STATISTICAL APPROACH OF STRUCTURE-ACTIVITY RELATIONSHIPS: A CASE STUDY

Sorana D. BOLBOACĂ and Lorentz JÄNTSCHI

Structure/Activity/Property Relationships (SARs, SPRs, and PARs) appears with the studies of Louis Plack HAMMETT in 1937 [1]. The most important applications of Hammett's equation were summarized in [2].Quantitative relationships (QSAR, QSPR, QPAR) occur when the property/activity is quantitative. Not all properties and activities of chemical compounds can be classified as quantitative. In fact, few properties meet all theoretical requirements to be quantitative [3]. From this reason in the last time are avoided to be used QSAR, QSPR, and QPAR, in their place being used (Q)SAR, (Q)SPR, and (Q)PAR, or more simple SAR, SPR, and PAR. Structure-based approaches have two levels (topological and geometrical). In the topological based level, an atom, a bond from a molecule can exist (and then are evidenced through electronic transitions and/or molecular vibrations and/or rotations) or not (being a matter of 0 and 1). Not so simple stays things related to molecular geometry (especially on liquid or gas phases). Heisenberg uncertainly principle [4] shows the uncertainly rules presented at micro level (molecular and atomic level). More than that, molecular geometry depends on the environment where the molecule is (vicinity of the molecule), temperature, pressure, so on, thus dealing with molecular geometry is both a matter of relativity and a matter of uncertainty. Thus, Structure-Property-Activity Relationships (SPARs) must deal with certainties (such as molecular topology), uncertainties (such as molecular geometry), relativities (such as biological activities) and evidences (such as physical and chemical quantitative properties). The Molecular Descriptors Family (MDF) is an original structure-based approach [5] which generates for given structure(s) a huge pool of quantum based [6] descriptors of structure (indices) using a unitary methodology [7] that incorporated both topological and geometrical approaches. SPARs MDF methodology [8] uses a genetic algorithm [9] in order to obtain so called MDF-SPARs (structure-property or structure-activity relationships with Molecular Descriptors Family members relating the structure). **AIM: to assess the potential of MDF-SPARs for drug design. IDEA: to develop, test, and use a complete statistical methodology in the evaluation of obtained relationships, to estimate and predict the desired activity/property. METHODS:** A (Q)SAR/(Q)SPR equation is often a Multiple Linear Regression (MLR) equation. Key statistics are given in table below.

