Spectral – Structure Activity Relationship (Spectral-SAR) Algorithm

Mihai V. Putz

Laboratory of Computational and Structural Physical Chemistry,
Chemistry Department, West University of Timişoara,
Str. Pestalozzi No.16, Timisoara, RO-300115, Romania

Tel: +40-256-592633; Fax. +40-256-592620,
Ems: mvputz@cbg.uvt.ro, mv_putz@yahoo.com

Abstract

With the present-day interest in correlating chemical structure with biological activity the quantitative structure-activity relationships (QSARs) are reviewed both on their fundamental statistical and advanced algebraic frameworks allowing for the so called Spectral-SAR reformulation of the classical Multilinear regression in terms of data vectors and orthogonal conditions, while being suited for inter-endpoint (computed activity) paths and maps of inter-conversion. This way there is presented a novel, fresh and fruitful picture of regression analysis aiming to closely approach the quantum interpretation of data and of ligand-receptor interaction by means of systematic orthogonal and scalar (dot) product of either molecular (chemicals or toxicants) descriptors between them and with the observed (recorded, measured) activities. The resulted Spectral- or Quantum- SAR widely employs the present data as a whole vectors, to be associated in principle with the eigen-states in quantum Hilbert space, opens the way for assigning a sort of wave function or wave packet for the congeneric active molecular series rather than for a single molecule as used to be; this way the specific interaction may be eventually modeled by structure (intrinsic)-metabolic (extrinsic) quantum rather quantitative correlation picture.

1. Introduction

In the last years the world scientific research was focused on the so called *green chemistry*, which consists in the efforts to reduce or eliminate the use or production of the dangerous substances (with toxic potential) in synthesis, main stream and application of the chemical compounds through pre-industrial or computational design [1].

As such on all meridians new specific organizations and laws of validation of the entered compounds in environment or everyday and medical life have raised: the first taxonomical groups emerged in United States by the *Environmental Protection Agency* [2] followed by the European agency *Umweltbundesamt* (1997) and by the *Environment Canada* (1999). However, at the level of European Union, since the Strategy on Management of Substances (SOMS, 2001) [3] program the first step was made towards establishing by the European Commission, on 23 October 2003, to the *Registration, Evaluation, Authorization and Restriction of Chemicals* (REACH) norms establishing, through its directive EC no. 1907/2006, that starting from 2009 any substance with carcinomic or mutagenic potential entering in the life-cycle through market to be made only with authorization of the *European Chemical Agency* (ECMA) at Helsinki [4, 5].

Also Romania, although from the legislative point of view has already the governmental directive OG no. 200/2000, approved by the law no. 451/2001, since 2003 was member of the *Rotterdam Convention* (10 September 2003) being part of the so called *Prior Informed Content* (PIC) procedure relating the priory consent about the risk or toxicity degree of specific chemical that will be circulating or imported across the country. Moreover, since the official membership of Romania at the European Union (1 January 2007), all chemicals on the Romanian territory have to agree with the REACH normative. In this context, the fundamental research is at its turn driven by the EU laws through the directives of the *Organization of Economical and Cooperation Development* (OECD) that already credits the quantitative structure-activity relationship (QSAR) methodology as the only and certain source of computational design for the tested compounds with bio-, eco-, and pharmaco-logical impact [6, 7].

Being used in Chemistry during the second half of 20th century as an extended statistical analysis [8-15], the quantitative structure-activity relationship (QSAR) method had attained in recent years a special status, officially certified by European Union as the main computational tool (within the so called "*in silico*" approach) for the regulatory assessments of chemicals by means of non-testing methods [2-7, 16-18].

However, while QSAR primarily uses the multiple regression analysis [8-15], alternative approaches as such neuronal-network (NN) or genetic algorithms (GA) have been advanced to somehow generalize the QSAR performance in delivering a classification of variables used, in the sense of principal component analysis (PCA) and partial least squares (PLS) methodologies; still, the claimed advantage of the NN over QSAR techniques is limited by the fact the grounding physical-mathematical philosophies are different since highly non-linear with basic multi-linear pictures are compared, respectively [19-26].

Actually, the chemical-physical advantage of QSAR stands in its multi-linearity correlation that resembles with superposition principle of quantum mechanics, which allow meaningful interpretation of the structural (inherently quantum) causes associated with the latent or unobserved variables (sometimes called as *common factors*) into the observed effects (activity) usually measured in terms of 50%-effect concentration (EC_{50}), associated with various types of bioaccumulation and toxicity [27].

Nevertheless, many efforts have been focused on applying QSAR methods to non-linearity features from where the "expert systems" emerged as formalized computer-based environments, involving knowledge-based, rule-based or hybrid automata able to provide rational predictions about properties of biological activity of chemicals or of their fragments; it results in various QSAR based databases: the model database (QMDB) - inventorying the robust summaries of QSARs that can be appealed by envisaged endpoint or chemical, the prediction database (QPDB) - when data from QMDB are used for further prediction to be stored, or together towering the chemical category database (CCD) documentation [28-34].

Therefore, a certain conceptual-computational analysis of a compound of a series of compounds in the view of assigning its toxicity degree naturally two levels: one addresses the atomic-molecular structure together with related quantum properties while the other envisages the correlations of these properties, e.g. hydrophobicity, polarizability, steric effects, etc., with the bio, eco- or pharmaco- logical observed activities. Finally, it gets out the molecular mechanistic <picture> of the reactions involved in the studied chemical-biological interaction or, with other words, of the quantum chemical strength established between the ligand (the effector or the chemical) and receptor (in the target site or organism). Still, either the structure or the quantum chemical binding aspects require the advanced studies upon them, firstly in a separate manner, and then combined both at the intrinsic structural level and for correlating the interaction, based on the versatility of the atomic and molecular world to generate surprisingly structures and interactions just because the quantum character involved (i.e. undulatory, thus allowing the tunneling even for the energetic inaccessible potential barriers) when forming new apparently not explicated or controllable compounds by means of macroscopic procedures.

Still, whatever the computational procedure approached, either of that of Hansch type [35-43], 3D [44-54], decisional [26, 55-66], or orthogonal ones [67-80], the problem of delivering the molecular interaction mechanism as a QSAR analysis result was only recently furnished by the so called Spectral-SAR that proposes a purely algebraic rethinking of the traditional statistic QSAR, which allows, through the new concepts introduced (e.g. the orthogonal space of variables, the vectorial length of the biological activity,

or the algebraic correlation factor as an intensity measure of the chemical-biological interaction) the building of an optimized chart of the molecular action pathways grounded on the *minimum spectral path principle*, $\delta[A,B]=0$ with A and B the endpoints, within a generalized space of the action norms and correlation factors [81-90].

The present review will present the Spectral-SAR method, developed at Timişoara (Romania), as a natural continuation and generalization of the classical standard (statistical) quantitative structure-activity relationship (QSAR) towards the quantum assessment of the ligand-receptor cellular specific interactions, paths and maps.

2. Statistic QSAR

2.1. Scalar (Dot) Product Basics

Often very useful for mathematical elegance but also with a deep insight for the present Spectral-SAR methodology the vectorial modeling of data may be associated with generalized classical-to-quantum description of variables on Hilbert space, beneficial for emphasizing many properties especially those related with orthogonality, i.e. independency of descriptors; this way the quantum most efficient description of a dynamical systems projected on the associated minimum set of commutative (independent) operators assure the maximum predictability in computation and viability in conceptual modeling. Skipping the formal mathematical details, while capping the essence of the computations, being the main operation on Hilbert space (a vectorial space) the scalar or dot product – its main features are shortly reviewed in what follows.

Given two vectors

$$\left|u\right\rangle=\left|u_1,u_2,...,u_n\right\rangle,\left|v\right\rangle=\left|v_1,v_2,...,v_n\right\rangle \quad (1)$$

their scalar (or dot) product writes as:

$$\left\langle u\middle|v\right\rangle \overset{def}{=}\sum_{i=1}^{n}u_iv_i=u_1v_1+u_2v_2+.....+u_nv_n\,. \quad (2)$$

Since the self scalar product looks like:

$$\left\langle u\middle|u\right\rangle=\sum_{i=1}^{n}u_i^2 \quad (3)$$

one may introduce the so called norm (or length) of the vector by:

$$\left\|u\right\rangle\right\|=\sqrt{\left\langle u\middle|u\right\rangle}=\sqrt{\sum_{i=1}^{n}u_i^2}\,. \quad (4)$$

The length property of the vectorial norm may be easily visualized through computing the modulus of an arbitrary 3D vector $\left|r\right\rangle=\left|u_1,u_2,u_3\right\rangle$:

$$\left|\vec{r}\right|=\sqrt{u_1^2+u_2^2+u_3^2}=\sqrt{\left\langle u_1,u_2,u_3\middle|u_1,u_2,u_3\right\rangle}=\sqrt{\left\langle r\middle|r\right\rangle}=\left\|r\right\rangle\right\|; \quad (5)$$

Consequently, the distance between two vectors is written in terms of their difference norm

$$d\left(|u\rangle,|v\rangle\right)=\left\||u\rangle-|v\rangle\right\|=\left\|u-v\right\rangle\|=\sqrt{\langle u-v|u-v\rangle}=\sqrt{\sum_{i=1}^{n}(u_i-v_i)^2}\ . (6)$$

From last relation (but also from the fact that scalar product is positively defined, see above) the distributivity and commutativity properties of scalar product may be employed for any $t\in\Re$ towards equivalent expressions

$$\langle u-tv|u-tv\rangle\geq 0,\ t\in\Re$$

$$\Leftrightarrow\left(\langle u|-\langle v|t\right)\left(|u\rangle-t|v\rangle\right)\geq 0$$

$$\Leftrightarrow\langle v|v\rangle t^2-2\langle u|v\rangle t+\langle u|u\rangle\geq 0\ . (7)$$

The last inequality says that the right side second order equation has no solution or has single equal solutions, a condition fulfilled when its discriminator is less or equal with zero, respectively, leading with the famous *Cauchy-Schwartz inequality*:

$$\langle u|v\rangle^2\leq\langle u|u\rangle\langle v|v\rangle\ (8)$$

rewritten as:

$$\sum_{i=1}^{n}u_iv_i\leq\sqrt{\sum_{i=1}^{n}u_i^2}\sqrt{\sum_{i=1}^{n}v_i^2}\ \ (9)$$

or as

$$\left|\langle u|v\rangle\right|\leq\left\||u\rangle\right\|\cdot\left\||v\rangle\right\|\ . (10)$$

Cauchy-Schwartz inequality is usually successfully employed in probability theory, variance theory and correlation factors, as will be illustrated soon in what following.

2.2. Basic Statistical Indices

Having a set of causes-effects covered "Universe" with either individual and coupled probabilities, as given in Table I, the ergodic statistical (or normalization) condition for their discrete realizations is expressed respectively as:

$$\sum_{i=1}^{M}p_i=1,\ (11a)$$

$$\sum_{j=1}^{N}p_j=1,\ (11b)$$

$$\sum_{i=1}^{M}\sum_{j=1}^{N} p_{ij} = 1\,. \quad (11c)$$

Table I: Schematic representation of the "Universe" by the probability table (and values) with which a certain cause x_k produces certain effect y_k..

Y \ X		x_1 ...	x_k ...	x_M
		p_1 ...	p_k ...	p_M
y_1	p_1	p_{11} ...	p_{1k} ...	p_{1M}
y_2	p_2	p_{21} ...	p_{2k} ...	p_{2M}
⋮	⋮	⋮	⋮	⋮
y_k	p_k	p_{k1} ...	p_{kk} ...	p_{kM}
⋮	⋮	⋮	⋮	⋮
y_N	p_N	p_{N1} ...	p_{Nk} ...	p_{NM}

Yet, in integral representation of the probability field extended to the "Universe" or actions, the condition (11c) rewrites in terms of probability density function $f(x,y)$ as

$$P(x,y)\big|_{UNIVERSE} = \int_x\int_y f(x,y)dxdy = 1\,, \quad (12)$$

while introducing the average of a given observable on a given domain of reality "*D*" in the same manner with the quantum mechanical measurement postulate [91]:

$$\left\langle \hat{\mathbf{A}} \right\rangle = \int_D \psi^* \hat{\mathbf{A}}\psi d\tau = \int \hat{\mathbf{A}}\psi^*\psi d\tau = \int \hat{\mathbf{A}}|\psi|^2 d\tau = \int_x\int_y \hat{\mathbf{A}} f(x,y)dxdy \quad (13)$$

where one easily recognizes the quantity $|\psi|^2$ as the probability density.

In these conditions the average for *x*-values writes as

$$\langle x \rangle = \iint_D x f(x,y)dxdy \quad (14)$$

producing the chain of individual values' departure from average

$$x_1 - \langle x \rangle,\ x_2 - \langle x \rangle, \ldots,\ x_n - \langle x \rangle, \quad (15)$$

and, even more, their squared (positive) counterparts

$$\left(x_1 - \langle x \rangle\right)^2, \left(x_2 - \langle x \rangle\right)^2, \ldots, \left(x_n - \langle x \rangle\right)^2 \quad (16)$$

for defining the *x*-dispersion (or *x*-variance) statements

$$D_x = \begin{cases} \left\langle \left(x - \langle x \rangle\right)^2 \right\rangle = \iint \left(x - \langle x \rangle\right)^2 f(x,y)dxdy \\ \overline{\left(x - \bar{x}\right)^2} = \sum_i p_i \left(x_i - \sum_i x_i p_i \right)^2 \\ \frac{1}{n}\sum_i \left(x_i - \frac{1}{n}\sum_i x_i \right)^2 \end{cases} \quad (17)$$

either expressed under integral, probability, or uniform probability

$$\begin{cases} p_{ij} = p_i = p_j = \frac{1}{n} \\ N = M = n \end{cases} \quad (18)$$

For alternative, more practical definition of variance, one may use the "quantum" average properties to successively get the forms

$$D_x = \left\langle \left(x - \langle x \rangle\right)^2 \right\rangle$$

$$= \left\langle \left(x - \langle x \rangle\right)\left(x - \langle x \rangle\right) \right\rangle = \left\langle x^2 - 2\langle x \rangle x + \langle x \rangle^2 \right\rangle$$

$$= \left\langle x^2 \right\rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2 = \left\langle x^2 \right\rangle - 2\langle x \rangle^2 + \langle x \rangle^2$$

$$= \left\langle x^2 \right\rangle - \langle x \rangle^2 \quad (19)$$

thus providing the celebrated dispersion form (used in Heisenberg indeterminacy principle) from where also its meaning as measuring the error in attributing the average (14) for the *x*-set of values in Table I [92]. Yet, the eq. (19) may be further adapted to the integral, probability and uniform variants, as before:

$$D_x = \begin{cases} \left\langle x^2 \right\rangle - \langle x \rangle^2 = \int\int x^2 f(x,y)dxdy - \left(\iint x f(x,y)dxdy\right)^2 \\ \overline{x^2} - \bar{x}^2 = \sum_i x_i^2 p_i - \left(\sum_i x_i p_i \right)^2 \\ = \frac{1}{n}\sum_i x_i^2 - \frac{1}{n^2}\left(\sum_i x_i \right)^2 \end{cases} \quad (20)$$

