOMPAK (Helsinki University of Technology), map size 20*20



Hitlist numbers 1-25: 1 = most similar spectrum.

Hit 2: Same situation as with PLS.

Only *X*-data (spectra) have been used for this mapping !

Unknown ● is surrounded by database compounds with similar structures.

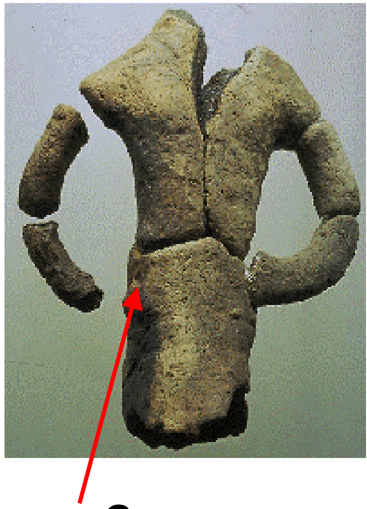
Distances in the Kohonen map can be used for a new ranking of the hitlist compounds. However, the result heavily depends on the parameters used.

PCA and PLS support the evaluation of hitlists

- by cluster analysis of chemical structures
- by selection of most relevant database structures

Archaeology - Chemometrics

A terracotta statuette was found in a prehistoric settlement near Vienna (Austria).



^{14}C dating: 5650 - 5100 B.C.

Seven fragments	
Preserved size	**14.2 cm**
Reconstructed size	**25 cm**

Prehistoric function

Maybe an idol (religious object)!
Maybe just a toy puppet?

Finding date 1989

Grooves were filled with an unknown dark material - obviously of organic origin.

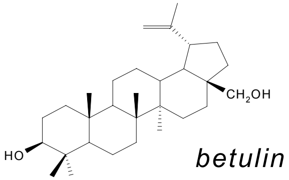
First examinations of the dark material and experiences with similar material found on other archaeological findings - for instance the Neolithic *Tyrolean Iceman* - lead to the idea

The dark material might be pitch produced by pyrolysis of birch wood.

Aim of the work was to evaluate this idea by chemotaxonomy + chemometrics.

Methods / Data / Conclusions

Compounds



Wood pitches can be characterized by **concentration patterns of triterpenoids**, such as *betulin* (characteristic for birch trees), or *friedelin* (characteristic for cork oak trees).

Samples

Reference samples were prepared by pyrolysis of wood and/or bark taken from four species of trees of the family *Betulaceae*.

Chemical Analysis

Analysis of pitch samples included several steps:

Distillation, solid phase extraction, gas chromatography / mass spectrometry, identification of main compounds by spectral similarity search, selection of 50 compounds for multivariate data analysis.

Data

33 objects	14	from *Betula*	(birch)	class 1	tribe *Betuleae*
(samples)	6	from *Alnus*	(alder)	class 2	
	7	from *Corylus*	(hazelnut)	class 3	tribe *Coryleae*
	5	from *Carpinus*	(hornbeam)	class 4	
	1	archaeological sample (unknown)			

50 features (relative concentrations, autoscaled)

Conclusions

The applied data analysis methods strongly indicate:

The dark material from the Neolithic statuette was prepared from wood or bark of birch trees (*Betula*).

This conclusion is consistent with other finds in prehistoric Europe.

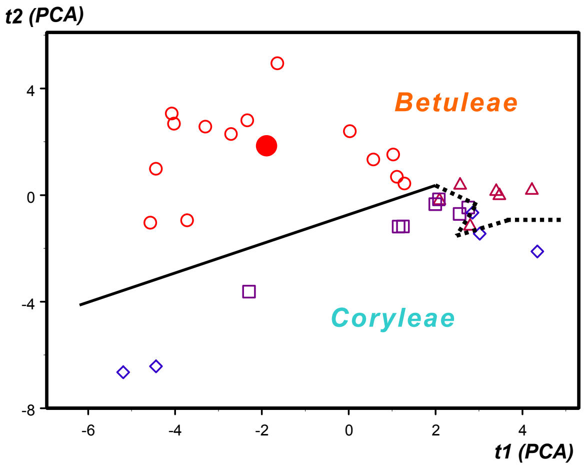
Pitch made from birch trees has been used as a multifunctional material (as coating of pottery, as glue, even as a gift).