Parameter	Mathematical formula	Remarks	General issues
Simple correlation analysis	$r_{SP}=r(Y,MDF_i)$, $p_{SP}=p(r_{SP},m,df=1)$	r_{SP}: correlation between Y and MDF_i; p_{SP}: probability of no linear dependence between Y and MDF_i; a larger p_{SP} (usually > 5%) leads to excluding of MDF_i from MLR equation	MLR: $\hat{Y} = \Sigma_i a_i MDF_i$; a_i: real coefficients (MLR coefficients); $\hat{Y}$: estimator of the measured activity/property Y; MDF_i: an MDF member (an array with m values); m: sample size; n: number of variables; i: vary from 1 to n; r: Pearson correlation coefficient; p: probability of wrong model (using either Fisher or Student distribution); df: degrees of freedom; $i \neq j$ ($i < j$ is enough);
Inter-correlation analysis	$r_{IP}=r(MDF_i,MDF_j)$, $p_{IP}=p(r_{IP},m,df=1)$	r_{IP}: correlation between MDF_i and MDF_j; p_{IP}: probability of no linear dependence between MDF_i and MDF_j; a larger r_{IP} (usually larger than r_{MP}) leads to a less predictive MLR equations; a solution can be excluding of MDF_i (if $r_{SP}(Y,MDF_i) < r_{SP}(Y,MDF_j)$ is true) or MDF_j (if $r_{SP}(Y,MDF_i) < r_{SP}(Y,MDF_j)$ is false) from MLR equation; same procedure can be applied for $p_{MP}>p_{IP}$	
Multiple correlation analysis	$r_{MP}=r(Y,\hat{Y})$, $p_{MP}=p(r_{MP},m,df=n)$	r_{MP}: Pearson multiple correlation coefficient; p_{SP}: probability of no linear dependence between Y and $\hat{Y}$; a larger p_{MP} (usually > 5%) leads to rejecting of MLR equation	
Qualitative vs. quantitative analysis	$r_{MS}=\rho(Y,\hat{Y})$, $p_{MS}=p(r_{MS},m,df=n)$ $r_{M\tau a}=\tau_a(Y,\hat{Y})$, $p_{M\tau a}=p_Z(r_{M\tau a},m,df=n)$ $r_{M\tau b}=\tau_b(Y,\hat{Y})$, $p_{M\tau b}=p_Z(r_{M\tau b},m,df=n)$ $r_{M\tau c}=\tau_c(Y,\hat{Y})$, $p_{M\tau c}=p_Z(r_{M\tau c},m,df=n)$ $r_{MГ}=Г(Y,\hat{Y})$, $p_{MГ}=p_Z(r_{MГ},m,df=n)$ $r_{MSP}=\sqrt{r_{MS}\cdot r_{MP}}$	r_{MX}: multiple qualitative correlation coefficients (X = S, τa, τb, τc, Γ); p_{MX}: probability of no linear dependence between ranks of Y and $\hat{Y}$; a larger p_{MX} (usually more than 5%) leads to rejecting of MLR equation r_{MSP}: multiple semi-quantitative correlation coefficient; p_{MSP}: probability of no linear semi-quantitative dependence between Y and $\hat{Y}$; a larger p_{MSP} (usually > 5%) lead to rejecting of MLR equation	ρ: Spearman ranks correlation coefficient; τ_a: Kendall tau-a ranks correlation coefficient; τ_b: Kendall tau-b ranks correlation coefficient; τ_c: Kendall tau-c ranks correlation coefficient; Γ: Goodman-Kruskal ranks correlation coefficient; P_Z: probability of wrong model (using normal distribution Z);
Leave-one-out cross-validation analysis	$r_{cv\text{-}loo}=r(Y,\hat{\hat{Y}})$, $p_{cv\text{-}loo}=p(r_{cv\text{-}loo},m,df=n)$	$r_{cv\text{-}loo}$: leave-one-out cross-validation correlation coefficient; $p_{cv\text{-}loo}$: probability of no predictive linear model; a larger p_{MP} (usually > 5%) leads to rejecting of MLR equation as predictive linear model;	$\hat{\hat{Y}}=(\hat{\hat{Y}}_k, k = 1..n)$; $\hat{\hat{Y}}_k$ results from the following algorithm: ÷ Remove molecule "k" from sample; ÷ Then $W:=Y\backslash Y_k$; $MDFW_i = MDF_i\backslash MDF_i(k)$; ÷ Apply MLR: $\hat{W} = \Sigma_i b_i MDFW_i$; b_i: real coefficients (MLR coefficients); $\hat{W}$ estimator of W; ÷ $\hat{W}$ predictor for Y_k: $\hat{\hat{Y}}_k = \Sigma_i b_i MDF_i(k)$
Training vs. test experiment	$r_{training}=r(Y\|_{training},\hat{Y}\|_{training})$, $p_{training}=p(r_{training},m_{training},df=n)$ $r_{test}=r(Y\|_{test},\hat{Y}\|_{test})$, $p_{test}=p(r_{test},m_{test},df=n)$	$r_{training}$: correlation between measured ($Y|_{training}$) and estimated ($\hat{Y}|_{training}$) into training subset; $p_{training}$: probability of no linear dependence into training subset; r_{test}: correlation between measured ($Y|_{test}$) and predicted ($\hat{Y}|_{test}$); p_{test}: probability of the no predictive ability of the MLR equation; a larger p_{test} (usually > 5%) combined with a small enough $p_{training}$ (usually < 5%) leads to rejecting of MLR equation as predictive linear model;	test - a random subset of the sample (usually of size of m/3); training - remaining subset of the sample after removing of the test set (usually of size of 2m/3); m_{test} - size of test subset of the sample; $m_{training}$ - size of training subset of the sample; $m_{training} = m - m_{test}$; $Y|_{test}$ - measured activity/property for test subset; $Y|_{training}$ - measured activity/property for training subset; $Y|_{training} = Y\backslash Y_{test}$; $\hat{Y}|_{test}$ results from the following algorithm: ÷ Apply MLR for training set: $\hat{U} = \Sigma_i c_i MDF_i$; c_i: real coefficients (MLR coefficients obtained from training set); ÷ $\hat{U}$ estimator for $Y|_{training}$: $\hat{Y}|_{training}(k) = \Sigma_i c_i MDF_i(k)$, $k \in$ training; ÷ $\hat{U}$ predictor for $Y|_{test}$: $\hat{Y}|_{test}(l) = \Sigma_i c_i MDF_i(l)$, $l \in$ test;

