Simple analytical method for evaluation of statistical importance of correlations in QSAR studies

J. Pecka [a] and R. Ponec [b,*]

[a] *Department of Organic Chemistry, Charles University, Prague 2 – Albertov, Czech Republic*
[b] *Institute of Chemical Process Fundamentals, Czech Academy of Sciences,*
Prague 6, Suchdol 2, 165 02, Czech Republic

Received 12 January 2000

A new general method allowing to evaluate the statistical importance of independent correlations from multilinear QSAR models was proposed. The method is applicable to any multilinear QSAR model involving N experimental points and M independent parameters. The application of this new method is demonstrated on several examples.

1. Introduction

One of the basic goals of any natural science is the formulation of simple models and concepts, in terms of which it is possible to describe, understand, and explain the observed phenomena. Sophistication of these models depends, of course, on the degree of development of a given science and, also, on the complexity of the studied problems. In this respect one of the fields resisting most strongly the rigorous theoretical description is represented by the broad area of the relations between the structure of molecules and their properties. For this reason the important role in formulating such relations still belongs to various empirical rules and concepts. Such is, e.g., the situation with the basically empirical Hammett and Taft equations [5,6,11–14], characterizing quantitatively the electronic effects of substituents on reaction rates and equilibria. The same idea of expressing the molecular property in terms of linear combination of certain parameters (descriptors) has found, however, wide acceptance in other areas of chemistry [1,4,16] and beyond. The typical example in this respect is represented by the systematic studies by Hansch [2,7–10], who extended the application of empirical structure–activity relationships to the correlation of biological activities. Since its introduction several decades ago, this approach, now known as the QSAR approach, has become a respected and widely used methodology in rational drug design. Because the factors determining the biological activity are extremely complex and often several of them act in parallel, the corresponding QSAR models usually have the form of multilinear correlations with several empirical descriptors (electronic and steric constants

[*] Corresponding author.

of substituents, hydrophobic parameter $\log P$ etc.). Parameters of these relations are usually determined using least-squares procedure. Within this approach, the quality of resulting empirical relations is usually classified using various statistical parameters as, e.g., the standard deviation, correlation coefficient, cross validated correlation coefficient [3] etc. Although these parameters certainly provide a certain measure of the "tightness" of the linear relationship between the correlated quantities in any particular case, the comparison of independent QSAR models is much more difficult. Thus while the statistical importance of two correlations with the same number of points and the same number of parameters can simply be done by comparing the values of correlation coefficients (the better correlation is the one for which the correlation coefficient r is higher), no such simple test exists for correlations in which either the number of points N or the number of parameters M (or both) are different. The typical situation in this respect is when one- and two-parameter correlations on the same set of data are to be compared. It is clear that the addition of the second parameter automatically results in better correlation coefficient, but the question is whether this increase is high enough to justify also the higher statistical importance of the two-parameter correlation. Our aim in this study is to propose a new simple method allowing one to address and to solve the above kind of problems.

The basic idea of our approach can best be demonstrated on the following simple example. Let us imagine that one has to correlate a set of N experimental points with some descriptors, and let us further assume that this correlation leads to the correlation coefficient R. Now, let us suppose that instead of correlating actual experimental data, we attempt at the same correlation with the set of N randomly chosen values λ_i. It is clear that in most cases the correlation coefficient r of such a random correlation will be smaller than R. But if we perform such a randomization test many times it is possible, especially when the value of R itself was not very high, that the correlation coefficient r of this random correlation could be equal to or higher than R. The statistical importance of our correlation can thus be naturally evaluated according to the probability that correlation coefficient $r > R$ is obtained accidentally. The above approach, performed numerically, is the basis of the randomization test [15] but in this study we propose a method replacing the above brute force numerical procedure by explicit analytical approach. In the following part the basic idea of our approach will be presented.