Closely related with the dispersion index stays the so called covariance index, which generalizes the variance for two different quantities, here viewed as x-causes and y-effects; it takes one of the forms

$$C_{xy} = \left\langle \left(x - \langle x \rangle\right)\left(y - \langle y \rangle\right) \right\rangle$$

$$= \left\langle xy - x\langle y \rangle - \langle x \rangle y + \langle x \rangle\langle y \rangle \right\rangle = \langle xy \rangle - \langle x \rangle\langle y \rangle - \langle x \rangle\langle y \rangle + \langle x \rangle\langle y \rangle = \langle xy \rangle - 2\langle x \rangle\langle y \rangle + \langle x \rangle\langle y \rangle$$

$$= \langle xy \rangle - \langle x \rangle\langle y \rangle \quad (21)$$

being immediately transcribed into hierarchical way from integral, discrete probabilities, and uniform probabilities, as above

$$C_{xy} = \begin{cases} \langle xy \rangle - \langle x \rangle\langle y \rangle = \int\int xyf(x,y)dxdy - \left(\int\int xf(x,y)dxdy\right)\left(\int\int yf(x,y)dxdy\right) \\ \overline{xy} - \overline{x}\,\overline{y} = \sum_{i,j} x_i y_j p_{ij} - \left(\sum_i x_i p_i\right)\left(\sum_j y_j p_j\right) \\ = \frac{1}{n}\sum_i x_i y_i - \frac{1}{n^2}\left(\sum_i x_i\right)\left(\sum_i y_i\right) \end{cases} \quad (22)$$

Note that the covariance meaning is best understood through imagine the case of its cancellation,

$$C_{xy} = 0 \Rightarrow \langle xy \rangle = \langle x \rangle\langle y \rangle \Rightarrow f(x,y) = f(x)f(y), \quad (23)$$

the case in which the bi-dimensional probability density factorizes into two one-dimensional ones, from where the covariance should account for the "non-separability" of x-causes and y-effects realization probabilities, being this another point where one statistical quantity is reflected by a quantum reality. Worth here remarking that someone would say that this is natural since the quantum theory is often interpreted in terms of probability and in statistical way in general; this is only partially true, while remarking the subtle difference that still exists between quantum mechanics and quantum statistics, to some degree equivalent, but manifestly distinct in regarding time and temperature dependence, respectively [93].

Next, by working with squares of the dispersions, i.e. by defining the standard deviations (σ)

$$\sigma_x = \sqrt{D_x}\ ,\ \sigma_y = \sqrt{D_y} \quad (24)$$

One can combine the covariance and dispersion into the ratio called as the (Pearson) correlation coefficient, written in simple way as:

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{\left\langle \left(x - \langle x \rangle\right)\left(y - \langle y \rangle\right) \right\rangle}{\sqrt{\left\langle \left(x - \langle x \rangle\right)^2 \right\rangle}\sqrt{\left\langle \left(y - \langle y \rangle\right)^2 \right\rangle}} = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_i \left(x_i - \bar{x}\right)^2}\sqrt{\sum_i \left(y_i - \bar{y}\right)^2}} \quad (25a)$$

or in equivalent scalar product fashion:

$$r_{xy}=\frac{\langle x-\bar{x}|y-\bar{y}\rangle}{\sqrt{\langle x-\bar{x}|x-\bar{x}\rangle}\sqrt{\langle y-\bar{y}|y-\bar{y}\rangle}} \quad (25b)$$

The last form gives the elegant opportunity to show its probabilistic character by applying the Cauchy-Schwartz inequality (8) for the vectors (states) $|x-\bar{x}>$, $|y-\bar{y}>$, that is

$$\left|\langle x-\bar{x}|y-\bar{y}\rangle\right| \le \sqrt{\langle x-\bar{x}|x-\bar{x}\rangle}\sqrt{\langle y-\bar{y}|y-\bar{y}\rangle} \quad (26)$$

thus proofing the realm of eq. (25b) as being sub-unitary

$$\left|r_{xy}\right| \le 1 . \quad (27)$$

If necessary, further discrete probability version of Pearson correlation coefficient (25)

$$r_{xy}=\frac{\sum_{ij}x_i y_j p_{ij}-\left(\sum_i x_i p_i\right)\left(\sum_j y_j p_j\right)}{\sqrt{\sum_i {x_i}^2 p_i-\left(\sum_i x_i p_i\right)^2}\sqrt{\sum_j y_j^2 p_j-\left(\sum_j y_j p_j\right)^2}}, \quad (28a)$$

or under uniform probability (18) variant

$$r_{xy}=\frac{n\sum_i x_i y_i-\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{\sqrt{n\sum_i {x_i}^2-\left(\sum_i x_i\right)^2}\sqrt{n\sum_i {y_i}^2-\left(\sum_i y_i\right)^2}} \quad (28b)$$

may be considered with the same interpretation as showing how much from the combined causes-effects probability may be represented as combining separated causes and effects' probabilities; if this relation is identity it means that the causes and effects are distinct realities and may be treated as such, otherwise, for correlation bellow unity there appears that causes are mixed with effects already in their stage of causes, being the effects less observable as distinct (measurable) reality. This heuristic (yet meaningfully) interpretation may be also "geometrically" treated through remembering the classical scalar product between two vectors

$$\vec{x}\cdot\vec{y}=|\vec{x}|\cdot|\vec{y}|\cos(\vec{x},\vec{y}) \quad (29)$$

furnishes the value of the angle between them as the cosines

$$\cos(\vec{x},\vec{y})=\frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} \quad (30a)$$

easily to be generalized for the present vectorial representation of causes and effects data recordings of Table I:

$$\cos\left(|x\rangle,|y\rangle\right)=\frac{\langle x|y\rangle}{\||x\rangle\|\,\||y\rangle\|}=\frac{\langle x|y\rangle}{\sqrt{\langle x|x\rangle}\sqrt{\langle y|y\rangle}}=\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \quad (30b)$$

Since the sub-unitary value of the cosines, the expression (30b) may be treated as another definition of the Pearson correlation (25b) involving just the original data (vectorial) sets, or, equivalently, when their averages are vanishing, $\bar{x}=0,\bar{y}=0$. From this point of view there is clear that the Pearson definition (25b) is more general since involving the average at whatever values, while the eq. (30b) fixes the angle between the causes and effects states: as the angle expresses orthogonality as the cosines goes unity and the two states are better correlated but in the sense of inferring one from other and not interfering one with other. This is a subtle message which we like to stress in synthesizing two conclusions:

- The orthogonality between two correlating sets of data is essential in establishing the qualitative degree of correlation and do not depend on the average of data sets but only by their vectorial length and scalar product through the angle cosines given by eq. (30b);
- Instead, the quantitative degree of correlation is established by invoking the average of concerned data sets through modifying/generalizing the eq. (30) towards the Pearson coefficient (25b).

These are fundamental ideas underlying the motivation and the "philosophy" of quantitative activity (for effects)-structure (for causes) relationships, to be in next step by step unfolded.

2.3. Linear Correlation

Going on with the correlation analysis let's explore the linear correlation between one-cause – the effect relationship and to see the role the correlation factor play on it. For that we consider the one-to-one data sets for x-cause and y-effect, within uniform probability realization of eq. (18), here summarized as the data rows:

X	x_1	x_2	...	x_n
Y	y_1	y_2	...	y_n

Basically, the regression problem consists in finding the best modeling of observed effects by the computed one

$$y_i^{obs}=\underbrace{ax_i+b}_{y^{comp}}+e_i=y^{comp}+e_i \quad (31)$$

through minimizing the errors of such approximation, that is:

$$\begin{cases} e_i^2=\left[y_i-(ax_i+b)\right]^2 \\ \sum_i e_i^2\rightarrow\min \end{cases} \quad (32)$$

Analytically, if the minimization function is introduced as the sum of squared errors

$$f(a,b)=\sum_i e_i^2=\sum_i\left(y_i-ax_i-b\right)^2\rightarrow\min, \quad (33)$$

Then the optimization procedure is to be done in respecting the linear parameters as the free terms and the slope of regression, i.e. providing the system

$$\begin{cases} \dfrac{\partial f(a,b)}{\partial a} = 0 \\ \dfrac{\partial f(a,b)}{\partial b} = 0 \end{cases} \quad (34a)$$

equivalently unfolded as

$$\begin{cases} 2\sum_i \left(y_i - ax_i - b\right)(-x_i) = 0 \\ 2\sum_i \left(y_i - ax_i - b\right)(-1) = 0 \end{cases} \quad (34b)$$

or even as:

$$\begin{cases} \sum_i y_i x_i = a\sum_i x_i^2 + b\sum_i x_i \Big| \cdot n \\ \sum_i y_i = a\sum_i x_i + bn \Big| \cdot (-\sum_i x_i) \end{cases} \quad (34c)$$

solved for the solutions:

$$a = \frac{n\sum_i x_i y_i - \left(\sum_i y_i\right)\left(\sum_i x_i\right)}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}, \quad (35)$$

$$b = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}. \quad (36)$$

Now, worth observing that by multiplying the second equation of the system (34c) with the factor $1/n$ one gets the meaningful expression

$$\frac{1}{n}\sum_i y_i = a\frac{1}{n}\sum_i x_i + b \quad (37)$$

telling that the linear correlation is in fact precisely fulfilled by the data set averages of cause and effect, respectively, i.e.

$$\bar{y} = a\bar{x} + b \quad (38)$$

However, looking to the slope expression (35) and comparing it with the Pearson coefficient (28b) one easily recognize that they are in different statistic quantities although linked by the *x*- and *y*- standard deviations (24) with dispersions of (20) type, namely as

$$a = r_{xy}\frac{\sigma_y}{\sigma_x} = \frac{C_{xy}}{\sigma_x\sigma_y}\frac{\sigma_y}{\sigma_x} = \frac{C_{xy}}{D_x}, \quad (39)$$

where the dependence of slope (35) by the *x*-*y* covariance of (22) and *x*-dispersion (20) was also emphasized.

With these, it is clear that the monovariate linear correlation has the correlation factor as the direct information included in its slope; indeed, if the *x*- and *y*-standard deviations are considered approximately the same,

$$\sigma_x = \sigma_y \quad (40)$$

that happens in the ideal case when both the *x*- and *y*- data sets are described by the same normal distribution, it results in the identity:

$$a = r_{xy}. \quad (41)$$

However, since, in general, we have the case

$$\frac{\sigma_y}{\sigma_x} \neq 1 \quad (42)$$

it is clear that this ratio "modulates" the correlation slope a of eq. (35) to provide the correct, sub-unitary, correlation factor; this explaining why, even in the practical cases of slope higher than unity ($a > 1$), the correlation factor still records sub-unitary values.

Returning to the general linear regression now we can consider the slope-Pearson correlation coefficient of eq. (39) as driven the *instantaneous* equation

$$y = r_{xy}\frac{\sigma_y}{\sigma_x}x + b \quad (40a)$$

along its averaged form, in accordance with eq. (38),

$$\overline{y} = r_{xy}\frac{\sigma_y}{\sigma_x}\overline{x} + b \quad (40b)$$

as well as their difference

$$y - \overline{y} = r_{xy}\frac{\sigma_y}{\sigma_x}\left(x - \overline{x}\right) \quad (41a)$$

from where the computed (predicted) instantaneous effects directly writes from averaged observed ones corrected in a *perturbation* sense by corresponding instantaneous cause departure fro its average

modulated by the Pearson coefficient, which is sub-unitary as earlier proofed by eq. (27), and the ratio of effect-to-cause standard deviations

$$y^{comp} = \bar{y} + r_{xy}\frac{\sigma_y}{\sigma_x}\left(x-\bar{x}\right) \quad (41)$$

Finally, worth giving a practical rule aiming to assure as much possible the premises of a good or relevant correlation in sense of increasing the Pearson correlation factor; it may look like

$$\left|r_{xy}\right|\sqrt{n-1} \geq 3 \,, \quad (45a)$$

which offers a quite reasonable framework depending of the dimension of the data sets accounted as causes and recorded as effects. As such, there is clear that even for data sets containing ten points the condition (42) is still not satisfied,

$$\text{n=10} \Rightarrow \sqrt{n-1} = 3 \,, \ \forall\left|r_{xy}\right|\cdot 3 < 3 \,, \quad (46a)$$

while at least from seventeen dimension of vectorial state with instantaneous cause-effect points the regression analysis may become reasonable:

$$\text{n=17} \Rightarrow \sqrt{n-1} = 4 \,, \ \exists \ \left|r_{xy}\right|\sqrt{n-1} \geq 3 \,. \quad (46b)$$

Nevertheless, from (45a) an even more relaxed condition may be inferred by squaring it to the condition:

$$r_{xy}^2\left(n-1\right) \geq 9 \quad (45b)$$

which may be satisfied even for data sets with cardinal laying in the range ≥ 10. Worth observing that as the data sets for causes is more restrained the correlation factor has to be closer to unity for goodness of the fit. Again, the present discussion has two subtle consequences, namely:

- The cause-effect (linear) regression is meaningful when the number of points included in analysis is significant, and in any case larger than ten;
- The correlation analysis is still relevant, even for lower square of Pearson coefficient as far the number of included cause-effects points is higher enough such that condition (45b) to be fulfilled; this consequence prevent the *ab initio* exclusion of the correlation models with correlation coefficient not laying in the unity vicinity, but when considerable large data set was assumed.

Further insight on correlation coefficient and of its alternative practical definition is to be in next exposed.