The investigated pitch from the statuette may have been used to fix some textile dressing.

Sauter F., Varmuza K., Werther W., Stadler P.: ARKIVOC **2002** [1] 54-60 (2002)
Free copy: http://www.arkat-usa.org/ark/journal/2002/General/1-343E/343E.pdf

PCA and HCA

PCA Principal Component Analysis Mapping

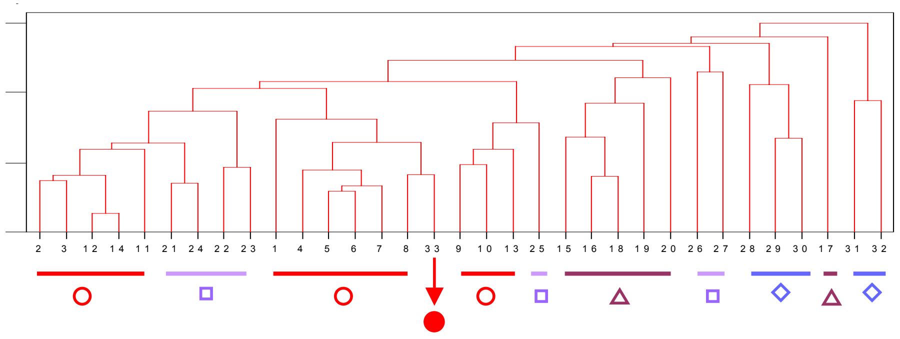


t1, *t2* are scores of PC1 and PC2, respectively, with variances 18.4 and 11.7% of total variance.

Separation of the tribes **Betuleae** (including *Betula* ○ and *Alnus* △), and **Coryleae** (including *Corylus* □ and *Carpinus* ◇).

● unknown material

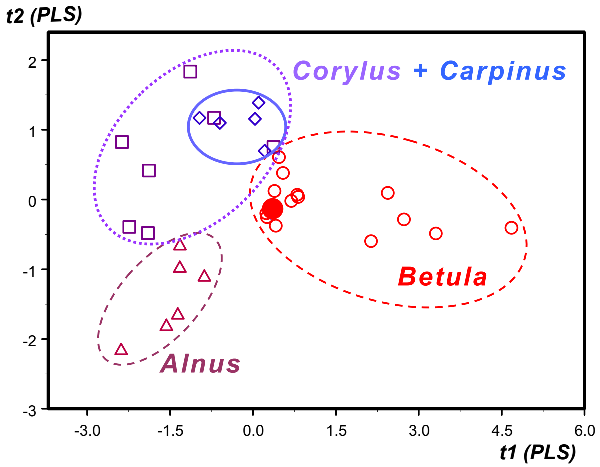
HCA Hierarchical Cluster Analysis



Similar clustering as obtained with PCA; however, less evident.

PLS and LDA

PLS Partial Least Squares Discriminant Mapping



t1, *t2* are *x*-scores of 1st and 2nd PLS component.

X: 50 concentrations

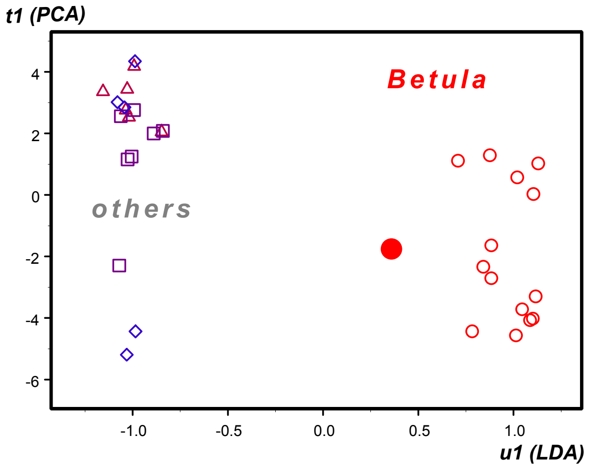
Y: 4 binary class variables

PLS: eigenvectors of

$$X^T Y Y^T X$$

Genus *Betula* is separated from genus A*lnus* and the overlapping genera *Corylus* and *Carpinus*.

LDA Linear Discriminant Analysis Mapping



u1 is the LDA discriminant variable, calculated from the first 25 PCA scores (for discrimination of genus *Betula* from the other classes).

t1 is the score of PC1.

