Statistical Evaluation of Molecular Descriptors

Used in Quantitative-Structure-Activity Relationships

Kurt Varmuza[1*], Peter Filzmoser[2], Matthias Dehmer[3]

[1] **Vienna University of Technology,**
Institute of Chemical Engineering, Laboratory for Chemometrics
Getreidemarkt 9/166, A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at, www.lcm.tuwien.ac.at

[2] **Vienna University of Technology,**
Institute of Statistics and Probability Theory
Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria
P.Filzmoser@tuwien.ac.at, www.statistik.tuwien.ac.at/public/filz

[3] **UMIT, University for Medicine and Information Technology,**
Institute for Bioinformatics and Translational Research
Eduard-Wallnöfer Zentrum 1, A-6060 Hall in Tyrol, Austria
matthias.dehmer@umit.at, www.umit.at

* Presenting author

Poster Presentation
First African-European Conference on Chemometrics
September 21–24, 2010, Rabat, Morocco

Acknowledgments. We gratefully acknowledge support by the Austrian Science Fund (FWF), project P22029-N13. We thank Katja Hansen (TU Berlin, Germany) for the AMES data, and Bettina Liebmann (TU Vienna) for collaboration.

Introduction

Molecular descriptors [1] are used to characterize chemical structures by numerical vectors. Successful applications in chemoinformatics [2] are

- empirical models for QSA(P)R, quantitative structure-activity- (property-) relationships;
- chemical structure searches, e.g. for similarity.

Several thousand molecular descriptors have been described, ranging from topological descriptors (the chemical structure is represented by a graph) to simulated physical/chemical/ biological properties. A crucial problem is to examine whether a "new group" of descriptors provides "new information" or not.

In this preliminary study, we investigate correlations between 1604 molecular descriptors and groups of descriptors. The descriptors are calculated for structures from a selected group of 6458 organic chemical compounds, with about half of them being mutagenic according to AMES tests.

Correlations between groups of variables are investigated by

Redundancy Analysis (RA)

a rather new method in chemometrics.

[1] Todeschini R., Consonni V., Mannhold R.: Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany (2002)
[2] Gasteiger J. (ed.): Handbook of Chemoinformatics - From Data to Knowledge (4 vol.), Wiley-VCH, Weinheim, Germany (2003)

Data and Software

Data

- **Objects (chemical structures)**

n = 6458 organic chemical compounds, with known AMES test results (mutagenic or not), collected by Katja Hansen, Berlin [3].

Data format of chemical structures: Molfile (2D) [2]

n_1 = 3488 (54 %) mutagenic compounds,
n_2 = 2970 (46 %) not mutagenic compounds.

Approximate 3-dimensional (3D) structures including all H-atoms have been created by software ***Corina*** [4].

- **Variables (molecular descriptors)**

m = 1604 molecular descriptors have been calculated by software ***Dragon*** [5].

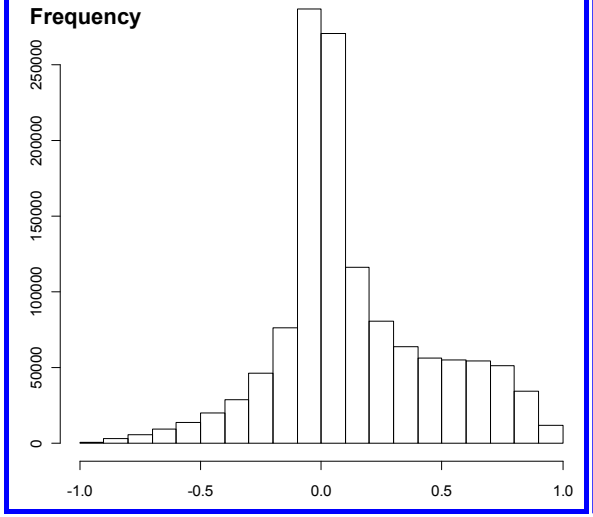
The descriptors are divided into **19 groups** according to definition:
(1) constitutional, (2) topological, (3) walk and path counts, (4) connectivity, (5) information indices, (6) 2D autocorrelations, (7) edge adjacency, (8) Burden eigenvalues, (9) topological charge, (10) eigenvalue-based, (11) Randic molecular profiles, (12) geometrical, (13) RDF, (14) 3D-MoRSE, (15) WHIM, (16) GETAWAY, (17) functional groups, (18) atom-centered fragments, (19) molecular properties.