2. Theoretical

Let us assume a general QSAR model in which a set of N experimental points $(y'_1, y'_2, \ldots, y'_N)$ is correlated with the set of M linearly independent parameters (descriptors) $(x'_{j1}, x'_{j2}, \ldots, x'_{jN};\ j = 1, 2, \ldots, M)$ using the empirical QSAR equation

$$y'_i = \sum_j^M a_j x'_{ji} + b. \tag{1}$$

Let us further assume that these quantities are subject to random experimental errors characterized by the Gaussian normal distribution. The parameters a_j are determined according to least-squares criterion:

$$\Delta = \sum_i^N \left(y'_i - \sum_j^M a_j x'_{ji} - b \right)^2, \tag{2}$$

$$\frac{\partial \Delta}{\partial a_j} = 0, \quad j = 1, 2, \ldots, M, \tag{3}$$

$$\frac{\partial \Delta}{\partial b} = 0. \tag{4}$$

Although the above least-square procedure can be applied directly to the set of experimental values y'_i and descriptors x'_{ji}, it is convenient to transform them into a new set of "centered" variables y_i and x_{ji} defined as

$$y_i = y'_i - \bar{y}, \qquad x_{ji} = x'_{ji} - \bar{x}_j, \tag{5}$$

where

$$\bar{y} = \frac{1}{N} \sum_i^N y'_i, \qquad \bar{x}_j = \frac{1}{N} \sum_i^N x'_{ji}. \tag{6}$$

Using these centered variables the correlation coefficient is defined as

$$r = \sqrt{\frac{\sum_j^M a_j \sum_k^N y_k x_{jk}}{\sum_k^N y_k^2}}, \tag{7}$$

and it is possible to show that its value is invariant to any linear scaling of variables y_i and x_{ji}. In addition to this invariance it is also possible to show that the centered variables have the same Gaussian distribution of errors as the original set of vectors y'_i and x'_{ji}, so that their statistical distribution also does not depend on the direction. Moreover, the centered variables are no longer linearly independent since because of equation (8) only $N - 1$ components can be chosen independently:

$$\sum_i^N y_i = 0, \qquad \sum_i^N x_{ji} = 0, \quad j = 1, 2, \ldots, M. \tag{8}$$

As a consequence, the vectors of centered variables can be regarded as a point on a surface of $(N - 1)$-dimensional sphere. The correlation coefficient then depends only on the angle between the vector $\vec{Y}$ and the (hyper)-plane spanned by M vectors $\vec{X}_j$:

$$Y_{\text{theor},i} = \sum_j^M \alpha_j X_{ji} + \beta. \tag{9}$$

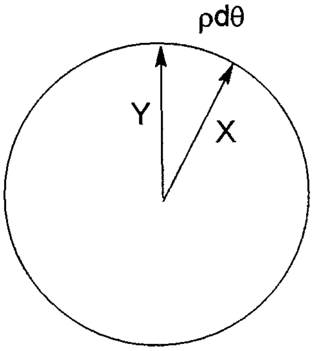

Figure 1.

Among all vectors lying in the (hyper)-plane (9), the one for which the coefficients α, β have been determined from least squares criterion represents the best statistical approximation to $\vec{Y}$.

The simplest situation is in the case of single-parameter linear correlation ($M = 1$) where the plane reduces to a vector $a\vec{X}$. In this case, because of invariance to linear scaling, the correlation coefficient is given by

$$r = \cos\bigl(\vec{X}, \vec{Y}\bigr). \tag{10}$$

This example is especially convenient for the demonstration of the basic idea of our approach. For this purpose let us imagine the simplest case of linear correlation ($M = 1$) with only $N = 3$ experimental points and let the actual correlation coefficient be equal to R. In this case the vectors $\vec{X}$ and $\vec{Y}$ can be regarded to lie on the surface of the two-dimensional ($N - 1$)-sphere, i.e. the circle (see figure 1). The angle of these vectors is then equal to $\vartheta = \arccos(R)$. Depending on the actual value of R, this angle can vary between zero, corresponding to $R = 1$, and $\pi/2$, corresponding to $R = 0$.