Genus *Betula* is well separated from the other classes.

The unknown ● can be assigned to class **Betula**.

Notes and References

Coworkers

W. Demuth, M. Karlovits, F. Müller, S. Qehaja, H. Scsibrany

Collaborations and Acknowledgments

A. Kerber, R. Laue (Univ. Bayreuth, Germany; graph theory); **R. Neudert** Chemical Concepts / VCH-Wiley, Weinheim, Germany; spectra databases); **J. Zupan** (Institute of Chemistry, Ljubljana, Slovenia; chemometrics); **Z. Hippe** (Univ. Rzeszow, Poland: computer chemistry); **P.N. Penchev** (Univ. Plovdiv, Bulgaria: chemometrics); **Z. Garkani-Nejad** (Univ. Rafsanjan, Iran; QSPR); **H. Masui, K. Funatsu** (Sumitomo Osaka, and Univ. Techn. Toyohashi, Japan; computer chemistry and spectra databases); **K.T. Fang, P. He** (Univ. Hongkong, China; statistics).

References

Beebe K.R., Pell R.J., Seasholtz M.B.: *Chemometrics: A practical guide*, Wiley, NewYork (1998).

Kramer R.: *Chemometric techniques for quantitative analysis*, Marcel Dekker, NewYork (1998).

Massart D.L., Vandeginste B. G. M., Buydens L. C. M., De Jong S., Smeyers-Verbeke J.: *Handbook of chemometrics and qualimetrics: Part A*, Elsevier, Amsterdam (1997).

Scsibrany H., Karlovits M., Demuth W., Müller F., Varmuza K.: *Chemom. Intell. Lab. Syst.*, in print (2003). Clustering and similarity of chemical structures represented by binary substructure descriptors.

Vandeginste B. G. M., Massart D. L., Buydens L. C. M., De Jong S., Smeyers-Verbeke J.: *Handbook of chemometrics and qualimetrics: Part B*, Elsevier, Amsterdam (1998**)**.

Varmuza K.: In *The Encyclopedia of Computational Chemistry*; Schleyer P. v. R., Allinger N. L., Clark T., Gasteiger J., Kollman P. A., Schaefer III H. F., Schreiner P. R., Eds.; Wiley & Sons: Chichester, Vol. 1, p. 346-366 (1998). Chemometrics: Multivariate view on chemical problems.

Varmuza K.: In *Encyclopedia of Spectroscopy and Spectrometry*; Lindon J. C., Tranter G. E., Holmes J. L., Eds.; Academic Press: London, p. 232-243 (2000). Chemical structure information from mass spectrometry.

Werther W., Demuth W., Krueger F. R., Kissel J., Schmid E. R., Varmuza K.: *J. Chemometrics*, 16, 99-110 (2002). Evaluation of mass spectra from organic compounds assumed to be present in cometary grains. Exploratory data analysis.

Makristathis A., Schwarzmeier J., Mader R.M., Varmuza K., Simonitsch I. Chavez J.C., Platzer W., Unterdorfer H., Scheithauer R., Derevianko A., Seidler H.: *J. Lipid Res.,* 43, 2056-2061 (2002). Fatty acid composition and preservation of the Tyrolean Iceman and other mummies.

Scsibrany H., Karlovits M., Demuth W., Müller F., Varmuza K.: *Chemom. Intell. Lab. Syst.*, in print (2003). Clustering and similarity of chemical structures represented by binary substructure descriptors.

Text from Proceedings

Chemists can often measure the properties of compounds or processes not directly. Examples of such problems are: identification of compounds, recognition of the chemical structure; quantitative analyses of complex mixtures, determination of the origin of samples, and prediction of properties or activities of chemical compounds or technological materials. In these cases a single variable is insufficient to model the desired data or to provide the required information. Therefore a multivariate approach is necessary for many problems in chemistry. The three main application areas are exploratory data analysis, classification, and calibration.