Software

All statistical computations were performed with programs written in ***R*** [6]. The free R-library "chemometrics" [7] has been used.

[3] Hansen K.: http://ml.cs.tu-berlin.de/toxbenchmark/index.html (TU Berlin)
[4] Corina software, Molecular Networks GmbH Computerchemie, www.mol-net.de, Erlangen, Germany (2004)
[5] Dragon software, 5.0, Talete srl, www.talete.mi.it, Milan, Italy (2004)
[6] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, 2010. www.r-project.org
[7] Varmuza K., Filzmoser P.. Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, FL (2009) Info: www.lcm.tuwien.ac.at/cm-book.pdf

Correlations of Variable Pairs, and with y

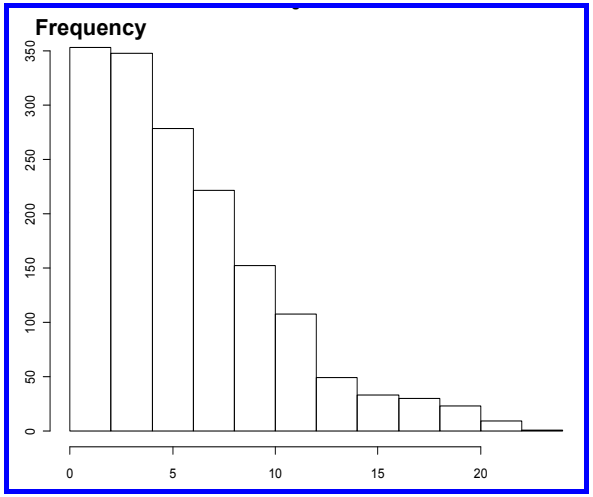


Limit of |R|	#(%) of variable pairs	
1	12	0.001%
0.999	205	0.016%
0.99	953	0.074%
0.98	1739	0.235%
0.95	4332	0.337%
0.9	12717	0.989%

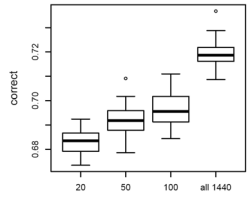
Total: 1 285 606 variable pairs.

Only ca 1% have $|R| \geq 0.9$

Pearson correlation coefficient, *R*, between descriptors



***t*-value from paired t-test**



Distribution of predictive ability for classification of mutagenic compounds (others similar).

Repeated double cross validation [8] with 30 repetitions (test set objects).
20, 50, 100 variables with maximum *t*-value, and all variables were used.

Univariate variable selection not recommended !

[8] Filzmoser P., Liebmann B., Varmuza K.: J. Chemometrics, **23**, 160-171 (2009)

Correlations between Variable Groups

Redundancy Analysis (RA)

The linear relationship between two groups of variables $x_1, ..., x_p$ ["independent variables", matrix $\boldsymbol{X}(n \times p)$] and $y_1, ..., y_q$ ["dependent variables", matrix $\boldsymbol{Y}(n \times q)$] is characterized by the **redundancy indices** $R_1, R_2, ..., R_\rho$ [9] with $\rho = \min(\text{rank}(\boldsymbol{X}), q)$. R_i (for $i = 1, ..., \rho$) is defined as

$$R_i = \sum_{j=1}^{q} [\text{ corr } (\boldsymbol{u}_i, \boldsymbol{y}_j)\ ]^2 / q \qquad \text{[range 0 ... 1]}$$

$\boldsymbol{u}_i = \boldsymbol{X}\,\boldsymbol{b}_i$ scores of a latent variable defined by loading vector $\boldsymbol{b}_i$ that maximizes R_i

A maximum of ρ redundancy indices can be derived under the conditions (1) no correlation between any score vectors $\boldsymbol{u}_i$ and (2) unit variances of the scores. For a robust RA see [10].