2.4. Correlation by Normal Distribution Function

After introducing the main statistical indices and concepts, worth generalizing them with the aid of the distribution function defining the so called (statistical) moment of k-th order for the x-variable (cause):

$$\mu_K = \left\langle \left(x-\langle x\rangle\right)^K \right\rangle = \int \left(x-\langle x\rangle\right)^K f(x)dx \,. \quad (47)$$

Consequently, the first three moments are easily recognized as:

- the normalization condition through the zero-th order moment

$$\mu_0 = \left\langle (x - \langle x \rangle)^0 \right\rangle = \langle 1 \rangle = 1 = \int f(x)dx \quad (48)$$

- cancellation of the first moment:

$$\mu_1 = \left\langle (x - \langle x \rangle)^1 \right\rangle = \int x f(x)dx - \int \langle x \rangle f(x)dx = \langle x \rangle - \langle x \rangle = 0 \quad (49)$$

- the standard deviation through the second moment:

$$\mu_2 = \left\langle \left(x - \langle x \rangle\right)^2 \right\rangle = \sigma_x^2 = \int (x - \bar{x})^2 f(x)dx \quad (50)$$

The remaining problem is the identification of the distribution function; it can be nevertheless chosen as the normal distribution, with the form

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \exp\left[-\frac{(x - \bar{x})^2}{2\sigma_x^2} \right], \; x \in \Re \quad (51)$$

for which the first three moments are verified from eqs. (49)-(51) with the help of Appendix, respectively as:

$$\mu_0 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} \exp\left[-\frac{(x - \bar{x})^2}{2\sigma_x^2} \right] d(x - \bar{x}) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \sqrt{\frac{\pi}{\frac{1}{2\sigma_x^2}}} = 1 , \quad (52a)$$

$$\mu_1 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} (x - \bar{x}) \exp\left[-\frac{(x - \bar{x})^2}{2\sigma_x^2} \right] d(x - \bar{x}) = 0 , \quad (52b)$$

$$\mu_2 = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \int_{-\infty}^{+\infty} (x - \bar{x})^2 \exp\left[-\frac{(x - \bar{x})^2}{2\sigma_x^2} \right] d(x - \bar{x}) = \frac{1}{\sqrt{2\pi} \cdot \sigma_x} \cdot \frac{1}{2 \cdot \frac{1}{2\sigma_x^2}} \cdot \sqrt{2\pi\sigma_x^2} = \sigma_x^2 . \quad (52c)$$

Next, having checked the reliability of the normal function for one variable , the generalized bi-dimensional form may be proposed through considering both the multiplication rule for independent probabilities (say for x-causes and y-effects) tuned by the degree of reciprocal correlation by the Pearson coefficient presence, with the working form:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r_{xy}^2}} \exp\left\{ -\frac{1}{2\left(1 - r_{xy}^2\right)} \left[\frac{(x - \bar{x})^2}{\sigma_x^2} + \frac{(y - \bar{y})^2}{\sigma_y^2} - 2r_{xy} \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x\sigma_y} \right] \right\} ; \quad (53)$$

Note that when the *x*- and *y*- distributions are really independent, i.e. with $r_{xy} = 0$, it reduced to the factorization of the distribution functions of the associate probability fields

$$f_{r_{xy}=0}(x,y)=\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_x}\cdot\exp\left[-\frac{(x-\bar{x})^2}{2\sigma_x^2}\right]\right)\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_y}\cdot\exp\left[-\frac{(y-\bar{y})^2}{2\sigma_y^2}\right]\right)=f(x)f(y) \quad (54)$$

in the same manner as the covariance behavior, previously quoted by the note (23), with the same meaning: when the causes and effects are at all correlated they may be true simultaneously, thus abolishing any ordering hierarchy between them.

Nevertheless, for better emphasizing on the cause role of the x-variable, the conditioned distribution function may be considered as the ratio of the bi-variate normal probability (51) reduced (normalized) by that corresponding to the cause probability, while better modeling the degree with which the effect probability arises when the cause manifestation is certain (and before it); thus the effect conditioned probability function by the cause appearance is successively written as:

$$g(y|x)=\frac{f(x,y)}{f(x)}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)}\left[\frac{(x-\bar{x})^2}{\sigma_x^2}+\frac{(y-\bar{y})^2}{\sigma_y^2}-2r_{xy}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right]+\frac{(x-\bar{x})^2}{2\sigma_x^2}\right\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)}\left[\frac{(x-\bar{x})^2}{\sigma_x^2}+\frac{(y-\bar{y})^2}{\sigma_y^2}-2r_{xy}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}-\frac{(1-r_{xy}^2)(x-\bar{x})^2}{\sigma_x^2}\right]\right\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)}\left[r_{xy}^2\frac{(x-\bar{x})^2}{\sigma_x^2}+\frac{(y-\bar{y})^2}{\sigma_y^2}-2r_{xy}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right]\right\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)}\left[r_{xy}^2\frac{(x-\bar{x})^2}{\sigma_x^2}+\frac{(y-\bar{y})^2}{\sigma_y^2}-2r_{xy}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right]\right\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)}\left[\frac{y-\bar{y}}{\sigma_y}-r_{xy}\frac{x-\bar{x}}{\sigma_x}\right]^2\right\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-r_{xy}^2}}\exp\left\{-\frac{1}{2(1-r_{xy}^2)\sigma_y^2}\left[y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right]^2\right\}; \quad (55)$$

It may, for instance, be used to check out the instantaneous computed/predicted effect, providing successively the expressions

$$y^{comp}=\langle y\rangle_{g(y|x)}=\int yg(y|x)dy$$

$$=\underbrace{\int_{-\infty}^{+\infty}\left[y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right]g(y|x)d\left(y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right)}_{=0}$$

$$+\left[\bar{y}+r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right]\underbrace{\int_{-\infty}^{+\infty}g(y|x)d\left(y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right)}_{=1}$$

$$=\bar{y}+r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \quad (56)$$

until recovering the perturbative form (41) earlier proofed within the linear regression context.

However, being with eq. (56) convinced by the usefulness of the conditioned probability function (55) one may use it for computing the important statistical quantity as the minimum of the squared errors sum, or the *sum of residues* SR in observing the effects from a set of causes, being practically equivalent with the variational calculation of the eq. (34). Indeed, through the following successive identities

$$SR_y=\min\sum_i\left(y_i^{obs}-y_i^{comp}\right)^2$$

$$=\left\langle\left(y_i^{obs}-y_i^{comp}\right)^2\right\rangle_{g(y|x)}=\left\langle\left[y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right]^2\right\rangle_{g(y|x)}$$

$$=\int\left[y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right]^2 g(y|x)d\left(y-\bar{y}-r_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right)$$

$$=\left(1-r_{xy}^2\right)\sigma_y^2 \quad (57)$$

one gets the equation

$$\frac{SR}{\sigma_y^2}=1-r_{xy}^2 \quad (58)$$

leaving with the so called standard or statistical correlation factor

$$R=\sqrt{1-\frac{SR}{\sigma_y^2}}=\sqrt{1-\frac{\sum_i\left(y_i^{obs}-y_i^{comp}\right)^2}{\sum_i\left(y_i^{obs}-\bar{y}\right)^2}}\ . \quad (59)$$

The result of eq. (59), although formally equivalent wit the Pearson correlation factor (28b) adds a very important feature: it describe the correlation cause-effect only through the observed and computed effects so that hiding the causes in the instantaneous computed/predicted effects based upon the regression equation effects-causes. Such formulation is of the first importance and use in evaluating the correlation factors when the multi-regression analysis is employed, since the presence of the many-causes probabilities and correlations – a problem that is avoided when the correlation factor is based only on computed and observed effects, as formula (59) display. Nevertheless, variants of it for may be formulated, for instance the corrected correlation factor that accounts for the dimension of causes and effect vector (state), i.e. the cardinals *M* and *N* of Table I, but this is relevant only for refining applicative discussions, while here we will restraint to only presenting and commenting the fundamental statistical regressions. In this line, in

next, the multi-linear correlation is analytically exposed to complete the statistical presentation of the cause-effect correlation paradigm for the structure-activity modeling, respectively.

2.5. Multilinear Correlation

The many-variable correlation problem may be resumed by finding the *b*'s parameters of the instantaneous equation

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_M X_M \quad (60)$$

when knowing the set of independent (*x*'s) and observed dependent (*y*) variables of Table II.

Table II: *The realization of the Table I within the uniform probability for evaluated (selected) causes of X_k and observed effects of Y columns, respectively.*

Y \ X	X_0	X_1 ...	X_k ...	X_M
y_1	1	x_{11} ...	x_{1k} ...	x_{1M}
y_2	1	x_{21} ...	x_{2k} ...	x_{2M}
⋮	⋮	⋮	⋮	⋮
y_k	1	x_{k1} ...	x_{kk} ...	x_{kM}
⋮	⋮	⋮	⋮	⋮
y_N	1	x_{N1} ...	x_{Nk} ...	x_{NM}

While recognizing the eq. (60) as being associated with the computed instantaneous effect (activity), when the observed counterparts is considered the corresponding errors appear provided the system (61) is fulfilled.

$$\begin{cases} y_1^{obs} = b_0 + b_1 x_{11} + b_2 x_{12} + ... + b_M x_{1M} + e_1 \\ y_2^{obs} = b_0 + b_1 x_{21} + b_2 x_{22} + ... + b_M x_{2M} + e_2 \\ .... \\ y_N^{obs} = b_0 + b_1 x_{N1} + b_2 x_{N2} + ... + b_M x_{NM} + e_N \end{cases} \quad (61)$$

The minimization of the squared sum of errors from (61) respecting each of the searched parameters looks like

$$\begin{cases} \dfrac{\partial}{\partial b_0}\left[\sum_{i=1}^{N} e_i^2\right] = 0 \\ \dfrac{\partial}{\partial b_1}\left[\sum_{i=1}^{N} e_i^2\right] = 0 \\ \ldots \\ \dfrac{\partial}{\partial b_M}\left[\sum_{i=1}^{N} e_i^2\right] = 0 \end{cases} \quad (62a)$$

as a generalization of the linear variational procedure of system (34a); it unfolds analytically firstly as

$$\begin{cases} -2\sum_{i=1}^{N}\left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_M x_{iM}\right)\right]\cdot 1 = 0 \\ -2\sum_{i=1}^{N}\left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_M x_{iM}\right)\right]\cdot x_{i1} = 0 \\ \ldots \\ -2\sum_{i=1}^{N}\left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_M x_{iM}\right)\right]\cdot x_{iM} = 0 \end{cases}, \quad (62b)$$

which can be then rearranged with the form

$$\begin{cases} \sum_{i=1}^{N} y_i = b_0 N + b_1 \sum_{i=1}^{N} x_{i1} + b_2 \sum_{i=1}^{N} x_{i2} + \ldots + b_M \sum_{i=1}^{N} x_{iM} \\ \sum_{i=1}^{N} y_i x_{i1} = b_0 \sum_{i=1}^{N} x_{i1} + b_1 \sum_{i=1}^{N} x_{i2}^2 + \ldots + b_M \sum_{i=1}^{N} x_{iM} x_{i1} \\ \ldots \\ \sum_{i=1}^{N} y_i x_{iM} = b_0 \sum_{i=1}^{N} x_{iM} + b_1 \sum_{i=1}^{N} x_{i1} x_{iM} + \ldots + b_M \sum_{i=1}^{N} x_{iM}^2 \end{cases} \quad (62c)$$

Yet, since the last system has to be solved for *b*'s coefficients a general (formal) solution may be furnished by recognizing it as the formal matrix equation

$$[X]^T[Y] = [X]^T[X][B] \quad (63)$$

with the notations:

$$[Y] = \begin{pmatrix} y_1 \\ y_2 \\ \ldots. \\ y_N \end{pmatrix}, [X] = \begin{pmatrix} 1\, x_{11}\, x_{12} \ldots x_{1M} \\ 1\, x_{21}\, x_{22} \ldots x_{2M} \\ \ldots.. \\ 1\, x_{N1}\, x_{N2} \ldots x_{NM} \end{pmatrix}, [B] = \begin{pmatrix} b_1 \\ b_2 \\ \ldots. \\ b_M \end{pmatrix}, [E] = \begin{pmatrix} e_1 \\ e_2 \\ \ldots. \\ e_N \end{pmatrix}; \quad (64)$$

Equation (63) can be nevertheless directly obtained by reconsidering the system (61) rewritten with notations (64) under the matrix form

$$[Y]=[X][B]+[E] \quad (65)$$

Upon which the optimization condition (62a) is now becoming formally as:

$$\frac{\partial}{\partial[B]}\left([E]^T[E]\right)=0 \quad (65)$$

and where $[X]^T$ stands for the transposition of the $[X]$ matrix.

In any case, the solution of the eq. (63) is immediately abstracted as

$$[B]=\left([X]^T[X]\right)^{-1}[X]^T[Y] \quad (67)$$

often known as the *Moore-Penrose* matrix. Yet, although elegantly obtained it involves the inverse matrix operation which may be quite cumbersome in cases of higher dimensions of the observed effects through the selected causes; it may suffers as well by the indeterminacy in cases in which the matrix inverse is not possible or with singularities. However, it allows for computer routines and is implemented in the majority of the statistical packages.

Since having somehow hidden structure the solution (67) should be checked for the linear-regression case for the knowing analytical solution as given by eqs. (35) and (36). For this special case the system (61) restrains to the simple one

$$\begin{cases} y_1^{obs}=b+ax_1+e_1 \\ y_2^{obs}=b+ax_2+e_2 \\ .... \\ y_N^{obs}=b+ax_N+e_N \end{cases} \quad (68)$$

whereas the involved matrices in general solution (67) are now shaped as

$$[Y]=\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{pmatrix}, [B]=\begin{pmatrix} b=b_0 \\ a=b_1 \end{pmatrix}, [X]=\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \quad (69)$$

Therefore, we firstly construct the matrix to be inversed, namely

$$[A]=[X]^T[X]=\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}=\begin{pmatrix} N & \sum_i^N x_i \\ \sum_i^N x_i & \sum_i^N x_i^2 \end{pmatrix} \quad (70)$$

whose determinant is immediately yielded

$$\det[A] = N\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2 \quad (71)$$

while providing also the minor determinants

$$\widetilde{A}_{11} = (-1)^{1+1}\sum_i^N x_i^2 = \sum_i^N x_i^2 \ , (72a)$$

$$\widetilde{A}_{12} = (-1)^{1+2}\sum_i^N x_i = -\sum_i^N x_i \quad (72b)$$

$$\widetilde{A}_{21} = (-1)^{2+1}\sum_i^N x_i = -\sum_i^N x_i \ , (72c)$$

$$\widetilde{A}_{22} = (-1)^{2+2} N = N \quad (72d)$$

entering the matrix

$$[A]^* = \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{pmatrix} = \begin{pmatrix} \sum_i^N x_i^2 & -\sum_i^N x_i \\ -\sum_i^N x_i & N \end{pmatrix}; (73)$$

all in all the inverse matrix of (70) is obtained as

$$[A]^{-1} = \frac{[A]^*}{\det[A]} = \begin{pmatrix} \dfrac{\sum_i^N x_i^2}{N\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2} & \dfrac{-\sum_i^N x_i}{N\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2} \\ \dfrac{-\sum_i^N x_i}{N\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2} & \dfrac{N}{N\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2} \end{pmatrix}, (74)$$

which together with the other matrices product

$$[X]^T[Y] = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_N \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \sum_i^N y_i \\ \sum_i^N x_i y_i \end{pmatrix} \quad (75)$$

construct the Moore-Penrose matrix for the mono-linear regression

$$[B]=[A]^{-1}\left([X]^{T}[Y]\right)=\begin{pmatrix}\dfrac{\sum_i^N x_i^2\sum_i^N y_i-\sum_i^N x_i\sum_i^N x_i y_i}{N\sum_i^N x_i^2-\left(\sum_i^N x_i\right)^2}\\ \dfrac{-\sum_i^N x_i\sum_i^N y_i+N\sum_i^N x_i y_i}{N\sum_i^N x_i^2-\left(\sum_i^N x_i\right)^2}\end{pmatrix}=\begin{pmatrix}b_0\\ b_1\end{pmatrix}=\begin{pmatrix}b\\ a\end{pmatrix} \quad (76)$$

recovering by its components the solutions (35) and (36). There is however clear by this presentation that as the linear regression is extended to many more (X_i) causes as the complexity and difficulty in analytically expression of the solution factors are increasing; moreover, through such algorithm there appears that no reference and control of the orthogonality or independency among the (X_i) causes are involved, or are so hidden to produce meaningful quantum interpretation for the ligand-receptor specific interaction.