The identification of chemical compounds is for instance an essential task in the characterization of materials from environmental chemistry, food chemistry, biology, medicine, and technology. Identification of a chemical compound is equivalent to the recognition or determination of its chemical structure. However, chemical structures cannot be measured or identified directly, but only by the evaluation of appropriate experimental data - usually of spectral data. A spectrum (for instance infrared spectrum, mass spectrum, nuclear magnetic resonance spectrum) can be represented by a vector, and such a vector is more or less characteristic for a chemical structure. Also a chemical structure can be characterized by a vector. Unfortunately, chemistry does not provide sufficient theory for the relationships between spectral data and chemical structure data.

Therefore, databases are widely used that contain chemical structure data and corresponding spectra, measured on reference compounds. Identification of an unknown compound is usually performed by automatic searches for reference spectra that are most similar to the measured spectrum. The similarity of vectors (representing spectra or chemical structures) is for instance defined by the Euclidean distance, the correlation coefficient or the Tanimoto index. The size of available spectral databases is some ten thousands to some hundred thousands of compounds; these numbers are much smaller than the number of known chemical compounds which is several millions. Therefore, additional strategies for chemical structure elucidation have been investigated in chemistry that are based on multivariate modeling, statistics, and computer science. Works in this field belong to a discipline in chemistry called chemometrics.

Chemometrics is considered as a chemical discipline, that uses statistical and mathematical methods, to design or select optimal procedures and experiments, and to provide maximum chemical information by analyzing chemical data. Today's chemometrics is dominated by applications of multivariate data analysis to chemistry-relevant problems.

Typical chemometric applications use the methods of principal component analysis, partial least squares regression and other concepts from multivariate data analysis and statistics - for instance to model relationships between spectral and structural data. Scatter plots - resulting from these methods, with a point for a chemical structure or for a spectrum - are helpful in the interpretation of measured spectra originating from compounds not present in available databases. Multivariate classification methods are

helpful for the prediction of parts of the unknown molecular structure.

Automatic searches for similar spectra is the most popular approach in computer-assisted evaluation of spectra. The resulting hitlist contains reference spectra (from the spectral database) that are most similar to the query spectrum; the hitlist is usually ordered by decreasing spectral similarity. If spectral data from the unknown compound are contained in the spectral library the correct solution is often given by the first hit or is among the first hits. If the unknown is not contained in the library, the hitlist data may be exploited with the aim to gather chemical structure information about the unknown.

The interpretative power of a spectral similarity search system is the ability to produce hitlists with chemical structures that are very similar to the structure of query compounds. For a systematic evaluation of library search systems it is necessary to define similarity criteria for spectra as well as for chemical structures.

Chemometrics can be considered as an interfacial discipline between measurement-oriented chemistry and applied statistics; it concerns the extraction of information from chemical data by mathematical and statistical tools. Chemometrics mainly focuses on the chemical model, rather than on random effects or distributions. The basic hypothesis suggests that complicated chemical systems can be characterized by a set of measured variables and that models (so called latent variables like for instance principal component scores) can help to find the essential information. Selection or creation of appropriate problem-relevant features is often more important than the method which is then applied for data interpretation. Actually, many parts of chemistry can be seen as indirect studies of latent concepts and therefore chemometric methods have been applied to a huge number of problems. Many applications, but by far not all, belong to analytical chemistry.

An important branch of chemometrics is pattern recognition with the aim of classifying unknowns to a class out of a set of pre-determined classes. A great variety of different types of samples or materials has been investigated, such as food samples, biological and medical samples, technological materials, environmental and archaeological samples. A typical goal of data analysis is to obtain information about the origin or quality of samples. The crucial point is the characterization of the objects by selecting problem-relevant measurements, such as for instance concentrations of elements or compounds or spectroscopic data.

Multivariate calibration (mainly based on principal component analysis, partial least squares regression, and artificial neural networks) has the largest number of applications of chemometric methods in routine work; for instance it became a widely used technique in quantitative analysis of complex mixtures. Typical examples are the determination of fat in meat or of water in protein by fast and cheap spectroscopic methods (instead of time- and chemicals-consuming wet-chemistry experiments). An important field is the investigation of quantitative chemical structure - activity relationships (QSAR); that means the search for mathematical model that are able to predict physical or biological properties of chemical compounds by using only chemical structure data (drug design).

A number of chemometric methods and software products are now routinely used in chemical laboratories. However, especially in the field of chemical structure recognition, many problems are unsolved. They require a deeper understanding how to model relationships between sets of multivariate data and what are appropriate statistical concepts for chemistry-relevant problems.