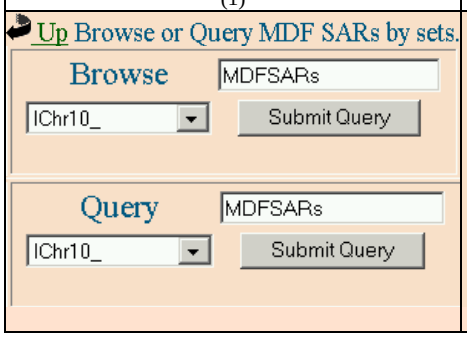
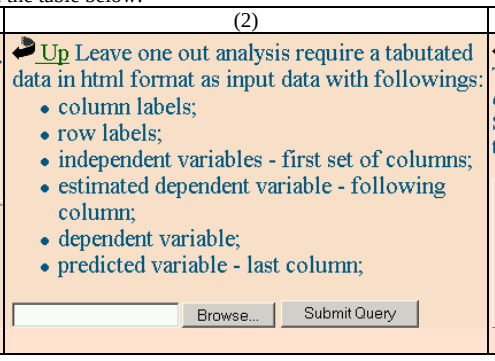
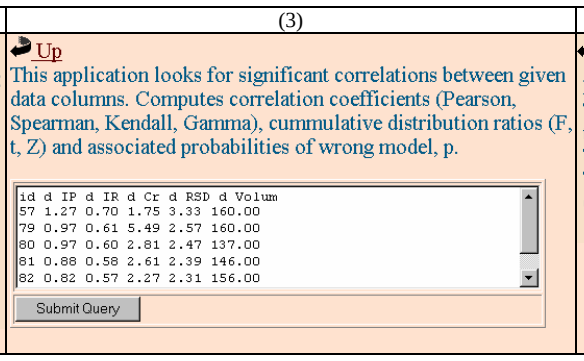
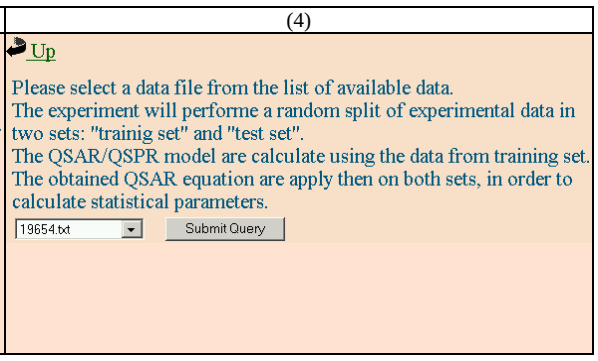
Correlated correlations analysis	$Z_{Steiger}(Y,\hat{Y}1,\hat{Y}2,df12)$	$Z_{Steiger}<Z(5\%)=1.96$: hypothesis of correlated correlations between the estimators $\hat{Y}1$ and $\hat{Y}2$ cannot be rejected with a confidence of 95%;
		$Z_{Steiger}$ can serve for comparing of two MDF-SPARs;
		$Z_{Steiger}$ can serve for comparing of a MDF-SPAR with previous reported SPARs;
			df1: model 1 degrees of freedom (m-n($\hat{Y}1$));
			df2: model 2 degrees of freedom (m-n($\hat{Y}2$));
			df12=min(df1,df2)-3;
			$Z_{Steiger}$ computes from r(Y,$\hat{Y}1$), r(Y,$\hat{Y}2$), r($\hat{Y}1$,$\hat{Y}2$), and df12;

EXPERIMENTAL: Following online applications were developed and used:

http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/k_browse_or_query.php?database=MDFSARs/	(1) Simple correlation analysis; Inter-correlation analysis; Multiple correlation analysis
http://l.academicdirect.org/Statistics/linear_dependence/	(2) Qualitative vs. quantitative analysis
http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/loo/	(3) Leave-one-out cross-validation analysis
http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/qsar_qspr_s/	(4) Training vs. test experiment
http://l.academicdirect.org/Statistics/tests/Steiger/	(5) Correlated correlations analysis

(5)

Snapshots of the applications (1)-(4) are presented in the table below:

(1)	(2)	(3)	(4)
			

RESULTS:

The model with one and two descriptors, respectively proved to has estimated and predictive abilities:

$\hat{Y}_{mono}$ =-0.58+iMDRoQg·8.53 Eq(1)

$\hat{Y}_{bi}$ =-1.36+iMDRoQg·6.03+ISPDwQg·0.08 Eq(2)

The application of the parameters presented in the table bellow leads to the results presented bellow:

Param.	Eq(1)		Eq(2)	
r_{SP}; p_{SP}	0.9514; $5.1\cdot10^{-8}$	✓	0.8806; $1.5\cdot10^{-5}$	✓
r_{MP}; p_{MP}	n.a.		0.9238; $6.2\cdot10^{-9}$	✓
r_{IP}; p_{IP}	n.a.		0.7726; $7.3\cdot10^{-4}$	✓
r_{MS}; p_{MS}	0.9429; $1.4\cdot10^{-7}$	✓	0.9643; $7.1\cdot10^{-9}$	✓
$r_{M\tau a}$; $p_{M\tau a}$	0.8286; $1.7\cdot10^{-5}$	✓	0.8857; $4.2\cdot10^{-6}$	✓
$r_{M\tau b}$; $p_{M\tau b}$	0.8286; $1.7\cdot10^{-5}$	✓	0.8857; $4.2\cdot10^{-6}$	✓
$r_{M\tau c}$; $p_{M\tau c}$	0.7733; $5.9\cdot10^{-5}$	✓	0.8267; $1.7\cdot10^{-5}$	✓
$r_{MГ}$; $p_{MГ}$	0.8286; $3.6\cdot10^{-4}$	✓	0.8857; $4.6\cdot10^{-5}$	✓
r_{MSP}; p_{MSP}	0.9471; $8.7\cdot10^{-8}$	✓	0.9714; $1.7\cdot10^{-9}$	✓
$r_{cv\text{-}loo}$; $p_{cv\text{-}loo}$	0.8744; $9.6\cdot10^{-7}$	✓	0.9158; $1.7\cdot10^{-7}$	✓
r_{tr}^2; p_{tr}	0.8619; $1.0\cdot10^{-4}$	✓	0.9572; $1.6\cdot10^{-5}$	✓
r_{ts}^2; p_{ts}	0.9862; $4.3\cdot10^{-3}$	✓	0.9629; $4.8\cdot10^{-2}$	✓
$Z_{Steiger}$; p	1.7847; 0.074			✗

$m_{training}$- Eq(1) = 10 (valine, cysteine, aspartate, methionine, isoleucine, threonine, glutamate, asparagine, glutamine, alanine); m_{test} = 5 amino acids.

$m_{training}$- Eq(2) = 10 (cysteine, alanine, threonine, leucine, glycine, glutamate, serine, aspartate, valine, phenylalanine)

where: ☑ = statistically significant & ☒ = no difference

MATERIALS:

The hydrophobicity on Hessa et al. scale [10] of fifteen standard amino acids was the property of interest.

The experimental values of hydrophobicity were as follows: alanine (0.11), asparagine (2.05), aspartate (3.49), cysteine (-0.13), glutamine (2.36), glutamate (2.68), glycine (0.74), isoleucine (-0.6), leucine (-0.55), lysine (2.71), methionine (-0.1), phenylalanine (-0.32), serine (0.84), threonine (0.52), and valine (-0.31).

DRUG DESIGN ►

This facility of MDF-SAR allows that having:

÷A set of compounds of interest with known values of property/activity and MDF-SARs obtained, validated, and stored into the database;

÷One of more similar/alike with selected compound(s) set by made of:

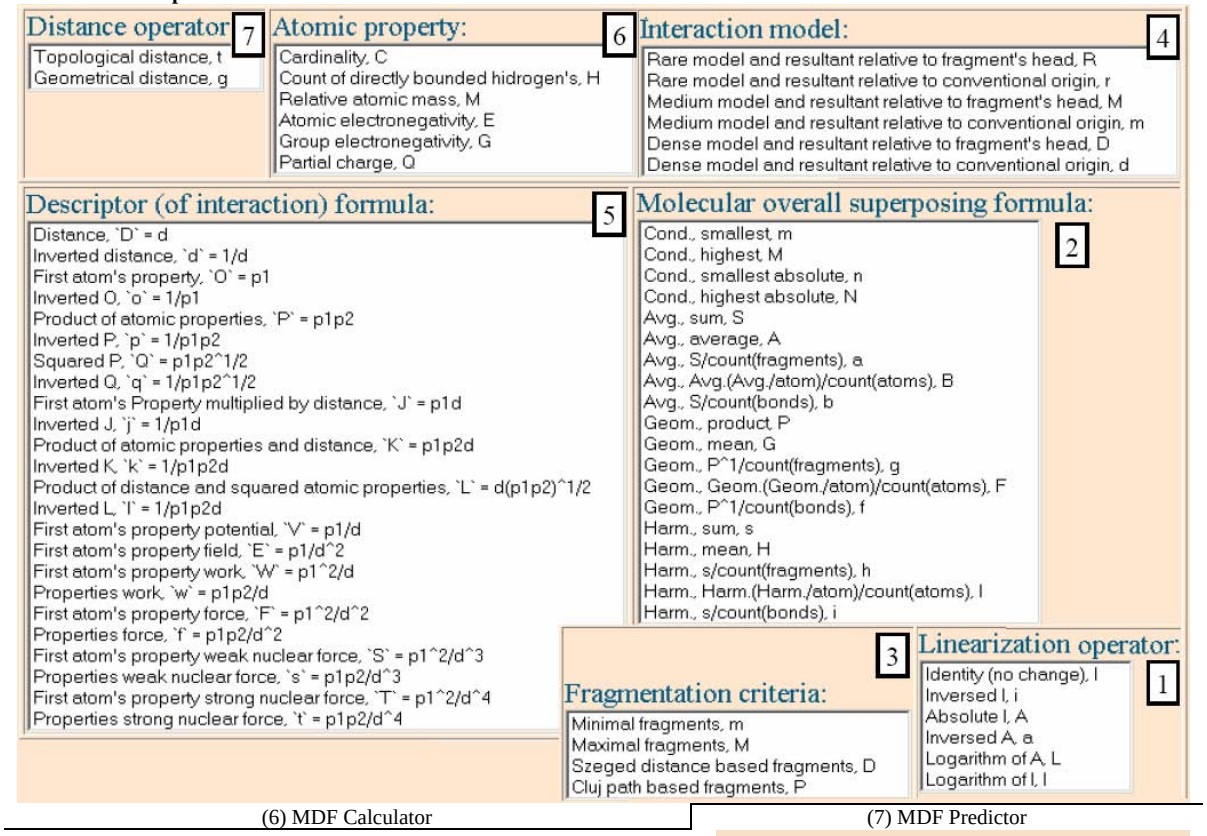
- MDF-SAR equation (MDF predictor);
- building (with HyperChem) of topological (2D) and geometrical (3D) through same choices as were build the selected set

to obtain predicted value(s) for the property / activity of the new compounds, even if this (these) compound(s) were not yet synthesized, in order to see if the new structure (virtual compound at this time) has or not improvements in desired property/activity.

CONCLUSION

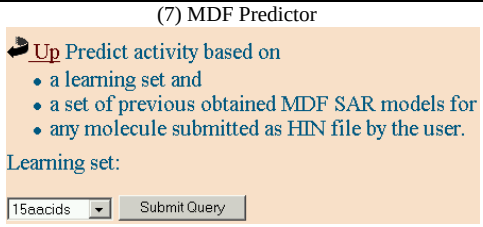
MDF method and MDF-SAR methodology proved to be a very good tool for design of chemical compounds.

MDF-SPAR completion: MDF Calculator & MDF Predictor.



(6) MDF Calculator

(7) MDF Predictor



ACKNOWLEDGEMENTS: [MDF] The MDF project was supported through ET36 research project (2005-2007). [MDF-SAR] The MDF-SAR of MDF is support through ET108 project (2006-2008).

[1] LP Hammett, The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives, J Am Chem Soc 1937;59(1):96-103.
[2] C Hansch, A Leo, RW Taft, A Survey of Hammett Substituent Constants and Resonance and Field Parameters, Chem Rev 1991;91(2):165-195.
[3] Bolboacă SD, Jäntschi L, Modelling the Property of Compounds from Structure: Statistical Methods for Models Validation, Env Chem Lett DOI 10.1007/s10311-007-0119-9.
[4] Heisenberg W, Over descriptive contents of the quantum-theoretical kinetics and mechanics (in German), Zeitschrift für Physik, 1927;43(3-4):172-198.
[5] Jäntschi L, MDF - A New QSAR/QSPR Molecular Descriptors Family, Leonardo J Sci 2004;3(4):68-85.
[6] Jäntschi L, Bolboacă SD, Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information, Int J Quantum Chem 2007;107(8):1736-1744.
[7] Jäntschi L, Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology, Leonardo El J Pract Technol 2005;4(6):76-98.
[8] Jäntschi L, Bolboacă SD, Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships, Int J Mol Sci 2007;8(3):189-203.
[9] Jäntschi L, Bolboacă SD, Diudea MV, Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization, Int J Mol Sci 2007;8(11):1125-1157.
[10] Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G, Recognition of transmembrane helices by the endoplasmic reticulum translocon, Nature 2005;433:377-381.