With these the statistical fundaments for treating and understanding the quantitative (regression) structure (cause) – activity (effects) relationships are exposed, while containing the "germens" for alternative and in some respects the generalized algebraic treatment of correlation in assessing for the ligand-receptor specific interaction an analytical pattern towards quantization, as will be presented din the sequel.

3. Algebraic QSAR

3.1. Multivariate Spectral Regression on Hilbert Space

The key concept in SAR discussion regards the independence of the considered structural parameters in Table III. As a consequence we may further employ this feature to quantify the basic SAR through an orthogonal space. The idea is to transform the columns of structural data of Table III into an abstract orthogonal space, where necessarily all predictor variables are independent, solve the SAR problem there and then referring the result to the initial data by means of a coordinate transformation.

Table III: *The vectorial descriptors in a Spectral-SAR analysis.*

Activity	*Structural predictor variables*					
$\left\|Y_{OBS(ERVED)}\right\rangle$	$\left\|X_0\right\rangle$	$\left\|X_1\right\rangle$	$\cdots$	$\left\|X_k\right\rangle$	$\cdots$	$\left\|X_M\right\rangle$
$y_{1\text{-}OBS}$	1	x_{11}	$\cdots$	x_{1k}	$\cdots$	x_{1M}
$y_{2\text{-}OBS}$	1	x_{21}	$\cdots$	x_{2k}	$\cdots$	x_{2M}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{N\text{-}OBS}$	1	x_{N1}	$\cdots$	x_{Nk}	$\cdots$	x_{NM}

Since QSAR models aims correlations between concerned molecular structures and measured (or otherwise evaluated) activity, appears naturally that the *structure* part of the problem to be accommodated within the quantum theory and of its formalisms. In fact, there are few quantum characters that we are using within the present approach [94]:

- Any molecular structural state (dynamical, since undergoes interaction with organism) may be represented by a $|ket\rangle$ *state vector*, in an abstract space of allowed states within Hilbert space, following the $\langle bra|ket\rangle$ Dirac formalism [95]; such states are to be here represented by any reliable molecular index, or, in particular in our study by hidophobicity $|LogP\rangle$, polarizability $|POL\rangle$, total optimized energy $|E_{tot}\rangle$, just to name only the so called Hansch parameters, usually employed for accounting the diffusion, electrostatic and steric effects for molecules acting within organisms' cells, respectively.
- The (quantum) *superposition principle* that assures that sum combinations of molecular states map on other resulting molecular state, here interpreted as bio-, eco- or toxico- logical activity, e.g. $|Y\rangle = |Y_0\rangle + C_{LogP}|LogP\rangle + C_{POL}|POL\rangle + ...$, with $|Y_0\rangle$ meaning the free or unperturbed activity (when all other influences are absent).
- The *orthogonalization feature* of quantum states, a crucial condition for that the superimposed molecular states generates other molecular state (here quantified as molecular-linking organism activity); analytically, the orthogonalization condition is represented by the $\langle bra|ket\rangle$ scalar product of two envisaged states (molecular indices) whom value if it is evaluated to be zero, $\langle bra|ket\rangle = 0$, then the states are said orthogonal and molecular descriptors independent, therefore suitable to be added as states in resulted activity state and as molecular indices in activity correlation. Further details on scalar product and related properties are given in Appendix A1, while in what follows the Spectral-SAR correlation method is resumed.

Therefore the analytical procedure is unfolded in three fundamental steps.

I. Given a set of *N* molecules being studies against biological activity they produce by means of their *M* – structural indicators, all input information (the states) may be vectorial expressed by the columns of the Table III and correlated upon equation

$$|Y_{OBS(ERVED)}\rangle = b_0|X_0\rangle + b_1|X_1\rangle + ... + b_k|X_k\rangle + ... + b_M|X_M\rangle + |prediction\ \ error\rangle$$

$$= |Y_{PRED(ICTED)}\rangle + |prediction\ \ \ error\rangle \quad (77)$$

whith the unity vector $|X_0\rangle = |1\,1\cdots 1_N\rangle$ added to account for the free term.

In order equation (77) to represent a reliable model of the given activities, the molecular states (indices) assumed should constitute an orthogonal set, having this constraint a quantum mechanically fundament, as above described. However, unlike other important studies addressing this problem [67-80], the present employed Spectral-SAR assumes the prediction error vector in eq. (77) as being from beginning orthogonal on all others, since it cannot be considered input data as the others,

$$\langle Y_{PRED}\,|\,prediction\ \ error\rangle = 0\ ; \quad (78)$$

being not known *apriori* any correlation is made. Moreover, from eqs. (77) and (78) there follows that the prediction error vector has to be orthogonal on all other descriptor states of predicted activity.

$$\left\langle X_{i=\overline{0,M}}\,\middle|\,prediction\ \ error\right\rangle = 0\ ; \quad (79)$$

for consistency of the present vectorial (quantum formalized by means of $|ket\rangle$ states) approach. In other terms, conditions (78) and (79) confirm the form (77) in the sense that prediction vector and the prediction activity $|Y_{PRED}\rangle$ (with all its sub-intended states $|X_{i=\overline{0,M}}\rangle$) belongs to disjoint (thus orthogonal) Hilbert spaces; or even more, one can say that the Hilbert space of the observed activity $|Y_{OBS}\rangle$ may be decomposed into a predicted and error independent Hilbert sub-spaces of states.

Therefore within *Timișoara Spectral-SAR* procedure the very beginning step in orthogonalization is prediction vector orthogonalization to prediction activity and of its predictor states, while the remaining orthogonalization algorithm do not search for optimizing the minimization of errors, but for producing the ideal correlation between $|Y_{PRED}\rangle$ and the given descriptors $|X_{i=\overline{0,M}}\rangle$.

II. Next, the Gram-Schmidt orthogonalization algorithm is applied through constructing the orthogonal set of descriptors by means of the consecrated iteration [96-98]:

$$|\Omega_0\rangle = |X_0\rangle, \quad (80a)$$

$$|\Omega_k\rangle = |X_k\rangle - \sum_{i=0}^{k-1} r_i^k |\Omega_i\rangle, \; r_i^k = \frac{\langle X_k | \Omega_i \rangle}{\langle \Omega_i | \Omega_i \rangle}, k = \overline{1,M}; \quad (80b)$$

providing the orthogonal correlation:

$$|Y_{PRED}\rangle = \omega_0 |\Omega_0\rangle + \omega_1 |\Omega_1\rangle + ... + \omega_k |\Omega_k\rangle + ... + \omega_M |\Omega_M\rangle, \; \omega_k = \frac{\langle \Omega_k | Y \rangle}{\langle \Omega_k | \Omega_k \rangle}, \; k = \overline{0,M}. \quad (81)$$

III. Remarkably, while the studies dedicated to orthogonal problem usually stops at this stage the Spectral-SAR uses it to provide the solution for the original searched correlation, eq. (77) with the error vector orthogonal on the predicted activity and on all its predictor states of Table III. This can be adequately achieved through rearranging eqs. (80) and (81) so that the system of all descriptors of Table III to be written in terms of orthogonal descriptors:

$$\begin{cases} |Y_{PRED}\rangle = \omega_0 |\Omega_0\rangle + \omega_1 |\Omega_1\rangle + ... + \omega_k |\Omega_k\rangle + ... + \omega_M |\Omega_M\rangle \\ |X_0\rangle = 1 \cdot |\Omega_0\rangle + 0 \cdot |\Omega_1\rangle + ... + 0 \cdot |\Omega_k\rangle + ... + 0 \cdot |\Omega_M\rangle \\ |X_1\rangle = r_0^1 |\Omega_0\rangle + 1 \cdot |\Omega_1\rangle + ... + 0 \cdot |\Omega_k\rangle + ... + 0 \cdot |\Omega_M\rangle \\ ..... \\ |X_k\rangle = r_0^k |\Omega_0\rangle + r_1^k |\Omega_1\rangle + ... + 1 \cdot |\Omega_k\rangle + ... + 0 \cdot |\Omega_M\rangle \\ ..... \\ |X_M\rangle = r_0^M |\Omega_0\rangle + r_1^M |\Omega_1\rangle + ... + r_k^M |\Omega_k\rangle + ... + 1 \cdot |\Omega_M\rangle \end{cases} \quad (82)$$

The system (82) has no trivial (orthogonal) solution if and only if the associated extended determinant vanishes; this condition introduces the Spectral-SAR determinant and of its equation [81-83]:

$$\begin{vmatrix} |Y_{PRED}\rangle & \omega_0 & \omega_1 & \cdots & \omega_k & \cdots & \omega_M \\ |X_0\rangle & 1 & 0 & \cdots & 0 & \cdots & 0 \\ |X_1\rangle & r_0^1 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \\ |X_k\rangle & r_0^k & r_1^k & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \\ |X_M\rangle & r_0^M & r_1^M & \cdots & r_k^M & \cdots & 1 \end{vmatrix} = 0 \,. \quad (83)$$

If the determinant of eq. (83) is expanded on it first column, and the result rearranged so that to have $|Y_{PRED}\rangle$ on left side and the rest of states/indicators on the right side the searched QSAR solution of the initial problem of eq. (77) is obtained as Spectral-SAR vectorial expansion (from where the "spectral" name is justified as well) with the error vector already absorbed in the orthogonalization procedure.

In fact Spectral-SAR procedure uses the double conversion passages: one forward from the given problem of eq. (77) to the orthogonal one of eq. (81) in which the error vector is orthogonally "dissolved"; and the reverse one, back from the orthogonal to the real descriptors throughout the system (82), leaving with the determinant (83) to be expanded as the QSAR solution.

The result is that now QSAR/Spectral-SAR equation is delivered directly by the determinant (83) and not through matrices products as in statistical Pearson approach, see Section 2.5, while furnishing directly the Spectral-SAR correlation equation and not only the parameters of multi-variate correlation [8-15]. Moreover, the Spectral-SAR algorithm is *invariant* also to the order of descriptors chosen in orthogonalization procedure, providing equivalent determinants just with rearranged lines, a matter that was not previously achieved by other orthogonalization techniques [67-80].

Remarkably, apart from being conceptually new through considering the spectral (orthogonal) expansion of the input data space (of both activity and descriptors) throughout the system (82), the present method also has the computational advantage of being simpler than the classical "standard" statistical way of treating SAR problem previously exposed. That because, one has nothing to do with computations of matrix of the coefficients (64) or (67), this being a quite involving and time consuming procedure for higher dimensional systems. Instead, one can write directly the Spectral-SAR solution (equation) as the expansion of a (M+2)-dimensional determinant of eq. (83) whose components are the activity and structural vectors involving the Gram-Schmidt and the spectral decomposition coefficients, r_i^k and ω_k, respectively.

However, although different from the mathematical procedure, both standard- and spectral-SAR give similar results due to the theorem that states that [96]: *if the matrix X*, as that from (64), *with dimension N×(M+1), N>M+1, has linear independent columns*, i.e. they are orthogonal as in the spectral approach, *then there exists an unique matrix* [Q] *of dimension N×(M+1) with orthogonal columns and a triangular matrix* [R] *of dimension (M+1)×(M+1) with the elements of the principal diagonal equal with* 1, as identified in the first small determinant in eq. (83), *so that the matrix* [X] *can be factorized as*

$$[X] = [Q][R] \,. \quad (84)$$

When combining equation (84) with the optimal equation (63) one can get, after straight algebraic rules, that the [B] vector of estimates takes the form

$$[B] = \left([Q]^T [Q]\right)^{-1} [Q]^T [Y] \quad (85)$$

in close agreement with previous normal one, see equation (67). However, by comparison of matrices $[X]^T[X]$ and $[Q]^T[Q]$ of equations (67) and (85), respectively, there is clear that the last case certainly furnishes a diagonal form which for sure is easier to handle (i.e. to take its inverse) when searching for the vector $[B]$ of SAR coefficients.

However, worth being convinced by the equivalence of the present Spectral method with the standard statistical one by specializing the general problem (77) to the linear case

$$\left|Y_{PRED}\right\rangle = b_0\left|X_0\right\rangle + b_1\left|X_1\right\rangle, \quad (86)$$

and to check whether this is unfolded through the Spectral-equation (83) as providing the parameters of linear regression given by eqs. (35) and (36). In this respect, actually, we deal with the particular equation

$$0 = \begin{vmatrix} \left|Y_{PRED}\right\rangle & \omega_0 & \omega_1 \\ \left|X_0\right\rangle & 1 & 0 \\ \left|X_1\right\rangle & r_0^1 & 1 \end{vmatrix} = \left|Y_{PRED}\right\rangle \begin{vmatrix} 1 & 0 \\ r_0^1 & 1 \end{vmatrix} - \left|X_0\right\rangle \begin{vmatrix} \omega_0 & \omega_1 \\ r_0^1 & 1 \end{vmatrix} + \left|X_1\right\rangle \begin{vmatrix} \omega_0 & \omega_1 \\ 1 & 0 \end{vmatrix}, \quad (87)$$

which is immediately rearranged as

$$\left|Y_{PRED}\right\rangle = \underbrace{\left(\omega_0 - r_0^1\omega_1\right)}_{b}\left|X_0\right\rangle + \underbrace{\omega_1}_{a}\left|X_1\right\rangle \quad (88)$$

so that to identify the actual with the previous linear coefficients of eqs. (35) and (36):

$$a = \omega_1, b = \omega_0 - r_0^1\omega_1 . \quad (89)$$

Going to evaluate the expressions of (89) within the Spectral-SAR algorithm, there is instructive to identify form Table III only the relevant actual variables, with convenient denotation of instantaneous structural ones as the columns:

$\left\|Y_{PRED}\right\rangle$	$\left\|X_0\right\rangle$	$\left\|X_1\right\rangle$
y_1	1	x_1
y_2	1	x_2
$\vdots$	$\vdots$	$\vdots$
y_N	1	x_N

Other working tools are the zero-th and the first orthogonal vectors, accordingly considered and computed respectively as

$$\left|\Omega_0\right\rangle = \left|1\ 1 \cdots 1_N\right\rangle, \quad (90a)$$

$$\left|\Omega_1\right\rangle = \left|X_1\right\rangle - r_0^1\left|\Omega_0\right\rangle$$

$$= \left|x_1\ x_2\ \ldots\ x_N\right\rangle - \frac{1}{N}\sum_{i=1}^{N} x_i \left|1\ 1\ 1\ \ldots\ 1\right\rangle = \left|x_1 - \frac{1}{N}\sum_{i=1}^{N} x_i\ \ldots\ x_N - \frac{1}{N}\sum_{i=1}^{N} x_i\right\rangle, \quad (90b)$$

with the help of coefficient

$$r_0^1 = \frac{\langle X_1 | \Omega_0 \rangle}{\langle \Omega_0 | \Omega_0 \rangle} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad (91)$$

specialized from the general definition (80b).