Now, let us imagine that instead of using actual (centered) experimental values of variables y_1, y_2, y_3, we calculate the correlation coefficient with the set of randomly generated values $\lambda_1, \lambda_2, \lambda_3$ (vector $\vec{\Lambda}$). It is very probable, of course, that the correlation coefficient r will be in most cases smaller than R, especially when the value of R itself is high enough. In other words, the probability of randomly generating a good correlation is low. If we now realize that a high value of correlation coefficient implies small angle between the vectors $\vec{X}$ and $\vec{Y}$, the low value of the correlation coefficient found in majority of random correlations implies that the angle between the vectors $\vec{X}$ and $\vec{\Lambda}$ is greater than $\arccos(R)$. But it is also possible (and the probability of such an event increases with the decrease of R) that the randomly found correlation will be equally good or better than the one with actual experimental values. The probability that this happens can be calculated in this case on the basis of simple geometrical considerations (figure 1). It is clear, namely, that the desired probability P is equal to the ratio of the length of the arc between the endpoints of the vectors $\vec{X}$ and $\vec{Y}$

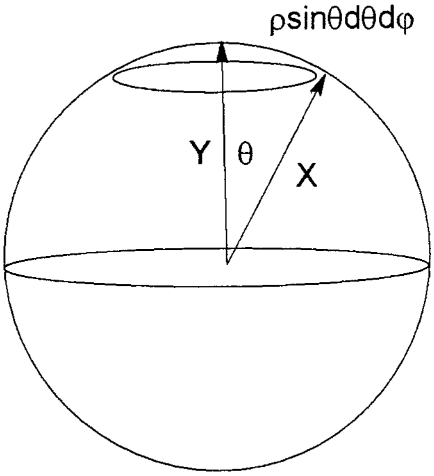

Figure 2.

(i.e. for the angle of these vectors equal to arccos(R)) and the length of the same arc for the angle $\theta = \pi/2$ corresponding to $R = 0$:

$$P = \frac{\int_0^{\operatorname{arccos}(R)} \rho \, \mathrm{d}\vartheta}{\int_0^{\pi/2} \rho \, \mathrm{d}\vartheta} = \frac{2}{\pi} \operatorname{arccos}(R). \tag{11}$$

Similarly transparent is the calculation of the probability for the case of a linear correlation with $N = 4$ experimental points. In this case, namely, the vectors $\vec{X}$ and $\vec{Y}$ can be regarded to lie on a surface of a three-dimensional sphere (figure 2). On the basis of analogous geometrical arguments as before it is clear that the desired probability is given by the ratio of the surfaces of spherical cap for $\theta = \operatorname{arccos}(R)$ and the half-sphere $\theta = \pi/2$:

$$P = \frac{\int_0^{2\pi} \mathrm{d}\varphi \int_0^{\operatorname{arccos}(r)} \rho \sin\vartheta \, \mathrm{d}\vartheta}{\int_0^{2\pi} \mathrm{d}\vartheta \int_0^{\pi/2} \rho \sin\vartheta \, \mathrm{d}\vartheta} = 1 - R. \tag{12}$$

The interpretation of these equations is simple. They say that if we found, for example, in the case of correlation with 4 points the correlation coefficient $R = 0.9$, then there is 10% probability $(1 - R)$ that such a correlation could be obtained accidentally. On the other hand, if the correlation coefficient is 0.99, the probability of finding equally good correlation accidentally is only 1%.