Vectors $\boldsymbol{b}_1, ..., \boldsymbol{b}_\rho$ are the eigenvectors of the matrix product $\boldsymbol{R}_x^{-1}\,\boldsymbol{R}_{xy}\,\boldsymbol{R}_{yx}$ with

- $\boldsymbol{R}_x$ matrix with Pearson correlation coefficients of $\boldsymbol{X}$, $(p \times p)$, (for the inversion a PCA-like transformation is performed);
- $\boldsymbol{R}_{xy}$ matrix with Pearson correlation coefficients between $\boldsymbol{X}$ and $\boldsymbol{Y}$, $(p \times q)$;
- $\boldsymbol{R}_{yx}$ matrix with Pearson correlation coefficients between $\boldsymbol{Y}$ and $\boldsymbol{X}$, $(q \times p)$.

The **used measure** for the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is

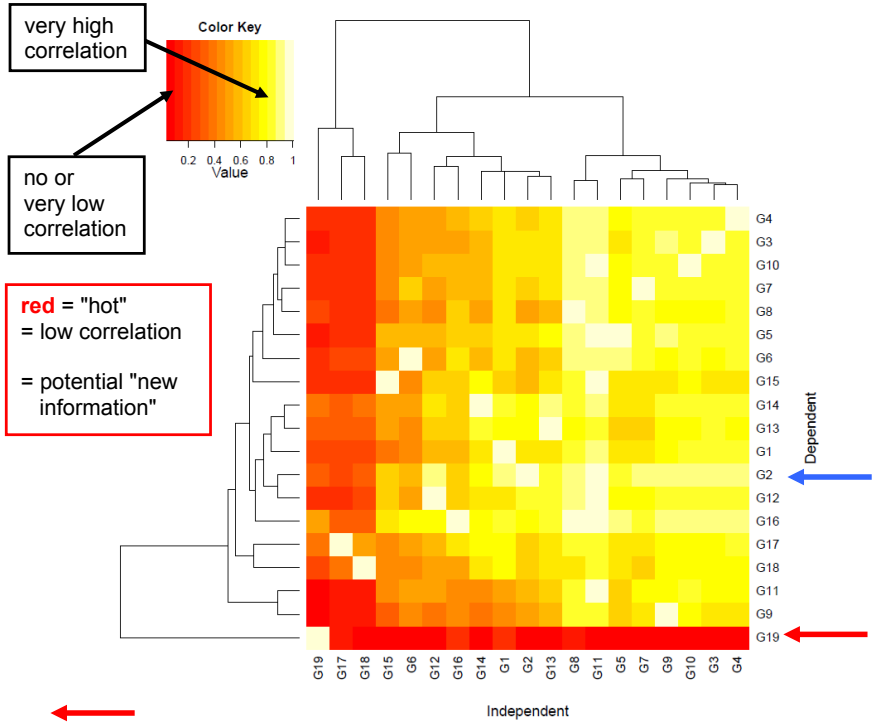
$$R_{SUM} = R_1 + R_2 + ... + R_\rho \qquad \text{[range 0 ... 1]}$$

Canonical correlation analysis is usually not suited for estimating relationships between groups of variables because even one pair of highly correlating variables x_k and y_l results in a high correlation measure for the variable sets.

[9] van den Wollenberg A.L.: Psychometrika **42**, 207-219 (1977)
[10] Oliveira M.R., Branco J.A., Croux C., Filzmoser P.: Statistics for Industry and Technology (ed.: Balakrishnan N.), p. 235-246, Birkhäuser Verlag, Basel, Switzerland (2004)

RA Results

The correlations between the 19 variable (descriptor) groups G1 ... G19 have been measured by the summed redundancy coefficient, R_{SUM} and displayed as a "heat map". The dendrograms indicate similar variable groups. Note the asymmetry because one variable group are independent variables, and the other dependent variables.



- Group G19 (*computed molecular properties*) shows very low correlations to all other groups.
- Group G2 (*topological descriptors*) and several other groups show high correlations to many other groups.

Application of RA for the evaluation of "new" descriptors, and for variable selection are future aims.