In the same manner, the other specific Spectral coefficients from the general orthogonal recipe (81) are now for linear regression computed as the zero-th order contribution

$$\omega_0 = \frac{\langle \Omega_0 | Y \rangle}{\langle \Omega_0 | \Omega_0 \rangle} = \frac{1}{N} \sum_i y_i \ , \quad (92)$$

while the first orthogonal one recovers precisely the previous linear slope of eq. (35):

$$\omega_1 = \frac{\langle \Omega_1 | Y \rangle}{\langle \Omega_1 | \Omega_1 \rangle}$$

$$= \frac{\langle x_1 - N^{-1} \sum x_i \ \ldots \ x_N - N^{-1} \sum x_i \, | \, y_1 \ \ldots \ y_N \rangle}{\sum_i \left(x_i - N^{-1} \sum_i x_i \right)^2}$$

$$= \frac{\sum_i y_i \left(x_i - N^{-1} \sum_i x_i \right)}{\sum_i \left(x_i - N^{-1} \sum_i x_i \right)^2} = \frac{\sum_i y_i x_i - N^{-1} \left(\sum_i y_i \right) \left(\sum_i x_i \right)}{\sum_i \left[x_i^2 + N^{-2} \left(\sum_i x_i \right)^2 - 2N^{-1} x_i \sum_i x_i \right]}$$

$$= \frac{N \sum_i y_i x_i - \left(\sum_i y_i \right) \left(\sum_i x_i \right)}{N \sum_i x_i^2 - \left(\sum_i x_i \right)^2} = a \ , \quad (93)$$

as prescribed by the the correspondence of (89). Additionally, also its companion free term coefficient of relationship (88) may be now straightly evaluated as

$$b = \omega_0 - r_0^1 \omega_1$$

$$= \frac{1}{N} \sum_i y_i - \frac{1}{N} \left(\sum_i x_i \right) \frac{N \sum_i y_i x_i - \left(\sum_i y_i \right) \left(\sum_i x_i \right)}{N \sum_i x_i^2 - \left(\sum_i x_i \right)^2}$$

$$= \frac{\left(\sum_i y_i\right)\left(\sum_i x_i\right)^2 - \left(\sum_i x_i\right)\left(\sum_i y_i x_i\right)}{N\sum_i x_i^2 - \left(\sum_i x_i\right)^2}, \quad (94)$$

as well successfully regaining the previously computed linear free terms counterpart as eq. (36), yet by means of variational statistical (optimization of errors' squares summation) procedure.

With this there is clear that the Timisoara Spectral-SAR algebraic SAR methodology not only recovers in great details the standard statistical QSAR routine but also generalizes to a great analyticity extent towards better assessment of mechanistically ordering and influences in practical eco- and bio- logical applications.

3.2. Algebraic Correlation Factor

Let's explore in next whether the present Spectral regression gives the opportunity in defining another correlation index, beyond the standard statistical one given by eq. (59) [94].

One starts with the simple connection between the observed, predicted and error vectors of eq. (77), however specialized on their instantaneous entries:

$$Y_{i-OBS} = Y_{i-PRED} + pe_i \quad (95)$$

where " pe "stays here as abbreviation for "prediction error".

Then, by means of squaring relation (95),

$$Y^2_{i-OBS} = Y^2_{i-PRED} + pe_i^2 + 2Y_{i-PRED} \cdot pe_i, \quad (96)$$

and summing for all working *N*-molecules (of Table III),

$$\sum_{i=1}^{N} Y^2_{i-OBS} = \sum_{i=1}^{N} Y^2_{i-PRED} + \sum_{i=1}^{N} pe_i^2 + 2\sum_{i=1}^{N} Y_{i-PRED} \cdot pe_i, \quad (97)$$

the last relation simplifies to:

$$\sum_{i=1}^{N} Y^2_{i-OBS} = \sum_{i=1}^{N} Y^2_{i-PRED} + \sum_{i=1}^{N} pe_i^2 \quad (98)$$

based on applying of scalar product definition (2) and of prediction error orthogonalization condition (78) upon the last term of (97), i.e.

$$\sum_{i=1}^{N} Y_{i-PRED} \cdot pe_i = \langle Y_{PRED} \,|\, pe \rangle = 0. \quad (99)$$

Now, substituting the prediction error values of (95) into remaining expression (98) one firstly gets:

$$\sum_{i=1}^{N} Y_{i-OBS}^2 = \sum_{i=1}^{N} Y_{i-PRED}^2 + \sum_{i=1}^{N} \left(Y_{i-OBS} - Y_{i-PRED} \right)^2 \quad (100)$$

or the equivalent identity

$$\sum_{i=1}^{N} Y_{i-PRED}^2 = \sum_{i=1}^{N} Y_{i-OBS} \cdot Y_{i-PRED} \ , \ (101)$$

which further rewrites, recalling the norm and scalar product definitions of eqs. (2)-(4), respectively, as:

$$\left\| Y_{PRED} \right\rangle \right\|^2 = \left\langle Y_{OBS} \left| Y_{PRED} \right. \right\rangle . \ (102)$$

Finally, the Cauchy-Schwarz form (10) is employed on the right side term of (102), noting that the observed and predicted activities are of the same nature for a given molecule – i.e. either both positive or both negative – thus providing their scalar product as positively defined; with these, the relation (102) immediately reads as the inequality:

$$\left\| Y_{PRED} \right\rangle \right\|^2 \le \left\| Y_{OBS} \right\rangle \right\| \cdot \left\| Y_{PRED} \right\rangle \right\| \ (103)$$

leaving with the predicted-observed norms' hierarchy

$$\left\| Y_{PRED} \right\rangle \right\| \le \left\| Y_{OBS} \right\rangle \right\| \ (104)$$

that guarantees the *consistent probability definition* while introduceing *algebraic correlation factor* with the form:

$$RA \equiv r_{ALGEBRAIC} = \frac{\left\| Y_{PRED} \right\rangle \right\|}{\left\| Y_{OBS} \right\rangle \right\|} \le 1 \ (105)$$

Nevertheless, there remains to compare this new correlation factor, written in algebraically manner as the ration of predicted – to – observed norms of investigated molecular activity or of their effects, with the fashioned statistical counterpart given by eq. (59); this issue will be addressed in what follows.

3.3. Algebraic vs. Statistic Correlations

Banater Ansatz on the algebraic Spectral-SAR correlation*: for any QSAR analysis, once considering the measured/observed and computed/predicted activity data as the vectors* $\left| Y_{OBS} \right\rangle$ *and* $\left| Y_{PRED} \right\rangle$ *with the associate norms through the scalar products of eqs. (2)-(4), the algebraic norm order* (105) *valid in defining the algebraic correlation factor (104), sets also the hierarchy at the levels of correlations factors in a sense that the algebraic one of always exceed the standard correlation factor (59):*

$$r_{S-SAR}^{ALGEBRAIC} \ge r_{QSAR}^{STATISTIC} . \ (106)$$

Proof: by straight algebraic translation the condition (106) firstly it rewrites as:

$$\frac{\langle Y_{PRED}\,|\,Y_{PRED}\rangle}{\langle Y_{OBS}\,|\,Y_{OBS}\rangle} \geq 1-\frac{\langle Y_{OBS}-Y_{PRED}\,|\,Y_{OBS}-Y_{PRED}\rangle}{\langle Y_{OBS}-\overline{Y}_{OBS}\,|\,Y_{OBS}-\overline{Y}_{OBS}\rangle}, \quad (107)$$

where we have introduced the averaged observed activity

$$\overline{Y}_{OBS}=\frac{1}{N}\sum_{i=1}^{N} y_{i-OBS}\ , \quad (108)$$

and its associate N-dimensional vector (state in Hilbert space):

$$\left|\overline{Y}_{OBS}\right\rangle=\left(\frac{1}{N}\sum_{i=1}^{N} y_{i-OBS}\right)\left|1\ 1\ \ldots\ 1_{N}\right\rangle. \quad (109)$$

Note that the inequality (107) becomes equality in the case of perfect identity between observed and predicted activity values, i.e. perfect correlation, the case in which the second term of the right hand side vanishes while that of the left hand side become unity. For all other non-perfect correlations strict inequality holds and this will be considered in next, for the equivalent expression

$$\langle Y_{PRED}\,|\,Y_{PRED}\rangle\langle Y_{OBS}-\overline{Y}_{OBS}\,|\,Y_{OBS}-\overline{Y}_{OBS}\rangle$$

$$>\langle Y_{OBS}\,|\,Y_{OBS}\rangle\left[\langle Y_{OBS}-\overline{Y}_{OBS}\,|\,Y_{OBS}-\overline{Y}_{OBS}\rangle-\langle Y_{OBS}-Y_{PRED}\,|\,Y_{OBS}-Y_{PRED}\rangle\right], \quad (110a)$$

which may be further rearranged as

$$\left[\langle Y_{PRED}\,|\,Y_{PRED}\rangle-\langle Y_{OBS}\,|\,Y_{OBS}\rangle\right]\left[\langle Y_{OBS}\,|\,Y_{OBS}\rangle-2\langle Y_{OBS}\,|\,\overline{Y}_{OBS}\rangle+\langle\overline{Y}_{OBS}\,|\,\overline{Y}_{OBS}\rangle\right]$$

$$+\langle Y_{OBS}\,|\,Y_{OBS}\rangle\left[\langle Y_{OBS}\,|\,Y_{OBS}\rangle-2\langle Y_{OBS}\,|\,Y_{PRED}\rangle+\langle Y_{PRED}\,|\,Y_{PRED}\rangle\right]>0\,. \quad (110b)$$

At this point, after obvious simplifications and factorization may easily recognize and employ both the identities (102) and (104), specific to algebraic correlation,

$$2\langle Y_{PRED}\,|\,Y_{PRED}\rangle\langle Y_{OBS}\,|\,Y_{OBS}\rangle-2\langle Y_{OBS}\,|\,Y_{OBS}\rangle\underbrace{\langle Y_{OBS}\,|\,Y_{PRED}\rangle}_{\langle Y_{PRED}\,|\,Y_{PRED}\rangle}$$

$$+\underbrace{\left[\langle Y_{OBS}\,|\,Y_{OBS}\rangle-\langle Y_{PRED}\,|\,Y_{PRED}\rangle\right]}_{\geq 0}\left[2\langle Y_{OBS}\,|\,\overline{Y}_{OBS}\rangle-\langle\overline{Y}_{OBS}\,|\,\overline{Y}_{OBS}\rangle\right]>0 \quad (110c)$$

the simplified expression is obtained

$$2\langle Y_{OBS}\,|\,\overline{Y}_{OBS}\rangle>\langle\overline{Y}_{OBS}\,|\,\overline{Y}_{OBS}\rangle \quad (111a)$$

that finally is analytically explicated with the aid of introduced vector (109) of the average activity to the unfolded scalar ordered products

$$2\sum_{i=1}^{N}\left(y_{i-OBS}\frac{1}{N}\sum_{i=1}^{N}y_{i-OBS}\right)>\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}y_{i-OBS}\right)\left(\frac{1}{N}\sum_{i=1}^{N}y_{i-OBS}\right) \quad (111b)$$

leaving with the equivalent strict inequality

$$\frac{2}{N}\left(\sum_{i=1}^{N}y_{i-OBS}\right)^{2}>\frac{1}{N}\left(\sum_{i=1}^{N}y_{i-OBS}\right)^{2} \quad (111c)$$

fully satisfied by the natural ordering as $2>1$. Therefore, there was proofed both the (qualitative) simplicity and the (quantitative) superiority of algebraic correlation factor. Many applications proof these statements also on dedicated molecular-biological or molecular-ecotoxicological cases. Yet, one modern bi-component molecular system concerned the *ionic liquids* (IL) toxicological actions are in next explained in its paradigmatic form.

3.4. Spectral-SAR for Ionic Liquids

Since their emergence a decade ago, ionic liquids have had a constantly growing influence on organic, bio- and green chemistry, due to the unique physico-chemical properties manifested by their typical salt structure: a heterocyclic nitrogen-containing organic cation (in general) and an inorganic or organic anion [99], with melting points below 100 ºC and no vapor pressure [100]. The latter property leads to the practical replacement of conventional volatile organic compounds (VOCs) from the point of view of atmospheric emissions, though they do present the serious drawback that a small amount of IL could enter the environment through groundwater [101]. This risk makes it necessary to perform further eco-toxicological studies of IL on various species, in order to improve the "design rules" for synthesized IL with minimal toxicity to environment integrated organisms.

Ionic liquids display variable stability in terms of moisture and solubility in water, polar and nonpolar organic solvents [102]. Various values of ionic liquid hydrophobicity and polarity may be tailored [101] with the help of nucleoside chemistry [103] according to the main principles of green chemistry [104, 105]: the new chemicals must be designed to preserve effectiveness of function while reducing toxicity, and not persisting in the environment at the end of their usage, but breaking down into inoffensive degradation products.

In this respect, the costs of all approaches for sustainable product design can be reduced using SAR and QSAR methods [84, 85, 89]. It has already been proved that the anti-microbial activity of quaternary ammonium chlorides is lipophilicity-dependent [106]. While the 1-octanol-water partition coefficient could be seen only as the first approximation for compound lipophylicity, bioaccumulation and toxicity in fish, as well as sorption to soil and sediments assumes that lipophylicity is the main factor of anti-microbial activity [107]. Nevertheless, aiming at a deeper understanding of the specific mechanistic description of IL eco-toxicity, it is worth considering that the ionic liquid properties are more comprehensively quantified through lipophylicity, polarizability and total energy as a unitarily complex of factors in developing appropriate structure-activity relationship (SAR) studies.

However, the main problem in assessing the viable QSAR studies to predict ionic liquid toxicities concerns the *anionic-cationic interaction* superimposed on the anionic and cationic subsystems containing ionic liquids. There are basically two complementary ways of attaining this goal. One may address the search of special rules for assessing the anionic-cationic structural separately from the individual anionic and cationic ones, and then generating the QSAR models. Yet, because the cationic and anionic effects on liquid toxicity are merely separately studied at the moment, the appropriate strategy would be to firstly derive the anionic and cationic QSARs and only then to move on to a QSAR of the ionic liquid viewed as an anionic-cationic interaction.