In connection with these formulas it is worth reminding that the value of probability does not depend on the "length" ρ of the vectors $\vec{X}$ and $\vec{Y}$. So it is possible, without loss of generality, to require them to be normalized to unity. This can always be done by simple scaling, since the value of the correlation coefficient is invariant with respect to such scaling. In a similar way it is possible to calculate the desired probabilities also for general linear correlation with N experimental points. The only important difference compared to both previous examples is that the vectors $\vec{X}$ and $\vec{Y}$

lie now on a surface of an ($N-1$)-dimensional hypersphere for which there is no possibility of simple graphical visualization. This, however, represents no serious obstacle and the necessary integration can advantageously be performed in generalized spherical coordinates defined by

$$\begin{aligned} y_1 &= \cos\vartheta, \\ y_2 &= v_1 \sin\vartheta, \\ y_3 &= v_2 \sin\vartheta, \\ &\vdots \\ y_{N-1} &= v_{N-2} \sin\vartheta, \end{aligned} \tag{13}$$

where the auxiliary parameters $v_1, v_2, \ldots, v_{N-1}$ are bound by the normalization

$$\sum_j^{N-2} v_j^2 = 1 \tag{14}$$

required for the normalization of the vector $\vec{Y}$. Thus, for example, in the case of linear correlation with $N = 4$ experimental points, the above formula for the transformation to generalized spherical coordinates reduces to

$$\begin{aligned} y_1 &= \cos\vartheta, \\ y_2 &= \cos\varphi \sin\vartheta, \\ y_3 &= \sin\varphi \sin\vartheta, \end{aligned} \tag{15}$$

which corresponds to identification $v_1 = \cos\varphi,\ v_2 = \sin\varphi$.

The area element on the surface of the abstract ($N-1$)-dimensional sphere is then given by

$$\mathrm{d}\Sigma = \frac{1}{v_{N-2}} \sin^{N-2}\vartheta \,\mathrm{d}\vartheta \,\mathrm{d}v_1 \,\mathrm{d}v_2 \ldots \mathrm{d}v_{N-3}. \tag{16}$$

Based on this expression, the final formula for the probability is given by

$$\begin{aligned} P &= \frac{\int_0^{\arccos(R)} \sin^{N-2}\vartheta \,\mathrm{d}\vartheta \int_\Delta (1/v_{N-2}) \,\mathrm{d}v_1 \,dv_2 \ldots \mathrm{d}v_{N-3}}{\int_0^{\pi/2} \sin^{N-2}\vartheta \,\mathrm{d}\vartheta \int_\Delta (1/v_{N-2}) \,\mathrm{d}v_1 \,dv_2 \ldots \mathrm{d}v_{N-3}} \\ &= \frac{\int_0^{\arccos(R)} \sin^{N-2}\vartheta \,\mathrm{d}\vartheta}{\int_0^{\pi/2} \sin^{N-2}\vartheta \,\mathrm{d}\vartheta}. \end{aligned} \tag{17}$$

These probabilities can be calculated for any particular value of N and their values, which depend on the value of the correlation coefficient R, determine the so-called level of importance of the correlation. These values are, of course, often used in evaluating the statistical importance of the correlations, but as far as we know, the critical values of the correlation coefficient for each pre-selected level (1%, 5% etc.) can be found in statistical tables only for one-dimensional case. What is new in our approach is that we report here the analytical formula allowing simple quantitative evaluation of the level of importance for any multi-linear correlation with M independent parameters.

In order to show the basic idea of such a generalization let us analyse in detail another simple example, for which the calculation of the probability is again possible on the basis of simple geometrical considerations. This example concerns the correlation of $N = 4$ experimental points using two-parameter ($M = 2$) correlation equation. Let us assume that this correlation gives the value of correlation coefficient R. In this case, the vector $\vec{Y}$ lies on the surface of an ordinary three-dimensional sphere and the vector Y_{theor} lies in the plane

$$Y_{\text{theor}} = a_1X_1 + a_2X_2 + b. \tag{18}$$

The correlation coefficient R is in this case given by the angle between the vector $\vec{Y}$ and the plane (18).

The calculation of the probability that the correlation coefficient r of the randomly generated correlation is greater or equal than the actual one R can be in this case calculated using simple geometrical considerations in the following spherical coordinates:

$$\begin{aligned} y_1 &= \sin\varphi\cos\vartheta, \\ y_2 &= \cos\varphi\cos\vartheta, \\ y_3 &= \sin\vartheta. \end{aligned} \tag{19}$$

As it is possible to see, the first two components define a plane in which lies the vector Y_{theor} (it can be generated for any given value of ϑ by systematically varying φ within the interval $0-2\pi$), while the third component defines the angle ϑ between the vector $\vec{Y}$ and the plane (see figure 3).

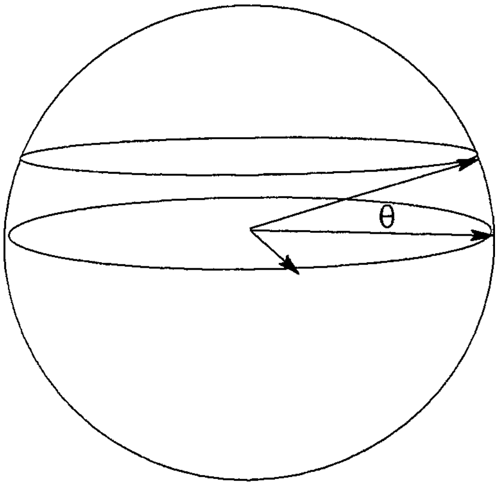

Figure 3.

Based on this scheme it is clear that the desired probability is given by the ratio of the surfaces of the spherical trip around the equator (for ϑ varying between 0 and $\arccos(R)$) and the half-sphere.

If we now take into account that in terms of coordinates (19), the surface element of the sphere with unit radius is given by

$$\mathrm{d}S = \cos\vartheta \, \mathrm{d}\vartheta \, \mathrm{d}\varphi, \tag{20}$$

the following final formula for the probability can straightforwardly be obtained:

$$P = \frac{\int_0^{\arccos(R)} \cos\vartheta \, \mathrm{d}\vartheta \int_0^{2\pi} \mathrm{d}\varphi}{\int_0^{\pi/2} \cos\vartheta \, \mathrm{d}\vartheta \int_0^{2\pi} \mathrm{d}\varphi} = \frac{\int_0^{\arccos(R)} \cos\vartheta \, \mathrm{d}\vartheta}{\int_0^{\pi/2} \cos\vartheta \, \mathrm{d}\vartheta}. \tag{21}$$

This choice of spherical coordinates, in which the angle ϑ is measured from the plane of the equator, is especially useful in this case since as it is possible to see, the value of the probability depends only on the angular variable ϑ related via (10) to the correlation coefficient R. Similar idea of using first M components of the vector Y to define the (hyper)-plane

$$Y_{\text{theor},i} = \sum_j^M a_j X_{ji} + b \tag{22}$$

allows one to introduce the following set of generalized spherical coordinates, in terms of which the calculation of the probability is especially straightforward:

$$\begin{aligned} y_1 &= u_1 \cos\vartheta, \\ y_2 &= u_2 \cos\vartheta, \\ y_3 &= u_3 \cos\vartheta, \\ &\vdots \\ y_M &= u_M \cos\vartheta, \\ y_{M+1} &= v_1 \sin\vartheta, \\ y_{M+2} &= v_2 \sin\vartheta, \\ &\vdots \\ y_{N-1} &= v_{N-M-1} \sin\vartheta, \end{aligned} \tag{23}$$

where again the auxiliary variables u and v satisfy the normalization conditions

$$\sum_i^M u_i^2 = 1, \qquad \sum_j^{N-M-1} v_j^2 = 1, \tag{24}$$

ensuring the proper normalization. Using the general notation, the previous example can be identified with the choice $u_1 = \cos\varphi$, $u_2 = \sin\varphi$, $v_1 = 1$ for which the normalization (24) is automatically satisfied. In terms of the above generalized coordinates, the area element on the surface of $(N-1)$-dimensional sphere is given by

$$\begin{aligned} \mathrm{d}\Sigma = {} & \frac{1}{u_M}\frac{1}{v_{N-M-1}}\cos^{M-1}\vartheta\,\sin^{N-M-2}\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}u_1\,\mathrm{d}u_2\ldots\,\mathrm{d}u_{M-1} \\ & \times\mathrm{d}v_1\,\mathrm{d}v_2\ldots\,\mathrm{d}v_{N-M-2} \end{aligned} \tag{25}$$

which straightforwardly leads to the following final formula for probability:

$$P = \frac{\int_0^{\arccos(R)}\cos^{M-1}\vartheta\,\sin^{N-M-2}\vartheta\,\mathrm{d}\vartheta}{\int_0^{\pi/2}\cos^{M-1}\vartheta\,\sin^{N-M-2}\vartheta\,\mathrm{d}\vartheta}. \tag{26}$$