As recently communicated [89], when the ionic liquids activity is evaluated two different additive models for modeling anionic-cationic interaction can be examined.

The first one is based on the vectorial summation of the produced anionic and cationic biological effects $\left|Y\right\rangle$, named the |1+> model, and which is constructed on the superposition of the anionic (subscripted with *A*) and cationic (subscripted with *C*) activities [84]:

$$\left|Y_{AC}\right\rangle^{1+} = \left|Y_C\right\rangle + \left|Y_A\right\rangle \quad (112)$$

The second S-SAR model, named |0+>, is employed when the additive stage is considered at the examined Hansch factors $\left|X = LogP, POL, E_{TOT}\right\rangle$, which are firstly combined to produce the anionic-cationic (subscripted with *AC*) indices that are further used to produce the spectral mechanistic map of the concerned interaction [85]:

$$\left|Y_{AC}\right\rangle^{0+} = \hat{O}_{S-SAR}\left|0+\right\rangle = \hat{O}_{S-SAR}\, f\left(\left\{\left|X_A\right\rangle\right\}, \left\{\left|X_C\right\rangle\right\}\right) \quad (113)$$

with the particular specifications of the spectral vectors:

$$f\left(LogP_A, LogP_C\right) \equiv LogP_{AC} = \log\left(e^{LogP_A} + e^{LogP_C}\right) \in \left\{\left|X_{1AC}\right\rangle\right\}, \quad (114a)$$

$$f\left(POL_A, POL_C\right) \equiv POL_{AC} = \left(POL_A^{1/3} + POL_C^{1/3}\right)^3 \in \left\{\left|X_{2AC}\right\rangle\right\} \text{ [Å}^3\text{]}, \quad (114b)$$

$$f\left(E_A, E_C\right) \equiv E_{AC} = E_A + E_C - 627.71\frac{q_A q_C}{POL_{AC}^{1/3}} \in \left\{\left|X_{3AC}\right\rangle\right\} \text{ [kcal/mol]}. \quad (114c)$$

The open issue addresses whether the |0+> & |1+> states yields with the same results or in which aspects they might differ in the IL ecotoxicity upon certain species. Nevertheless, a practically criteria of deciding upon activity or structure additivity models, between eqs. (112) and (114), respectively, may be set respecting the so called *ionic liquid internal angle* between the anion-cationic activity vectors, with $y_{iA}, y_{iC}, i = \overline{1,N}$ components, abstracted from the general definition (30b), following the prescription [89]:

$$\cos\theta_{AC} = \frac{\sum_{i=1}^{N} y_{iC} y_{iA}}{\sqrt{\sum_{i=1}^{N} y_{iC}^2 \sum_{i=1}^{N} y_{iA}^2}} \begin{cases} \geq 0.707107 \ldots \left|0+\right\rangle \ MODEL \\ < 0.707107 \ldots \left|1+\right\rangle \ MODEL \end{cases} \quad (115)$$

The illustration of the presented S-SAR-IL models was already performed by studying the aquatic species *Vibrio fischeri*, *Daphnia magna* and *Electric El* recorded ecotoxicity against a given tested ionic liquids, appropriately chosen so that containing a wide variety of heads, side chains, and anions. This way, the present methodology may be extended over a wide range of organisms towards designing specific eco-toxicological ionic liquid batteries [87, 108].

4. Spectral-SAR Paths and Quantum-SAR Maps

Having in deep presented the way in which structure – activity correlations may be realized from N recorded activity viewed as effects of M-structural causes, there remains to explore the combinatory of the models (endpoints) obtained along considering different sets of predictor variables in Table III; this is nothing than the QSAR counterpart for what in quantum theory is known as the *complete set of commutative operators* (CoSCOpe) – since in both cases the discussion is to find the minimum (however complete) operators in quantum theory and structural variables in QSAR to behave as independent one each other so that to be independent or orthogonal one each other. Therefore the discussion and analysis based on the various possibilities a QSAR is realized from different structural indices implicitly or explicitly targets the quantum description of the correlation space; here we try to show the first step in exploitation this possibility [109].

Given a set of *N*-molecules, one can chose to correlate their observed activities $A_{i=\overline{1,N}}$ with *M*-selected structural indicators in as many combinations as:

$$C = \sum_{k=1}^{M} C_M^k \ , \ C_M^k = \frac{M!}{k!(M-k)!}, \text{(116a)}$$

linked by different endpoint paths, as many as:

$$K = \prod_{k=1}^{M} C_M^k \ \text{(116b)}$$

indexing the numbers of paths built from connected distinct models with orders (dimension of correlation) from k=1 to k=M.

Basically, for each of the *C*-combinations a correlation (endpoint) QSAR equation is determined, say $Y_{l=\overline{1,C}} = \{y_{il}\}_{l=\overline{1,C}}^{i=\overline{1,N}}$, containing all computed activities for all considered *N*-molecules within the *l*-selected correlation.

Note that the Spectral-SAR version of QSAR analysis computes these activities in a complete non-statistical way, i.e. by assuming the vectors for both observed (activities) and unobserved (latent variables) quantities while furnishing their correlation throughout the specific Spectral-SAR determinant, see eq. (83), obtained from the transformation matrix between the orthogonal (desirable) and oblique (input) correlations. Yet, besides producing essentially the same results as the statistical least-square fit of residues the Spectral-SAR method introduces new concepts reviewed here within three families as follows [109]

I. The Spectral-SAR concepts:

- *The endpoint (computed) spectral norm*

$$\left\| Y_l \right\rangle \| = \sqrt{\langle Y_l \left| Y_l \right\rangle} = \sqrt{\sum_{i=1}^{N} y_{il}^2} \ , \ \ l = \overline{1,C} \ ; \text{(117a)}$$

allowing the possibility of the unique assignment of a number to a specific type of correlation, i.e. performing a sort of resumed quantification of the models;

- *The algebraic correlation factor* of eq. (105) here rewritten as

$$R_{ALG,l}=\frac{\left\|Y_l\right\rangle\right\|}{\left\|A\right\rangle\right\|}=\sqrt{\frac{\sum_{i=1}^{N}y_{il}^2}{\sum_{i=1}^{N}A_i^2}}\ ,\ l=\overline{1,C}\ ,\ (117b)$$

viewed as the ratio of the spectral norm of the predicted activity to that of the measured one, giving the measure of the overall (or summed up) potency of the computed activities respecting the observed one rather than the local (individual) molecular distribution of activities around the mean statistical yields; thus, it is a specific measure of the molecular selection under study, always with a superior value to that yielded from statistical approach, however preserving the same hierarchy in a shrink (less dispersive) manner being therefore better suited for intra-training set molecular analysis.

II. The QSAR map of end-points [109]:

- *The spectral path*, with the distance defined in the Euclidian sense as:

$$\left[l,l'\right]=\sqrt{\left(\left\|Y_l\right\rangle\right\|-\left\|Y_{l'}\right\rangle\right\|\right)^2+\left(R_l-R_{l'}\right)^2}\ ,\ \forall(l,l')=\overline{1,C}\ \ (118)$$

allows for defining complex information as path distances in norm-correlation space with norms computed from eq. (117a) while correlation free to be considered either from statistical (local) or algebraically (global) – eqs. (59) and (117b), respectively; note that as far as computed activity Y_l corresponds to the measured activity A_l defined as logarithm of inverse of 50%-effect concentration (EC50), see bellow, both modulus of Y_l vectors and R values have no units so assuring the consistency of the eq. (118).

- *The least spectral path principle*, formally shaped as:

$$\delta\left[l_1,...l_k...,l_M\right]=0;\ \ l_1,...,l_k,...,l_M : ENDPOINTS\ \ (119)$$

that provides a practical tool in deciding the dominant $\{\alpha,...\}$ hierarchies along the paths constructed by linking all possible k-models (i.e. models with k correlation factors) from (116a) combinations selected one time each on a formed path – generating the so called "M-endpoints containing ergodic path on K-paths assembly" of (116b). However, the implementation of the principle (119) is recursively performed through selecting the least distance computed upon systematically application of eq. (118) on ergodic paths; if, by instance, two paths are equal there is selected that one containing the first two models with shorter norm difference in accordance with the natural least action; the procedure is repeated until all C-models where connected on shortest paths; there was already conjectured that only the first M-shortest paths (called as $\alpha_1,...,\alpha_M$) are enough to be considered for a comprehensive (and self-consistent) mechanistic analysis [34-40].

III. The Quantum-SAR indices and analysis [109]:

- *The inter-endpoint norm difference (IEND)*,

$$\Delta Y_{l|l'}=\left\|Y_{l'}\right\rangle\right\|-\left\|Y_l\right\rangle\right\|,\ (l,l')\in\{\alpha_1,...,\alpha_M\}\ \ (120)$$

that accounts for norm differences of the models lying on the M-shortest spectral paths linking M- from the C-models of Equation (116a);

- *The inter-endpoint molecular activity difference (IEMAD)*,

$$\Delta A_{i|j}^{l|l'} = A_j^{l'} - A_i^{l} = \ln\frac{1}{\left(EC_{50}\right)_j^{l'}} - \ln\frac{1}{\left(EC_{50}\right)_i^{l}} = \ln\frac{\left(EC_{50}\right)_i^{l}}{\left(EC_{50}\right)_j^{l'}} \quad (121)$$

is considered from activity difference between the fittest molecules (*i, j*), in the sense of minimum residues, for the models (*l, l'*) belonging to the shortest paths $\alpha_1,...,\alpha_M$ for which the inter-endpoint norm difference is given by eq. (120).

This way, we can interpret the two fittest molecules (*i, j*) as reciprocally activated by the models (*l, l'*) through the spectral path whom they belong; put in analytical terms, the difference between quantities of eqs. (120) and (121) may assure the "jump" or *transition activity* that turns the effect of *i* molecule on that of *j* molecule across the least spectral (here revealed as metabolization) path connecting the models *l* and *l'*:

$$\ln\frac{1}{q_{i|j}^{l|l'}} \equiv \Delta Y_{l|l'} - \Delta A_{i|j}^{l|l'} . \quad (122)$$

Note that if we rearrange eq. (122) in terms of 505 - effect concentrations of eq. (121) one gets the wave-like form of molecular EC_{50} inter-molecular transformation:

$$\left(EC_{50}\right)_i^{l} = \left(EC_{50}\right)_j^{l'} q_{i|j}^{l|l'} \exp\left(i\Delta Y_{l|l'}\right) \quad (123)$$

providing the analytic continuation in the complex plane for the *IEND* of eq. (120) was assumed, i.e. $\Delta Y_{l|l'} \to i\Delta Y_{l|l'}$ outside the factor $q_{i|j}^{l|l'}$. Remark that although the differences in eqs. (120) and (121) were considered mathematically along the "arrow" *i-to-j* the "quantum transformation" of eq. (123) suggests that the bio-chemical-physical equivalence (metabolization) of the concentration effects evolves *from j-to-i*, revealing a typical quantum behavior with the factor $q_{i|j}^{l|l'}$ playing the propagator role as the quantum kernels in path integral formulation of quantum mechanics [48].

This way, we may assert that eq. (123) stands as the present "quantum"-SAR equation because:

- it involves *the wave-type* expression of molecular effect of concentration, however, for special selected molecules (the fittest out of the *C*-models) and for special selected paths (the least for the *M*-ergodic assembly), being *M* and *C* related by eq. (116a);
- it provides the *specific transition* or specific transformation of the effect of a certain molecule into the effect of another special molecule out from the *N*-trained molecules, paralleling the phenomenology of consecrated quantum transitions;
- it has the amplitude of transformation driven by the so called *quantum-SAR factor* of an exponential form

$$q_{i|j}^{l|l'} = \exp\left(\Delta A_{i|j}^{l|l'} - \Delta Y_{l|l'}\right) \quad (124)$$

  defining the specific quantum-SAR wave;

- it allows the *identity*

$$\left(EC_{50}\right)_i^{l} = \left(EC_{50}\right)_i^{l} \quad (125)$$

  when the reverse effects is considered

$$\left(EC_{50}\right)_j^{l'} = \left(EC_{50}\right)_i^{l} \frac{1}{q_{i|j}^{l|l'}} \exp\left(-i\Delta Y_{l|l'}\right) \quad (126)$$

and substituted in the direct one (123), as absorption and emissions stand as reciprocal quantum effects;

- it has a "phase" with unity norm, in the same manner as ordinary quantum wave functions, allowing the inter-molecular *"real" quantum-SAR transformation*

$$\left|\left(EC_{50}\right)_i^{l}\right| = q_{i|j}^{l|l'} \cdot \left|\left(EC_{50}\right)_j^{l'}\right| \quad (127)$$

exclusively regulated by the quantum-SAR factor of eq. (124), in the same fashion as quantum tunneling is characterized by the transmission coefficient;

- when *multiple transformations* take place across paths with multiple linked models, say (*l*, *l'*, *l''*), the inter-molecular transformation $i \to j \to t$ is characterized by the overall quantum-SAR factor (124) written as product of intermediary ones

$$q_{i|t}^{l|l''} = q_{i|j}^{l|l'} \cdot q_{j|t}^{l'|l''}; \quad (128)$$

due to the two-equivalent ways the $\left(EC_{50}\right)_i^{l}$ effect may be described directly from *t* or intermediated by *j* molecular effect transformations, respectively:

$$\begin{aligned} \left|\left(EC_{50}\right)_i^{l}\right| &= q_{i|t}^{l|l''} \cdot \left|\left(EC_{50}\right)_t^{l''}\right| \\ &= q_{i|j}^{l|l'} \cdot \left|\left(EC_{50}\right)_j^{l'}\right| = q_{i|j}^{l|l'} \cdot \left(q_{j|t}^{l'|l''} \cdot \left|\left(EC_{50}\right)_t^{l''}\right|\right) \end{aligned} \quad (129)$$

in the same way as the quantum propagators behave along quantum paths [48]; certainly, such contraction scheme may be generalized for least paths connecting the *M*-contained *k*-endpoints giving an overall quantum-SAR (*metabolization power*) factor as:

$$q_{i_1|i_M}^{l_1|l_M} = \prod_{w=2}^{M} q_{i_{w-1}|i_w}^{l_{w-1}|l_w} \quad (130)$$

- Equation (123) supports the *self-transformation* as well, with the driven qua-SAR factor given by:

$$q_{i|j=i}^{l|l'} = \exp\left(-\Delta Y_{l|l'}\right) \quad (131)$$

during its evolution along the least paths when the same molecule ($i=j$) is metabolized by activating certain structural features ($l \neq l'$) though specific indicators (variables) in correlation (bindings with receptor site); this case resembles the stationary quantum case according which even isolated (or with free motion), the molecular structures suffer dynamical wave-corpuscular or fluctuant transformation along their quantum paths.

With the present Quantum-SAR methodology one can appropriately identify the molecular pairs that drive certain bio-/eco- activities against given receptor by means of selected descriptors in a "wave"- or "quantum" mechanistic formal way. The ultimate goal will be the computation of quantum-SAR factors along the least paths of actions that give the quantum-map information of the conversion power of the

fittest molecules in their specific bindings [109]. This line is to be in the near future more applied and refined.

5. Conclusion

Paradoxically, the main problem for QSAR resides not in performing the correlation itself but setting the variable selection for it; the mathematical counterpart for such problem is known as the "factor indeterminacy" [110-114] and affirms that the same degree of correlation may be reached with in principle an infinity of latent variable combinations. Fortunately, in chemical-physics there are a limited (although many enough) indicators to be considered with a clear-cut meaning in molecular structure that allows for rationale of reactivity and bindings [115, 116].

Therefore, although undoubtedly useful, the "official" trend in employing QSAR methods is to classify, over-classify and validate through (external or molecular test set) prediction, a gap between the molecular computed orderings and the associate mechanistic role in bio-/eco- activity assessment remains as large as the QSAR strategy has not turned into a versatile tool in identifying the inter-molecular role in receptor binding sites through recorded activities by means of structurally selected common variables; that is to use QSAR information for internal mechanistic predictions among training molecules to see their inter-relation respecting the whole class of observed activities employed for a specific correlation. Such an approach will also be helpful for checking the chemical domain spanned by training molecules – a feature of the paramount importance also for further external tests.

The modern *in silico* (computational) chemical analysis respecting the bio- activity and availability of analogues substances, potentially beneficial or detrimental for specific interaction in organs and organisms, faces with a paradoxical dichotomy: if searching for the best correlation useful for *prediction* of specific molecular bio- or eco- activity QSAR models involving un-interpretable many latent variables may be produced, while always remaining the question of correlation factor indeterminacy (i.e. the assumed descriptors can be at any time replaced with other producing at least the same correlation performances); instead, when restricting the analysis to search for molecular design and mechanisms throughout performing SARs by means of special structural indicators for a given class of relevant molecules, arises the price of limiting the use of generated models for further prediction.

The present review aims filling this gap by deepening the modeling of inter-molecular activity through extending the main concepts of recent developed Spectral-SAR [81-90, 94, 109], developed the fully algebraic version of traditional statistically optimized QSAR picture, targeting the quantification of the competition between molecular inter-activity and inter-endpoints records. As such, the present review was mainly oriented in presenting and developing the second (Q)SAR facet by rationalized the recent introduced notion of spectral-path-linking-endpoints and the associate least action principle to spectral path quantification, in terms of the best fitted molecules, along the contained computed models, by means of the introduced q(uantum)-SAR factor within the generally called Quantum-SAR (QuaSAR) methodology.

On the other side, the so called *green chemistry* stands as a priority field of research which is approached by the research programs of United States and European Commission as well. It has the goal of characterization, prediction and the control of the chemical structures acting as toxicants on organisms and environment. The main reason for such research links the economical, ecological and public health issues in a general paradigm: *method* $\rightarrow$ *data* $\rightarrow$ *information* $\rightarrow$ *knowledge* $\rightarrow$ *use*. Within this epistemological chain *the method* relates the involved procedure in obtaining the experimental data and is regulated by the chemical-physical and biological scientific laws; *the data* represent the chemicals and their toxic or carcinogenic values; *information* refers to elaboration of models through the recorded data; the *knowledge* means the prediction or the final model of the molecular action mechanisms; *the use* is defined by the legal boundaries for the toxic values or classes of chemicals admitted.

In this context, the actual Spectral-to-Quantum SAR project propose an advanced study based on the epistemological bulk data-information-knowledge of the chemicals used in green chemistry in order to

asses: a specific model of quantum characterization of concerned active substances at the bio-, eco- and pharmaco-logic levels through unitary formulation of the atomic-molecular indices for the effector-receptor binding degree potential of the logistic type (including the temporal dependency); a computational consistent model aiming to minimize the residual recorded activities in the experiments studying the enzymic, ionic liquid, antagonists and allosteric inhibition interactions. The methodology allows pattering both the controlling as well as the design of new compounds for synthesis this way eventually covering also the method-and-use segments of the economical-social life in XXI.

Acknowledgements

Authors are truly indebt to Prof. Dr. Eduardo A. Castro from National University of La Plata (UNLA) and La Plata Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), Argentina, for fruitful ideas exchanged on statistical and algebraic correlation analysis during his visit on summer 2009 at Chemistry Department of West University of Timisoara, as well to Dr. Francisco M. Fernández from UNLA-INIFTA for the follow-up useful comments on orthogonality statements of Spectral-SAR algorithm.

Appendix: Common Poisson Integrals

- $I_0(a)=\int_{-\infty}^{+\infty} e^{-ax^2}dx=\sqrt{\frac{\pi}{a}}$ … the 0th order Poisson integral

$$\textit{Proof: } I_0^2(a)=\left(\int_{-\infty}^{+\infty} e^{-ax^2}dx\right)\left(\int_{-\infty}^{+\infty} e^{-ay^2}dy\right)=\int_{-\infty}^{+\infty}\int e^{-a\left(x^2+y^2\right)}dxdy$$

$$=\int_0^{\infty}\int_0^{2\pi} e^{-ar^2}rdrd\varphi=\left(\int_0^{\infty} e^{-ar^2}rdr\right)\left(\int_0^{2\pi} d\varphi\right)$$

$$=-\frac{2\pi}{2a}\int_0^{\infty} d\left(e^{-ar^2}\right)=-\frac{\pi}{a}\left(e^{-ar^2}\right)_0^{\infty}=\frac{\pi}{a}$$

- $I_1(a)=\int_{-\infty}^{+\infty} xe^{-ax^2}dx=0$ … the 1st order Poisson integral

$$\textit{Proof: } I_1(a)=\int_{-\infty}^{+\infty} xe^{-ax^2}dx=-\frac{1}{2a}\int_{-\infty}^{+\infty} d\left(e^{-ax^2}\right)=-\frac{1}{2a}\left(e^{-ax^2}\right)_{-\infty}^{+\infty}=0$$

- $I_2(a)=\int_{-\infty}^{+\infty} x^2e^{-ax^2}dx=\frac{1}{2a}\sqrt{\frac{\pi}{a}}$ … the 2nd order Poisson integral

$$\textit{Proof: } I_2(a) = \int_{-\infty}^{+\infty} x^2 e^{-ax^2}\,dx = \int_{-\infty}^{+\infty} x\left(xe^{-ax^2}\right)dx = -\frac{1}{2a}\int_{-\infty}^{+\infty} x\frac{d}{dx}\left(e^{-ax^2}\right)dx$$

$$= -\frac{1}{2a}\left[\int_{-\infty}^{+\infty}\frac{d}{dx}\left(xe^{-ax^2}\right)dx - \int_{-\infty}^{+\infty} e^{-ax^2}\,dx\right] = -\frac{1}{2a}\int_{-\infty}^{+\infty} d\left(xe^{-ax^2}\right) + \frac{1}{2a}\sqrt{\frac{\pi}{a}}$$

$$= -\frac{1}{2a}\underbrace{\left(xe^{-ax^2}\right)_{-\infty}^{+\infty}}_{(l'Hospital)\to 0} + \frac{1}{2a}\sqrt{\frac{\pi}{a}} = \frac{1}{2a}\sqrt{\frac{\pi}{a}}.$$

References

1. Benfenati E., Predicting toxicity through computers: a changing world, *Chem. Central J.* 2007, 1:32, DOI: 10.1186/1752-153X-1-32
2. EPA US EPA AQUIRE (AQUatic toxicity Information REtrieval). U.S. Environmental Protection Agency.2002. ECOTOX User Guide: ECOTOXicology Database System. Version 3.0 [http://www.epa.gov/ecotox/]. *ECOTOX User Guide: ECOTOXicology Database System*, 2002.
3. SOMS (Strategy on Management of Substances) Ministry of Housing Spatial Planning and Environment; The Hague. http://www2.minvrom.nl/Docs/internationaal/soms_engels.pdf., 2001.
4. European Commission. Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 Dec. 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European Chemicals Agency, amending directive 1999/45/EC and repealing Council Regulation (EC) No. 1488/94 as well as Council Directive 76/769/EEC and commission directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Off. J. Eur. Union, L 396/1 of 30.12.2006*; Office for Official Publication of the European Communities (OPOCE): Luxembourg, 2006.
5. European Commission. Directive 2006/121/EC of the European Parliament and of the Council of 18 Dec. 2006 amending Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances in order to adapt it to Regulation (EC) No. 1907/2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH) and establishing a european chemicals agency. *Off. J. Eur. Union, L 396/850 of 30.12.2006*; Office for Official Publication of the European Communities (OPOCE): Luxembourg, 2006.
6. OECD, Report on the regulatory uses and applications in OECD member countries of (quantitative) structure-activity relationship [(Q)SAR] models in the assessment of new and existing chemicals. Organization of Economic Cooperation and Development: Paris, France, 2006; Available online: http://www.oecd.org/, accessed January 2009.
7. OECD, Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. OECD series on testing and assessment No. 69. ENV/JM/MONO (2007) 2. Organization for Economic Cooperation and Development: Paris, France, 2007; Available online: http://www.oecd.org/, accessed January 2009.
8. Anderson, T.W. *An Introduction to Multivariate Statistical Methods*; Wiley: New York, USA, 1958.
9. Draper, N.R.; Smith, H. *Applied Regression Analysis*, Wiley: New York, USA, 1966.
10. Shorter, J. *Correlation Analysis in Organic Chemistry: An Introduction to Linear Free Energy Relationships*; Oxford Univ. Press: London, UK, 1973.
11. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters*; John-Wiley: New York, USA, 1978.
12. Green, J.R.; Margerison, D. *Statistical Treatment of Experimental Data*; Elsevier: New York, USA, 1978.

13. Topliss, J. *Quantitative Structure-Activity Relationships of Drugs*; Academic Press: New York, USA, 1983.
14. Seyfel, J.K. *QSAR and Strategies in the Design of Bioactive Compounds*; VCH Weinheim: New York, USA, 1985.
15. Chatterjee, S.; Hadi, A.S.; Price, B. *Regression Analysis by Examples*, 3rd Ed.; John-Wiley: New-York, USA, 2000.
16. Worth, A.P.; Bassan, A.; Gallegos Saliner, A.; Netzeva, T.I.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vracko, M. The characterization of quantitative structure-activity relationships: Preliminary guidance. European Commission - Joint Research Centre: Ispra, Italy, 2005; Available online: http://ecb.jrc.it/qsar/publications/, accessed January 2009.
17. Worth, A.P.; Bassan, A.; Fabjan, E.; Gallegos Saliner, A.; Netzeva, T.I.; Patlewicz, G.; Pavan, M.; Tsakovska, I. The characterization of quantitative structure-activity relationships: Preliminary guidance. European Commission - Joint Research Centre: Ispra, Italy, 2005; Available online: http://ecb.jrc.it/qsar/publications/, accessed January 2009.
18. Benigni, R.; Bossa, C.; Netzeva, T.I.; Worth, A.P. Collection and evaluation of [(Q)SAR] models for mutagenicity and carcinogenicity. European Commission - Joint Research Centre: Ispra, Italy, 2007; Available online: http://ecb.jrc.it/qsar/publications/, accessed January 2009.
19. So, S.S.; Karpuls, M. Evolutionary optimisation in quantitative structure-activity relationship: An application of genetic neural network. *J. Med. Chem*. 1996, 39, 1521-1530.
20. Kubinyi, H. Evolutionary variable selection in regression and PLS analysis. *J. Chemometr*. 1996, 10, 119-133.
21. Teko, I.V.; Alessandro, V.A.E.P.; Livingston, D.J. Neutral network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci.* 1996, *36*, 794-803.
22. Kubinyi, H. Variable selection in QSAR studies. 1. An evolutionary algorithm. *Quant. Struct.-Act. Relat*. 1994, 13, 285-294.
23. Haegawa, K.; Kimura, T.; Fanatsu, K. GA strategy for variable selection in QSAR Studies: Enhancement of comparative molecular binding energy analysis by GA-based PLS method. *Quant. Struct.-Act. Relat*. 1999, 18, 262-272.
24. Zheng, W.; Tropsha, A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest neighbour principle. *J. Chem. Inf. Comput. Sci.* 2000, 40, 185-194.
25. Lucic, B.; Trinajstic, N. Multivariate regression outperforms several robust architectures of neural networks in QSAR modelling. *J. Chem. Inf. Comput. Sci.* 1999, 39, 121-132.
26. Duchowicz, P.R.; Castro, E.A. *The Order Theory in QSPR-QSAR Studies*; Mathematical Chemistry Monographs, University of Kragujevac: Kragujevac, Serbia, 2008.
27. Zhao, V.H.; Cronin, M.T.D.; Dearden, J.C. Quantitative structure-activity relationships of chemicals acting by non-polar narcosis - theoretical considerations. *Quant. Struct.-Act. Relat*. 1998, 17, 131-138.
28. Pavan, M.; Netzeva, T.; Worth, A.P. Review of literature based quantitative structure-activity relationship models for bioconcentration. *QSAR Comb. Sci*. 2008, 27, 21-31.
29. Pavan, M.; Worth, A.P. Review of estimation models for biodegradation. *QSAR Comb. Sci*. 2008, 27, 32-40.
30. Tsakovska, I.; Lessigiarska, I.; Netzeva, T.; Worth, A.P. A mini review of mammalian toxicity (Q)SAR models. *QSAR Comb. Sci*. 2008, 27, 41-48.
31. Gallegos Saliner, A.; Patlewicz, G.; Worth, A.P. A review of (Q)SAR models for skin and eye irritation and corrosion. *QSAR Comb. Sci*. 2008, 27, 49-59.
32. Patlewicz, G.; Aptula, A.; Roberts, D.W. Uriarte, E. A mini-review of available skin sensitization (Q)SARs/Expert systems. *QSAR Comb. Sci*. 2008, 27, 60-76.
33. Netzeva, T.; Pavan, M.; Worth, A.P. Review of (quantitative) structure-activity relationship for acute aquatic toxicity. *QSAR Comb. Sci*. 2008, 27, 77-90.
34. Cronin, M.T.D.; Worth, A.P. (Q)SARs for predicting effects relating to reproductive toxicity. *QSAR Comb. Sci*. 2008, 27, 91-100.
35. Ogihara, N. Drawing out drugs. *Mod. Drug Discovery* 2003, 6 (9), 28-32.
36. Hansch, C.; Hoekman, D.; Gao, H. Comparative QSAR: toward a deeper understanding of chemicobiological interactions. *Chem. Rev.* 1996, 96, 1045-1075.
37. Kubinyi, H. Der Schlüssel zum Schloß I. Grundlagen der Arzneimittelwirkung. *Pharmazie in unserer Zeit* 1994, 23 Jahrg. Nr.3, 158-168.