Using this formula the value of probability can be calculated (using, for example, MATHCAD, MAPLE or other similar programs) for any particular combination of N, M and R and these probabilities can consequently be used for comparing the statistical importance of different QSAR models. As an example, let us consider two following hypothetical correlations:

(1) simple linear correlation ($M = 1$) with 10 experimental points and the correlation coefficient $R = 0.95$,

(2) two-parameter correlation ($M = 2$) with the same number of experimental points and the value of $R = 0.97$, and let us ask what of these correlations is statistically more important (“better”). Such a decision can unambiguously be done just on the value of probabilities for each particular case. These values, calculated according to formula (19) using MATHCAD program, are $2.5 \cdot 10^{-5}$ and $5.0 \cdot 10^{-5}$, respectively. As the first probability is smaller than the second one, it is possible to conclude that the first, single-parameter correlation is statistically more important than the second one and so it is not necessary in this case to invoke the second parameter. In order to be more significant than the single-parameter correlation, the value of the correlation coefficient R of the second correlation should be at least 0.98.

Acknowledgements

The authors thank Prof. Ramon Carbó-Dorca (University of Girona) and Dr. Subhash Basak (University of Minnesota) for critically reading the manuscript and for fruitful comments.

References

[1] M.M. Bursey, in: *Advances in LFER*, eds. N.B. Chapman and J. Shorter (Plenum Press, New York, 1978) p. 445.

[2] A. Cammarata and K.S. Rogers, in: *Advances in LFER*, eds. N.B. Chapman and J. Shorter (Plenum Press, London, 1972) p. 401.

[3] S. Clementi and S. Wold, in: *Chemometric Methods of Molecular Design*, ed. Van der Waterbeemd (VCH Publishers, New York, 1995) p. 319.
[4] D.F. Ewing, in: *Advances in LFER*, eds. N.B. Chapman and J. Shorter (Plenum Press, New York, 1978) p. 357.
[5] O. Exner, *Korelační Vztahy v Organické Chemii* (SNTL, Praha, 1981).
[6] L.P. Hammett, Trans. Faraday Soc. 34 (1938) 96.
[7] C. Hansch, in: *Drug Design*, Vol. 1, ed. E.J. Ariens (Academic Press, New York, 1971) p. 271.
[8] C. Hansch, in: *Correlation Analysis in Chemistry – Recent Advances*, eds. N.B. Chapman and J. Shorter (Plenum Press, London, 1978) p. 397.
[9] C. Hansch and A. Leo, *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, ACS Professional Reference Book (Washington DC, 1995).
[10] C. Hansch, D. Rockwell, P.Y. Jow and E.E. Steller, J. Med. Chem. 20 (1977) 304.
[11] C.D. Johnson, *The Hammett Equation* (University Press, Cambridge, 1973).
[12] V.A. Palm, *Osnovy Kolichestvennoj Teorii Organicheskih Reakcij* (Khimiya, Leningrad, 1977).
[13] J. Shorter, *Correlation Analysis in Organic Chemistry* (Clarendon Press, Oxford, 1973).
[14] R.W. Taft, J. Am. Chem. Soc. 75 (1953) 4231.
[15] S. Wold and L. Ericsson, in: *Chemometric Methods of Molecular Design*, ed. Van der Waterbeemd (VCH Publishers, New York, 1995) p. 309.
[16] P. Zuman, *Substituent Effects in Organic Polarography* (Plenum Press, New York, 1967).