38. Liwo, A.; Tarnowska, M.; Grzonka, Z. Tempczyk, A. Modified Free-Wilson method for the analysis of biological activity data. *Computers Chem.* 1992, 16, 1-9.
39. Schmidli, H. Multivariate prediction for QSAR. *Chemometrics and Intelligent Laboratory Systems* 1997, 37, 125-134.
40. Lhuguenot, J.-C. Relation quantitative structure-activité (QSAR): une méthode mal reconnue car trop souvent mal utilisée. *Ann. Fals. Exp. Chim.* 1995, 88, 293-310.
41. Crippen, G. M.; Bradley, M. P.; Richardson, W. W. Why are binding-site models more complicated than molecules? *Perspectives in Drug Discovery and Design* 1993, 1, 321-328.
42. Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, 1986.
43. Balaban, A.T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure-activity correlations. *Top. Curr. Chem.* 1983, 114, 21-55.
44. Navia, M. A.; Peattie, D. A. Structure-based drug design: applications in immunopharmacology and immunosuppression. *Immunology Today* 1993, 14, 296-301.
45. Perkins, T. D. J.; Dean, P. M. An exploration of a novel strategy for superposing several flexible molecules. *J. Comput.-Aided Mol. Design* 1993, 7, 155-172.
46. Lemmen, C.; Lengauer, T. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Design* 1997, 11, 357-368.
47. Balaban, A. T.; Chiriac, A.; Motoc, I; Simon, Z. *Steric Fit in QSAR*; Springer, Berlin (Lecture Notes in Chemistry Series), 1980.
48. Simon, Z; Chiriac, A.; Holban, S.; Ciubotariu, D.; Mihalas, G. I. *Minimum Steric Difference. The MTD Method for QSAR Studies*; Res. Studies Press (Wiley), Letchworth, 1984.
49. Duda-Seiman C., Duda-Seiman D., Dragoş D., Medeleanu M., Careja V., Putz M.V., Lacrămă A.-M., Chiriac A., Nuţiu R., Ciubotariu D. Design of anti-HIV ligands by means of minimal topological difference (MTD) Method, *Int. J. Mol. Sci.* 2006, 7, 537-555.
50. Cramer, R.D.III; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 1988, 110, 5959-5967.
51. Cramer, R.D.III; DePriest, S.A.; Patterson, D.E.; Hecht, P. The developing practice of comparative molecular field analysis. In *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), Escom, Leiden, 1993, pp. 443-485.
52. Sun, J.; Chen, H.F.; Xia, H.R.; Yao, J.H.; Fan, B.T. Comparative study of factor Xa inhibitors using molecular docking/SVM/HQSAR/3D-QSAR methods. *QSAR Comb. Sci.* 2006, 25, 25-45.
53. Randić, M.; Jerman-Blazic, B.; Trinajstić, N. Development of 3-dimensional molecular descriptors. *Comput. Chem.* 1990, 14, 237-246.
54. Randić, M.; Razinger, M. Molecular topographic indices. *J. Chem. Inf. Comput. Sci.* 1995, 35, 140-147.
55. Manallack, D. T.; Livingstone, D. J. Artificial neural networks: application and chance effects for QSAR data analysis. *Med. Chem. Res.* 1992, 2, 181-190.
56. Manallack, D. T.; Livingstone, D. J Limitations of functional-link nets as applied to QSAR data analysis. *Quant. Struct-Act. Relat.* 1994, 13, 18-21.
57. Marchant, C. A.; Combes, R. D. Artificial intelligence: the use of computer methods in the prediction of metabolism and toxicity, in *Bioactive Compound Design: Possibilities for Industrial Use*, M. G. Ford, R. Greenwood (eds.), G. T. Brooks and R. Franke BIOS Scientific Publishers Limited, 1996.
58. Moriguchi, I.; Hirono, S.; Matsushita, Y.; Liu, Q.; Nakagome, I. Fuzzy adaptive least squares applied to structure-activity and structure-toxicity correlations. *Chem. Pharm. Bull.* 1992, 40, 930-934.
59. Moriguchi, I.; Hirono, S. Fuzzy adaptive least squares and its use in quantitative structure-activity relationships, in *QSAR and Drug Design – New Developments and Applications*, T. Fujita (ed.), Elsevier Science B. V., 1995.
60. Vapnik, V.N. *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
61. Vapnik, V.N. *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, Berlin, 1982.
62. Schölkpof, B.; Burges, C.J.C.; Smola, A.J. (eds.) *Advances in Kernel Methods. Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
63. Schölkpof, B.; Smola, A.J. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.
64. Mangasarian, O.L.; Musicant, D.R. Succesive overrelaxation for support vector machines. *IEEE Trans. Neural Networks* 1999, 10, 1032-1036.

65. Mattera, D.; Palmieri, F.; Haykin, S. Simple and robust methods for support vector expansions. *IEEE Trans. Neural Networks* 1999, 10, 1038-1047.
66. Luan, F.; Ma, W.P.; Zhang, X.Y. ; Zhang, H.X. ; Liu, M.C. ; Hu, Z.D. ; Fan, B.T. QSAR study of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls using the Heuristic method and support vector machine. *QSAR Comb. Sci.* 25, 25, 46-55.
67. Sutter, J. M.; Kalivas, J. H.; Lang, P. K. Which principal components to utilize for principal component regression. *J. Chemometrics* 1992, 6, 217-225.
68. Nendza, M.; Wenzel, A. Statistical approach to chemicals classification. *Environ. Toxicol. Chem.* 1993, *Supplement*, 1459-1470.
69. Cash, G. G.; Breen, J. J. Principal component analysis and spatial correlation: environmental analytical software tools. *Chemosphere* 1992, 24, 1607-1623.
70. Hemmateenejad, B.; Miri, R.; Jafarpour, M.; Tabarzad, M.; Foroumadi, A. Multiple linear regression and principal component analysis-based prediction of the anti-tuberculosis activity of some 2-aryl-1,3,4-thiadiazole derivatives. *QSAR Comb. Sci.* 2006, 25, 56-66.
71. Randić, M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.*1991, 31, 311-320.
72. Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* 1991, 15, 517-525.
73. Amić, D.; Davidović-Amić, D.; Trinajstić, N. Calculation of retention times of anthocyanins with orthogonalized topological indices. *J. Chem. Inf. Comput. Sci.* 1995, 35, 136-139.
74. Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D. The structure-property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* 1995, *35, 532-538.*
75. Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. A Structure-property study of the solubility of aliphatic alcohols in water. *Croatica Chem. Acta* 1995, 68, 417-434.
76. Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. A Novel QSPR approach to physicochemical properties of the α-amino acids. *Croatica Chem. Acta* 1995, 68, 435-450.
77. Šoškić, M.; Plavšić, D.; Trinajstić, N. Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.* 1996, 36, 829-832.
78. Klein, D.J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. Hierarchical orthogonalization of descriptors. *Int. J. Quantum Chem.* 1997, 63, 215-222.
79. Ivanciuc, O.; Taraviras, S.L.; Cabrol-Bass, D. Quasi-orthogonal basis sets of molecular graph descriptors as chemical diversity measure. *J. Chem. Inf. Comput. Sci.* 2000, 40, 126-134.
80. Fernandez, F. M.; Duchowicz, P. R.; Castro E. A. About orthogonal descriptors in QSPR/QSAR theories, *MATCH Commun. Math. Comput. Chem.* 2004, 51, 39-57.
81. Putz, M.V. A spectral approach of the molecular structure – biological activity relationship part I. The general algorithm. *Ann. West Univ. Timişoara Ser. Chem.* 2006, 15, 159-166.
82. Putz, M.V.; Lacrămă, A.-M. A spectral approach of the molecular structure – biological activity relationship part II. The enzymatic activity. *Ann.West Univ. Timişoara Ser. Chem.* 2006, 15, 167-176.
83. Putz, M.V.; Lacrămă, A.-M. Introducing spectral structure activity relationship (S-SAR) analysis. Application to ecotoxicology. *Int. J. Mol. Sci.* 2007, 8, 363-391.
84. Lacrămă, A.-M.; Putz, M.V.; Ostafe, V. A Spectral-SAR model for the anionic-cationic interaction in ionic liquids: Application to *Vibrio fischeri* ecotoxicity. *Int. J. Mol. Sci.* 2007, 8, 842-863.
85. Putz, M.V.; Lacrămă, A.-M.; Ostafe V. Spectral-SAR ecotoxicology of ionic liquids. The *Daphnia magna* case. *Res. Lett. Ecol.* 2007, Article ID12813/5 pages, DOI: 10.1155/2007/12813.
86. Putz, M.V.; Duda-Seiman, C.; Duda-Seiman, D.M.; Putz A.-M. Turning SPECTRAL-SAR into 3D-QSAR analysis. application on H^+K^+-ATPase inhibitory activity, *Int. J. Chem. Model.* 2008, 1, 45-62.
87. Lacrămă, A.-M.; Putz, M.V.; Ostafe, V. Designing a spectral structure-activity ecotoxico-logistical battery, in *Advances in Quantum Chemical Bonding Structures*, Putz M.V., Ed.; Transworld Research Network: Kerala, India, 2008; Chapter 16, pp. 389-419.
88. Putz, M.V.; Putz (Lacrămă) A.-M. Spectral-SAR: Old wine in new bottle. *Studia Universitatis Babeş-Bolyai Chemia*, 2008, 53, 73-81.
89. Putz, M.V.; Putz, A.-M.; Ostafe, V.; Chiriac A. Application of spectral-structure activity relationship (S-SAR) method to ecotoxicology of some ionic liquids at the molecular level using acethylcolinesterase. *Int. J. Chem. Model.* 2009, 2, 85-96.

90. Putz, M.V.; Putz, A.M.; Lazea, M.; Chiriac, A. Spectral vs. statistic approach of structure-activity relationship. Application on ecotoxicity of aliphatic amines. *J. Theor. Comput. Chem.* 2010, 8 in press.
91. Daudel R.; Leroy G.; Peeters D.; Sana M. *Quantum Chemistry*, John Wiley & Sons, New York, 1983.
92. Messiah, A. *Quantum Mechanics*, Vols. I and II, North-Holland: Amsterdam, Holland, 1961.
93. Weiss, U. *Quantum Dissipative Systems*, World Scientific, Singapore, 1993.
94. Chicu, S.A.; Putz, M.V. Köln-Timişoara molecular activity combined models toward interspecies toxicity assessment. *Int. J. Mol. Sci.* 2009, 10, accepted.
95. Dirac, P.A.M. *The Principles of Quantum Mechanics*, Oxford University Press: Oxford, UK, 1944.
96. Fadeeva V. N. *Computational Methods of Linear Algebra*, Dover Publications, New York, 1959.
97. Steen, L.A. Highlights in the history of spectral theory. *Amer. Math. Monthly* 1973, *80*, 359-381.
98. Siegmund-Schultze, R. Der Beweis des Hilbert-Schmidt Theorem. *Arch. Hist. Ex. Sc.* 1986, 36, 251-270.
99. Pernak, J.; Chwala, P. Synthesis and anti-microbial activities of choline-like quaternary ammonium chlorides. *Eur. J. Med. Chem.* 2003, 38, 1035-1042.
100. Bernot, R.J.; Brueseke, M.A.; Evans-White, M.A.; Lamberti, G.A. Acute and chronic toxicity of imidazolium-based ionic liquids on *Daphnia Magna*. *Environ. Toxicol. Chem.* 2005, 24, 87-92.
101. Sheldon, R.A. Green solvents for sustainable organic synthesis: state of the art. *Green Chem.* 2005, 7, 267-278.
102. Docherty, K.M.; Kulpa, C.F.Jr. Toxicity and antimicrobial activity of imidazolium and pyridinium ionic liquids. *Green Chem.* 2005, 7, 185-189.
103. Freemantle, M. New frontiers for ionic liquids. *Chem. Eng. News* 2007, 1, 23-26.
104. Anastas, P.T.; Warner, J.C. *Green Chemistry Theory and Practice*, 1998, Oxford University Press, New York.
105. Jastorff, B.; Molter, K.; Behrend, P.; Bottin-Weber, U.; Filser, J.; Heimers, A.; Ondurschka, B.; Ranke, J.; Scaefer, M.; Schroder, H.; Stark, A.; Stepnowski, P.; Stock, F.; Stormann, R.; Stolte, S.; Welz-Biermann, U.; Ziegert, S.; Thoming, J. Progress in evaluation of risk potential of ionic liquids—basis for an eco-design of sustainable products. *Green Chem.* 2005, 7, 362-372.
106. Jastorff, B.; Stormann, R.; Ranke, J.; Molter, K.; Stock, F.; Oberheitmann, B.; Hoffmann, W.; Hoffmann, J.; Nuchter, M.; Ondruschka, B.; Filser, J. How hazardous are ionic liquids? Structure – activity relationship and biologic testing as important elements for sustainability evaluation. *Green Chem.* 2003, 5, 136-142.
107. Pernak, J.; Sobaszkiewicz, K.; Mirska, I. Antimicrobial activities of ionic liquids. *Green Chem.* 2003, 5, 52-56.
108. Lacrămă, A.M. *Ecotoxicological Batteries with Organisms from Different Species* (in Romanian), PhD dissertation, West University of Timişoara, Romania, 2007.
109. Putz, M.V.; Putz, A.M.; Lazea, M.; Ienciu, L.; Chiriac A. Quantum-SAR extension of the Spectral-SAR algorithm. Application to polyphenolic anticancer bioactivity. *Int. J. Mol. Sci.* 2009, 10, 1193-1214.
110. Steiger, J.H.; Schonemann, P.H. A history of factor indeterminacy. In *Theory Construction and Data Analysis in the Behavioural Science*, Shye, S., Ed.; Jossey-Bass Publishers: San Francisco, CA, USA, 1978.
111. Spearman, C. *The Abilities of Man*; MacMillan: London, UK, 1927.
112. Wilson, E.B. Review of the abilities of man, their nature and measurement, by Spearman, C. *Science* 1928, 67, 244-248.
113. Wilson, E.B.; Hilferty, M.M. The distribution of chi-square. *Proc. Nat. Acad. Sci. USA* 1931, 17, 684.
114. Wilson, E.B.; Worcester, J. A note on factor analysis. *Psychometrika* 1939, 4, 133-148.
115. Topliss, J.G.; Costello, R.J. Chance correlation in structure-activity studies using multiple regression analysis. *J. Med. Chem.* 1972, 15, 1066-1068.
116. Topliss, J.G.; Edwards, R.P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* 1979, 22, 1238-1244.